Toivo, Taru Iris Wilhelmiina (2020) *Reduced emotional resonance in bilinguals' L2: Potential causes, methods of measurement, and behavioural implications*. PhD thesis.

University of Glasgow | College of Science & Engineering

Reduced emotional resonance in bilinguals’ L2: Potential causes, methods of measurement, and behavioural implications

Taru Iris Wilhelmiina Toivo (M.Sc., M.A.)

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

February 2020

School of Psychology
University of Glasgow
62 Hillhead Street
Glasgow
G12 8QB

Abstract

Bilinguals often report feeling “less” in their second language (L2). While a speaker might be fully proficient in their L2, it may not feel the same as one’s L1 does; in some cases bilinguals feel like their L2 is more emotionally distant, or even fake. This phenomenon, called *reduced emotional resonance of L2*, has been studied using a number of different methodologies ranging from questionnaire-based approaches to physiological measurement of emotion. The field, while truly interdisciplinary, lacks consensus on measurement practices. This thesis aims to address some of the most prevalent methodological issues in studying reduced emotional resonance of L2, namely how the word stimuli should be selected and normed, and provide guidance to conducting studies with word stimuli.

This thesis presents six studies, which investigate the causes, measurement methods and implications of reduced emotional resonance in bilinguals’ L2. Chapter 2 focuses on the causes of reduced emotional resonance, and measures it with pupillometry. The potential causes of reduced emotional resonance are examined by trying to predict bilinguals’ physiological responses to emotional language from their language background information.

Chapters 3-5 focus on the methodological aspects of reduced emotional resonance. Chapter 3 attempts to contrast different physiological measurement techniques of emotion. Comparing pupillometry and skin-conductance measurement, the chapter points out differences in paradigm design and sensitivity of these two techniques. Chapter 4 investigates the reliability of cognitive paradigms as measures of bilingual emotion, points out the importance of including stimulus item covariates in both stimulus selection, as well as analysis stage, and discusses why the use of translation equivalents is problematic. In this chapter, we compare a Lexical Decision Task to a pupillometry task in bilinguals’ L1 vs. L2, and in bilinguals vs. monolinguals. Chapter 5 looks into metacognitive measurement and compares affective word ratings with a pupillometry task to establish whether physiological responses to, and conscious evaluations of emotional words are related.

Chapter 6 focuses on the behavioural implications of reduced emotional resonance of L2. Behavioural implications have typically been studied in the context of moral decision-making. Here, we expand this literature to attributions. Through two experiments, this chapter investigates whether Optimality bias (assigning more blame to actors who make suboptimal choices) will be mediated by the Foreign Language Effect. In other words, whether doing the experiment in one's L2 will mitigate the Optimality bias.

Finally, chapter 7 discusses the key findings and common themes to stem from the experiments, as well as the limitations and potential future directions for the field. The main contribution of this thesis is to provide systematic, methodology-focused work on reduced emotional resonance in bilinguals' L2, to point out methodological inconsistencies, and to provide more robust alternatives for stimulus selection processes and statistical analyses of bilingual data.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost I would like to thank my supervisor Christoph Scheepers. Thank you for always being there for a chat (be it political or model-related), data wizardry, and pointing out the ambiguities in my syntax. It has been a pleasure working with you.

I would also like to thank my second supervisor Sara Sereno for her input and support, Catherine Caldwell-Harris for welcoming me in Boston and for her expertise, help and inspiration, and the whole UofG Psychology teaching team for always making me feel welcome and keeping me sufficiently distracted from this thesis.

A big thank you to all my participants, who have given their time to provide not only data, but also very interesting discussions about bilingualism.

Thank you to all my friends and family for your continuous support. And a special thank you to the Labless lab group (best lab mates are the ones you have chosen), to Holly and Maggi for being there through all parts of the process, to all my amazing flatmates (past and present) who always take care of me and make me laugh, to Riccardo and Laura for all their help, to Violeta for absolutely everything, and to Agnieszka, who inspired me to do research on bilingual emotions in the first place.

Chapter 1 Introduction

Bilinguals often feel “less” when speaking in their second language (L2). Despite being proficient and having an excellent command of the language, it may not *feel* the same as one’s first language (L1) does (Pavlenko, 2006). In some cases, second language can feel detached or even fake (Pavlenko, 2005). Bilinguals may also refer to an emotional distance in L2 (Degner, Doycheva, & Wentura, 2011), as opposed to their L1, which usually feels more like the language of emotions (Dewaele, 2010a). This emotional distance may make it easier to discuss embarrassing topics (Bond & Lai, 1986) or swear in one’s L2.

The literature does not have one consistent term to discuss bilinguals “feeling less” in their L2. Usually terms such as reduced emotionality, perceived emotional intensity or emotional distance are used. In order to keep the terminology consistent across this thesis, we will use the term *reduced emotional resonance of L2,* or in L2 to describe the phenomenon. This term is chosen to reflect bilinguals’ feeling of emotional intensity, rather than expressing or experiencing different emotional states, as reduced emotionality would suggest. Doing mostly physiological measurement, we chose this term to reflect physiological response to emotional stimuli, as well as the possible perceived emotional response.

Empirical research and theoretical considerations of reduced emotional resonance in bilinguals have vastly expanded over the past decade, but the field is lacking consensus and rigorous methodological investigations. This thesis aims to address this with a focus on methodological issues. We also hope to expand the literature on underlying reasons and implications of reduced emotional resonance of L2

1.1 The aims and structure of this thesis

In this introduction (Chapter 1) we will outline the concepts and definitions for the following chapters, and briefly discuss relevant background information.

The main body of this thesis is divided into three parts examining different aspects of reduced emotional resonance of second language. The first part investigates underlying reasons of reduced emotional resonance; what makes bilinguals feel less in their second language, and which factors in one's language background predict this? The aim of this part is to expand on the scarce literature on the predictors of reduced emotional resonance and do this with a physiological measure of emotion. The first part includes chapter 2 (Experiment 1).

The second, most substantial part of this thesis looks into different measurement methods of reduced emotional resonance and attempts to compare them. The three larger measurement strands we will focus on are physiological measurement (pupillometry and skin-conductance measurement), behavioural measurement with cognitive paradigms (we will use Lexical Decision Task), and meta-cognitive assessment (here, affective word ratings). Thus far, several different measurement methods and stimuli selection techniques have been used, but there is very little work on exploring the methodological aspects of the topic. The aim of this part is to provide a systematic account of the three main measurement strands and emphasise the importance of including lexical covariates in the stimuli selection process, as well as during the analysis stage. This part consists of chapters 3, 4 and 5 (Experiments 2-4).

The third part of this thesis is focused on implications of reduced emotional resonance, looking at optimality bias in bilingual decision-making. The aim of the last part is to extend implications literature from moral decision-making into attributions and to investigate whether the foreign language effect mediates how we judge other people. The third part of this thesis includes chapter 6 (experiments 5 and 6). Chapter 6 has been published in the Journal of Cross-Cultural and Cognitive Psychology (Bodig, E., Toivo, W., & Scheepers, C. (2019). Investigating the foreign language effect as a mitigating influence on the 'optimality bias' in moral judgements. *Journal of Cultural Cognitive Science*, 1-15.)

The final chapter (7) will provide a discussion of the common themes, findings and future directions drawn from the experimental chapters.

1.2 Definitions of bilingualism

This thesis will define bilingualism as the regular use of two or more languages (Grosjean, 2008). Here, *bilingualism* refers both to bi- and multilingualism. As many of our participants speak more than two languages, an umbrella term will be used for clarity. *A bilingual* refers to a person, who speaks two or more languages in their everyday lives, regardless of their level of fluency in these languages (Grosjean, 2008; Pavlenko, 2012).

1.2.1 Early vs. late bilinguals

For some of the experiments in this thesis (3, 5 and 6), the distinction between early and late bilinguals is crucial. The concept of early and late bilinguals has been much debated in the literature – there is no one clear cut-off age and the definitions differ widely. According to some definitions (Pavlenko, 2005; Tao , Marzecová, Taft, Asanowicz, & Wodniecka, 2011), learning a language any time before puberty will make an L2 user an early bilingual. Other definitions require the speaker to have learnt their L2 before a certain age cut-off, such as the age of 6 (Gathercole & Moawad, 2010), school age (Pelham & Abrams, 2014), or the age of 3 (Kapa & Colombo, 2013). Sometimes there is a distinction between simultaneous bilinguals and early bilinguals. Being exposed to multiple languages from birth makes the speaker a simultaneous bilingual, whilst acquiring an L2 before starting school makes the speaker an early bilingual (De Houwer, 2005). Some studies also distinguish between language exposure and language use (e.g. Kapa & Colombo, 2013).

In experiments 1, 2 and 4, we included speakers who fit the definition of speaking two or more languages in their everyday lives. These samples include speakers from a wide range of linguistic backgrounds. In experiments 3, 5 and 6 we only tested late bilinguals. In these experiments, speakers who have learnt their L2 before the age of 6 years were excluded.

To explain this specific age cut-off, the difference between foreign language (FL) and L2 should be noted. Foreign language typically refers to a language studied in an educational context, whereas L2 is a language used in the speaker’s daily life (Pavlenko, 2005). This distinction is important to how we have selected the early and late bilinguals for experiments 3, 5 and 6; we excluded any bilinguals who had learnt English before the age of 6. This specific cut-off was chosen, as we wanted to test participants who have not acquired English in a naturalistic setting at an early age (for example from family or by immigrating to the country where the language is spoken). These participants may have started to learn English as an FL before puberty (typically at school around the age of 7-9) but have only later in life moved to an English-speaking country and started to use it as an L2.

1.2.2 Defining a bilingual vs. monolingual speaker

According to some scholars' and lay definitions, only speakers who have been exposed to all their languages from birth, and have achieved native level of fluency, are bilinguals (Pavlenko, 2012). These views stem from a monolingual normative definition of bilingualism, and essentially consider bilingualism as "fractional"; they argue that a bilingual speaker is two monolinguals in one (Grosjean, 2008). Because monolinguals have been long considered the model of a "normal" speaker, the study of bilingualism has largely been focused on its cognitive and developmental consequences, trying to evaluate the good, the bad and the ugly. This thesis attempts to look beyond the functions and forms of bilingualism. Our aim, even in the chapter discussing implications of reduced emotional resonance, is not to assess the harm or utility of speaking multiple languages, but to simply examine a phenomenon integral to the experience of many bilingual speakers.

The tradition of monolingual speaker as the norm not only affects the direction of theory and empirical research are taking, but also reflects the terms used in the literature. Dewaele (2017) has argued against the use of terms native vs. non-native, as it posits one superior to the other. To avoid this type of normative language, this thesis will use the term *L1 speaker* or *monolingual* when discussing the "native" speakers, namely the control groups in experiments 1-5. These speakers only speak English in their daily lives. When discussing speakers who use two or more languages, we will use the terms *L2 speaker/user* or *bilingual*.

1.2.3 L2 vs. LX

Recently, the emphasis on defining bilingualism has moved away from traditional, categorical thinking to seeing bilingualism more as a spectrum of different language backgrounds, each unique in their own way. The more modern take on bilingualism is also moving away from defining a clear “order” of languages for the speaker (Dewaele, 2017). This idea allows for the concept of first and consequent languages to be more dynamic, based on the context of learning, context of current use and which language the speaker feels to be more dominant or their ”main language”.

Sometimes the distinction between first and consequent languages is marked as L1 and LX, instead of L2 (Dewaele, 2017). This allows for a more inclusive take on individual’s languages, as they may speak multiple languages, and their dominance order and proficiency may switch dynamically. In this thesis, we will use the markers L1 and L2. Our participants often speak multiple languages and come from a variety of backgrounds, but in most cases only speak their respective L1 and English daily. Hence, most of our participants will classify English as their L2 with high proficiency. When discussing specific experiments within this thesis, the term L2 will always refer to English. When discussing general theory, L2 is an umbrella term for “foreign” languages (Dewaele, 2017). It should be noted that L2 does not necessarily refer to the language that was chronologically acquired second but any language that was acquired later in life.

We acknowledge there are individuals, particularly in Experiments 1 and 4, who have more complex language backgrounds and may consider themselves balanced bilinguals – we attempt to capture this variety with our language background questionnaires and extensive demographic details in method sections and in the appendices. However, for clarity, we will use L1 and L2, with L2 here always referring to English unless otherwise stated. Some of our participants have acquired their two languages from birth – L2 in their case will also refer to English, even though they may have acquired it concurrently with their other language. This is to keep definitions constant across all participants and all experiments.

1.2.4 Comparing monolinguals and bilinguals?

From the monolingual normative point of view, the most intuitive experimental design is to compare bilingual speakers with monolingual speakers. However, the holistic view of bilingualism considers the bilingual speaker as an integrated person and not an individual decomposed into parts (Grosjean, 2008). For this reason, it has been argued that when studying bilingual speakers, we should move beyond the monolingual-bilingual comparison, and not to examine the bilingual speakers' one language without examining the other (Grosjean, 2008). Surrain & Luk (2019) also argue against simplistic monolingual-bilingual comparison; they highlight that bilingualism is a dynamic and an interactive experience, and that researchers should attempt to report the social context and other factors of their participants' language use and ability.

For experimental design reasons, this thesis will include experiments where we compare bilingual speakers to a monolingual control group (L1 speakers of English). However, where possible, we will also attempt to compare across the languages of the bilingual speakers, rather than just across speaker groups. Where possible, we will include a detailed account of the participants to reflect their language background.

1.3 What do we mean by emotion?

This thesis will adopt a definition suggested by Pavlenko (2008): emotionality refers to autonomic arousal elicited by particular languages or words, examined directly and indirectly, through verbal and non-verbal behaviours and self-perceptions. The key concept here is *autonomic arousal,* as mostly measured with physiological methods. For the most part of this thesis, we will focus on the emotional arousal dimension of the stimuli words, not assessing the quality or direction of the participants' emotional response, but rather the magnitude of it. In experiment 3 we will briefly delve into the valence dimension and explore whether there is a positivity bias in bilingual emotion processing.

One could argue this is somewhat consistent with the universalist view of emotion – the focus is on body experience, and emotions are biologically determined processes (Pavlenko, 2005). According to this view emotions share common experiential qualities recognised universally (such as the Ekman six). This view suggests that both language and concepts are secondary, and that emotions are strictly rooted in bodily experience (Pavlenko 2005). It should be noted this thesis will not take a stance on the universality of emotions, or any specific emotion states – we are simply interested in the autonomic arousal, whether that is reduced in bilinguals, and how this can be seen through word processing.

1.4 Grounded cognition and language embodiment

While this thesis is not focused on grounded cognition, it will be touched upon in some of our methodological considerations in the later chapters. Consequently, it is useful to provide a brief account of embodiment here.

Traditionally, semantic systems have been viewed as separate from modal perception systems. According to this view, they operate on abstract, amodal symbols (Barsalou, 2008; Zwaan, 2014). Grounded cognition theorists have challenged this view, suggesting that cognition, in fact, is grounded in our perceptual experience across multiple domains such as sensorimotor and affective processing (Barsalou, 2010).

Embodiment is a facet of grounded cognition, focused on how cognition is grounded through the human body. Embodiment is defined as the grounding of cognition in systems, which process low level perceptual and action information (Monaco, Jost, Gygax, & Annoni, 2019). Embodied theories of cognition claim that higher cognitive processing, including language, activates the same brain sensorimotor structures involved when experiencing the environment (Monaco et al., 2019). In bilingualism research, the term embodiment is typically used to describe the contextual and autobiographical relationship a speaker has with a language (Pavlenko, 2005).

Here, we will use the term embodiment instead of grounded cognition. This term is chosen to keep it consistent with previous bilingualism research (see Pavlenko, 2012), but also because in all the physiological experiments (experiments 1-4) we are measuring how the body responds the emotional dimension of words. It should be noted that these body responses are likely to be grounded through situations as well, and not just through the bodily experience. Grounded cognition theorists have argued that cognition is not dependent on bodily states, even though they can be closely related.

The key question to address is whether bilinguals' L2 is disembodied. Most of the language embodiment literature in bilinguals has investigated embodiment of motion and somatic simulation. There is no clear consensus on whether L2 is disembodied or not. For example, Dudschig and colleagues (2014) found no differences in motor responses to spatially associated words (such as star) between L1 and L2 conditions, suggesting both languages are embodied. Xue and colleagues (2015), on the other hand, found sensorimotor differences between the L1 and L2. Monaco and colleagues (2019) suggest in their review that L2 is at least partially embodied, but the mechanisms of embodiment and disembodiment are not fully understood yet.

Some research has also investigated embodiment through mental imagery in bilinguals (Hayakawa & Keysar, 2018). Across three different mental imagery tasks they found that the use of foreign language reduced the vividness of mental imagery. While we are mainly focusing on the emotional aspect of language processing, the findings summarised here demonstrate that L2 embodiment is a much wider concept – Hayakawa and Keysar (2018) also related their embodiment findings to bilingual decision making, which has traditionally been considered a natural consequence from reduced emotionality.

Foroni (2015) looked at motor simulation of emotional language processing, which is more relevant to the questions this thesis is examining. They measured muscle activation in the participants' faces upon being exposed to emotional language and found that the L2 words were only partially simulated, suggesting they may be disembodied. Another study (Baumeister, Foroni, Conrad, Rumiati, & Winkielman, 2017) used a memory task to examine bilingual emotional responses (facial muscle activation and skin-conductance response). They found that the enhanced memory effect for emotional content was stronger in L1, and they found partial evidence for decreased facial motor resonance in L2, aligning with the previous findings (Baumeister et al., 2017). They suggested that this increased affective encoding and retrieval of L1 content is due to the embodied knowledge, which is involved in emotional memory processes (Baumeister et al., 2017).

Some scholars argue that disembodied affective processing in second language is a possible cause of reduced emotional resonance (Pavlenko, 2012). The differences between L1 and L2 in affective processing may be due to the languages being embodied differentially. Particularly for late bilinguals and FL users, it is possible the words are processed only on a semantic level (Pavlenko, 2012).

Kuhne and Gianelli (2019) argue in that abstract vocabulary can carry emotional load, and there is substantial evidence for embodiment happening through emotion, and not only through motion. This is consistent with the idea that reduced emotional resonance in L2 occurs because the emotion in L2 words is not fully embodied. This notion is discussed in more detail below in the theoretical accounts section.

Pavlenko (2005) argues that when language is embodied, it will elicit physiological responses as well as sensory images. Language becomes embodied through two simultaneous processes. Conceptual development occurs when the speaker acquires denotative meaning, and words occurring across multiple contexts will form conceptual categories for them. In parallel to this, words and phrases will acquire affective connotations when they are integrated with emotionally arousing memories and experiences (Pavlenko, 2005).

FLs that are acquired through formal education are often disembodied and de-contextualised, as the speaker does not acquire affective connotations or conceptual development of words and phrases. This view naturally posits that language learning is a continuum ranging from naturally acquired, always contextualised L1 to de-contextualised classroom FL (Pavlenko, 2005). Hence, language embodiment is a dynamic concept, which can change through L2 socialisation. It should also be noted that bilinguals undergo various socialisation processes, which in turn lead to different affective conditioning, different conceptualisations across speakers’ languages, and differences in how the speakers perceive the affective status of these languages (Pavlenko, 2005).

1.4.1 Words vs. concepts

When discussing the possibility of disembodied cognition in bilinguals, it is important to outline the difference between word meanings and concepts. Pavlenko argues (2005) that a speaker can recognise and understand a word without having any conceptual representation of it. Conceptual representation is a prototypical script which is formed through repeated experiences of the word in several contexts. Across these contexts, the word is then associated with consequences, different means of display and regulation, as well as physiological responses (Pavlenko, 2005). Hence, bilinguals may understand the meaning of affective words, but lack the full conceptual representation of it.

This is consistent with the early models of bilingual word processing. Initially proposed by Kroll and Stewart (1994), the revised hierarchical model of bilingual language processing suggests that words in L2 often have fewer conceptual representations than words in L1. First language words are directly linked to conceptual representations. On the other hand, L2 words, especially early in the language learning process, are learnt through L1 translations and hence only weakly linked to conceptual representations or linked to them through the L1 translations. According to this model, as proficiency increases, the links become stronger.

It should be noted that the effects of conceptual representation and consequently language embodiment are not only due to increased proficiency. This thesis argues that reduced emotional resonance in fact is a separate construct, and not just a function of language proficiency.

1.5 Underlying reasons

1.5.1 Theoretical background

There are three main theoretical approaches to bilingual emotion processing: brain maturation accounts, Emotional context of learning theory (Harris, Gleason, & Ayçiçeği, 2006) and Language Embodiment Theory (Pavlenko, 2005).

The brain maturation accounts argue that L1 is typically more emotional because L1 is most often learnt in childhood, and thus language learning and development of brain areas associated with emotion happen concurrently (Harris et al., 2006). This makes the emotional meaning of words more deeply encoded in the brain (Pavlenko, 2005).

The theory of language embodiment (Pavlenko, 2005) posits that language acquisition consists of two processes, as briefly outlined above: conceptual development and affective linguistic conditioning. These processes are interrelated, and both contribute to the experience of language embodiment (words elicit both physiological reactions and sensory images). Denotative meaning of words and phrases develops through conceptual development, while affective linguistic conditioning creates affective connotations and personal meanings of words via association and integration with emotional experiences and memories. Fundamentally, the theory suggests that emotional processing of language is based on a form of operant conditioning; words become the conditioned stimuli, which then elicit an emotional reaction as a conditioned response. Consequently, the basis for emotional resonance in a language is autobiographical.

In relation to reduced emotionality in L2, the language embodiment theory suggests that L2 rarely becomes embodied due to less experience of emotional situations and the non-naturalistic setting of learning. This approach explains the role of the Context of Acquisition (CoA): as bilinguals are exposed to L2 words in more emotional contexts, the links to autobiographical memory and conceptual representations are strengthened. Furthermore, the idea of disembodied L2 would explain why L2 is often experienced as more emotional if the CoA is naturalistic - primary language is linked to childhood memories, traumas and anxieties (Pavlenko, 2005). Thus, when L2 is acquired primarily through declarative memory, the L2 self can be perceived detached or unemotional (Pavlenko, 2005).

Emotional context of learning theory argues that L2 has reduced emotional resonance due to the context it has been learnt in (Harris et al., 2006). The theory largely agrees with the other two approaches but adds to them based on the empirical evidence obtained from studies directly examining the processes behind emotional resonance (Dewaele, 2008; Harris et al., 2006). It claims that age of acquisition and brain maturation are not the only explanations for language emotionality, yet largely influential due to greater emotionality in the language one is exposed to at family and childhood settings. Further, the theory argues “language is stored with its context of occurrence” (Harris et al., 2006) – human learning is associative, hence exposure to multiple examples across different contexts facilitates greater emotional connotations. This would explain why more naturalistic settings of language acquisition usually result in greater emotionality of L2. The theory argues that the context affects emotional resonance of a language more widely, not only as autobiographical memories.

The language embodiment theory and the Emotional context of learning theory are both consistent with the *contextual learning hypothesis* of how language acquires emotional meaning (Barrett, Lindquist, & Gendron, 2007; Braun, 2015). This hypothesis suggests that the process of linking verbal and emotional information is mediated by learning and experience.

1.5.2 Empirical accounts

Somewhat surprisingly, the literature on why reduced emotional resonance occurs and what predicts it is scarce. To date, the Bilingualism and Emotions questionnaire (BEQ) (Dewaele & Pavlenko, 2001-2003) is the largest systematic account of the predictors of reduced emotional resonance. The questionnaire study was conducted online in 2001-2003, capturing responses from nearly 1600 multilingual speakers around the world.

In the Bilingualism and Emotions Questionnaire, L1 was typically rated more emotional and more likely to be used for expressing emotion. The perceived emotional resonance of a language was mediated by four factors: Age of Acquisition (AoA), perceived language dominance, Context of Acquisition (CoA) and order of acquisition (Dewaele, 2010). A skin-conductance experiment looking at physiological responses to phrases in the participants’ L1 and L2 found a similar pattern; the earlier the acquisition, or the more naturalistic the learning context, the more likely L2 was found to have equal emotional resonance when compared to L1 (Harris, 2004).

In a sub-analysis of one of the questions of the BEQ, assessing the weight of the phrase “I love you” in the speakers’ different languages, language dominance, AoA and CoA also predicted the perceived emotional weight of the phrase. Additionally, degree of socialization in L2, the network of interlocutors in their L2, and self-assessed oral proficiency in L2 predicted the perception of the emotional weight. Another study specifically looking at the phrase “I love you” tested Polish-English bilinguals who were immersed in their L2 (English) (Ożańska-Ponikwia, 2017). Their results align with those of the BEQ. The emotionality the phrase was perceived stronger in the participants’ L1, but the perception in L2 was mediated by the length of stay in an English-speaking country. Further, self-perceived L2 proficiency and frequency of L2 use were found to affect the perceived emotionality, as well as socialisation into the L2 culture and the degree of L2 use.

Other studies have also found that frequency of language use affects emotional resonance (Degner et al., 2011). In an affective priming task, it was found that the priming effects were larger in L1, and only appeared in L2 in participants with high L2 use frequency and high level of immersion in their L2. The frequency effect was replicated in a study looking at Finnish L1 speakers and comparing their L2 and L3 – it was found that the participants self-ratings of language emotionality were mediated by the frequency of everyday exposure to the language (Räsänen & Pine, 2012). Pavlenko (2012) divided the predictors identified in these studies into two superordinate groups. The order of acquisition and AoA can be combined into *age effects*, whereas CoA, frequency of use and language dominance are *context effects*.

There is also some conflicting evidence in relation to the predictors outlined above. For example, Eilola, Havelka and Sharma (2007) used an emotional Stroop task and found similar interference effects independent of language (L1 or L2) for highly proficient late bilinguals. Furthermore, in qualitative research and self-ratings bilinguals who have had a shift in their dominant language and use L2 in highly emotional contexts (e.g. with family or spouse) sometimes report higher emotionality in L2 (Pavlenko, 2008). These findings suggest that that age effects per se are not a sufficient explanation for reduced emotional resonance of L2.

The second large-scale, systematic investigation of bilingual emotion processing also contrasts the theory and BEQ findings. Ponari and colleagues (2015) found that none of the above factors were predictive of the emotional response to words differs between L1 and L2. Testing L1 speakers of several typologically different L1s with English as their L2, and English L1 speakers, they found that the processing of high and low valence words was not different between the two groups. Reaction times were not mediated by frequency of use, language immersion, or age of English acquisition, which contrasts with all three bilingual emotion processing theories discussed above. However, it is possible that cognitive paradigms simply do not capture the reduced emotional resonance effect very well – see section Cognitive Measurement below and chapter 4 for further discussion.

This thesis attempts to assess the underlying reasons of reduced emotional resonance systematically, and through an experimental approach, measuring autonomous emotional response rather than self-reports. This will be discussed in more detail in chapter 2.

1.6 Methods of measuring

The most substantial part of this thesis will focus on the different measurement methods of reduced emotional resonance of L2. Here, we will only provide a brief description of each of the methods; the literature and experimental evidence will be discussed in more detail under each of the experimental chapters.

1.6.1 Stimulus selection

One of the main aims of this thesis is to address the issues in stimulus selection in the literature. Typically, the stimuli are selected manually and lexical covariates are not included in the analyses or even during the initial stimulus selection process. This is problematic as other areas of psycholinguistics research have established that word-processing depends on the lexical properties of the word (Brysbaert, Warriner, & Kuperman, 2014; Kousta, Vinson, & Vigliocco, 2009; Scott, O’Donnell, & Sereno, 2014)

The main lexical feature we are manipulating is emotional arousal of the words, as rated by participants of previous norming studies (for example, Warriner, Kuperman, & Brysbaert, 2013). In study 3, both emotional arousal and valence are manipulated. Then, for each stimulus set we have collected lexical covariates from a number of different databases. The stimuli set for experiments 1, 2 and 4 has been matched on nine lexical covariates, and the stimuli set for experiment 3 has been matched on seven lexical covariates. The specific covariates are detailed under the method sections of each of the experiments. These covariates have also been included in the statistical analyses - even though we attempt to control for them, some variance remains and addressing this will help to increase modelling accuracy.

Including the covariates increases the complexity of the models, and this in turn may cause issues with power as well as convergence. We have addressed this by running a Principal Component Analysis on each of the stimuli sets to reduce the number of predictors. The principal components have then been entered as predictors in the mixed effects models. All stimulus sets can be found in full in appendix B, and with covariate information on the Open Science framework (https://osf.io/9rqbj/).

Through the experimental chapters 3-5, we attempt to illustrate the discrepancies in stimulus selection and statistical analyses in the current literature and provide an alternative method to more robust examination of bilingual affective language processing.

1.6.2 Emotion words vs. emotion-laden words

There is an ongoing debate on how emotional words should be defined in bilingualism research (Wierzbicka, 2008). Some scholars argue emotion words (words directly describing an emotion, for example “happy” or “sad”) should be distinguished from words that bear emotional meaning, i.e. emotion-laden words. Some studies have indeed distinguished the two into separate categories (see: Kazanas & Altarriba, 2016), or different categories of emotionally arousing words (for example, Caldwell-Harris, Tong, Lung, & Poo, 2010). Whilst we agree this is an important aspect of emotional processing of language, the way it complicates experimental design is outwith the scope of this thesis – we are interested in the emotional arousal dimension, not the meaning of the words per se.

Wiezbicka (2008) argues that the field of bilingual emotions is lacking in precision in how emotion concepts are defined, and which exact part of lexicon we are studying. This thesis aims to address this conceptual - and as we argue, also methodological – issue. In order to standardise and match our stimuli sets as strictly as possible, we are not separating the emotionally arousing words into further categories. Hence, this thesis will simply focus on the emotional arousal aspect of the stimuli used, given that previous literature has established that this dimension can capture emotional responses in bilinguals (Toivo & Scheepers, 2019).

1.6.3 Brain imaging methods and EEG

Investigating reduced emotional resonance through ERPs and brain imaging is one of the four measurement strands. This thesis will mostly focus on physiological techniques based on the Autonomous Nervous System (ANS) activity, rather than measuring lexical activation in the brain, but it is worth briefly reviewing some of the ERP and brain imaging findings here.

ERP studies focusing on semantic integration and lexical activation have found strong supporting evidence for reduced emotional resonance in L2. For example, Wu and Thierry (2012) found that words associated with low (negative) emotional valence tend to block simultaneous activation of L1 and L2. Similarly, Jonczyk and colleagues (2016) showed lower N400 amplitudes for negative valence sentences in L2 than in L1. Other ERP studies using a Lexical Decision Task suggest weaker (Conrad, Recio, & Jacobs, 2011) or delayed (Opitz & Degner, 2012) automatic affective processing in L2. These findings suggest that there is a difference in how bilinguals process affective words, and that this in turn may affect lexical access.

In an fMRI study looking at reading passages of Harry Potter in L1 and L2, there were stronger hemodynamic responses in amygdala and the left pre-frontal cortex to happy vs. neutral passages, but this was only found in participants’ L1 (Hsu, Jacobs, & Conrad, 2015). The researchers suggested that these findings show that reading emotionally loaded texts in one’s L1 elicits a stronger emotional experience than reading in one’s L2 does.

Overall, findings from brain imaging studies suggest that they are a good method of detecting and quantifying reduced emotional resonance. Studies using EEG and a cognitive paradigm concurrently (for example: Chen, Lin, Chen, Lu, & Guo, 2015; Conrad et al., 2011; Opitz & Degner, 2012) can provide us with insight into the effectiveness of each of these measurement techniques, and are easier to conduct simultaneously than physiological measurement and a cognitive task.

1.6.4 Physiological measurement of emotion

Physiological techniques are based on the activation of the Autonomous Nervous System, which is automatic and involuntary. When exposed to emotionally arousing stimuli, the ANS activates and produces different bodily responses. The first physiological method used in the context of studying reduced emotional resonance is skin-conductance measurement. When the ANS activates, the skin produces a galvanic response and conducts more electricity than it does when exposed to neutral stimuli.

Caldwell-Harris and colleagues have studied bilingual skin-conductance responses (SCRs) across multiple modalities (visual and auditory), and multiple speaker groups. Typically, their stimuli have been split into multiple categories: childhood reprimands, insults, neutral phrases, taboo phrases, and positive and negative phrases. In their first experiment, Harris and colleagues (2003) compared the SCRs of Turkish-English late bilinguals in their L1 and L2. They found that L1 taboo words and childhood reprimands elicited higher SCRs in the speakers’ L1.

In later experiments with Spanish-English (Harris, 2004) and Mandarin-English bilinguals (Caldwell-Harris et al., 2010), the effect was replicated, but there were some noteworthy differences. In the Spanish-English study an Age of Acquisition effect was found on the SCRs – late L2 learners had stronger SCRs for childhood reprimands in their L1, but for early bilinguals this difference was not found (Harris, 2004). On the other hand, in a study with Mandarin-English speakers (Caldwell-Harris et al. 2010), the L2 endearments elicited higher SCRs than the L1 phrases did. This was interpreted as a reflection of cultural differences in emotion expression.

Followed by the skin-conductance response and skin conductance level (SCL) measurement, physiological techniques have been expanded to pupillometry. The underlying mechanism is very similar – pupillary response has been shown to be sensitive to emotionally arousing stimuli, such as pictures, sounds or words (Bradley, Miccoli, Escrig, & Lang, 2008; Partala & Surakka, 2003). As the ANS activates through being exposed to an emotional stimulus, pupils dilate as a response (Partala & Surakka, 2003). It is important to note here, that pupillary response is also sensitive to increased cognitive effort, for example through an increased memory load (see: Schmidtke, 2014 for review), and to surprise (Kloosterman et al., 2015; Scheepers, Mohr, Fischer, & Roberts, 2013). These are possible confounds we are assessing at every step of experimental design and analysis; it has been found that bilinguals' pupillary responses may be larger when stimuli are presented in their L2 (Schmidtke, 2014), as there is often an increased cognitive load associated with speaking in one's L2.

Previous pupillometry experiments looking at bilinguals' responses to emotionally arousing language have investigated the difference in pupillary responses between L1 and L2 words in single words (Toivo & Scheepers, 2019) and sentences (Iacozza, Costa, & Dunabeitia, 2017). In the first pupillometry experiment using pupillometry as a measurement method of reduced emotional resonance of L2, Iacozza and colleagues (2017) tested Spanish-English participants in either their L1 (Spanish) or L2 (English). Target words, each low in valence but high in arousal, were embedded in sentences that the participants read. They found that the difference between pupillary response to emotional vs. neutral stimuli was smaller in L2 as opposed to L1, suggesting that the participants showed reduced emotional resonance in their L2. This experiment only included low valence words as their target stimuli.

In our first pupillometry experiment (Toivo & Scheepers, 2019), we tested Finnish-English and German-English bilingual speakers in their respective L1 and L2 (English), and a control group of monolingual English speakers. The participants were shown high arousing, low arousing words and neutral distractor words, while their pupillary response was recorded. We found that the difference between pupillary response to HA and LA words was smaller in the participants' L2 as opposed to L1. This interaction, significant both when comparing monolinguals vs. bilinguals and bilinguals' L1 vs. L2, was interpreted as the reduced emotional resonance effect. This experiment is the basis for all the pupillometry work in this thesis, and we are using the same paradigm slightly altered.

1.6.5 Cognitive measurement

Another large measurement strand are standard cognitive paradigms such as Stroop task, Implicit Association Test or Lexical Decision Task. Psycholinguistics research on other domains has established that depending on the test type, emotional arousal and/or valence of the words will either interfere with the response time or facilitate it (Kousta et al., 2009). The studies which use cognitive measurement techniques rely on the assumption that emotional activation is faster and more automatic in speakers' L1. Hence, the hypotheses of reduced emotional resonance studies using cognitive paradigms typically expect an L1 advantage effect, or an L2 advantage effect depending on the paradigm type (Pavlenko, 2012)

The *L1 advantage effect* should be prevalent in paradigms in which the response time is facilitated by the emotional charge of the words, such as in the Implicit Association Test or Lexical Decision Task. Participants should have stronger emotional associations in their L1 as opposed to L2, which should lead to less pronounced differences between word types in L2. The *L2 advantage effect*, on the other hand, means that the *interference effects* of emotion are smaller in L2. This should happen in paradigms where the emotional aspect of the stimuli is expected to interfere with the task, such as in the Stroop Task. This again, would lead to a smaller difference in RTs between high arousal or extreme valence stimuli and neutral stimuli in participants' L2.

Studies using cognitive paradigms to measure bilingual emotion processing in language have typically produced very inconsistent results. Rather surprisingly, such paradigms are still being used as measures of reduced emotional resonance, even though the field has failed to convincingly answer whether these tests actually *can* detect reduced emotional resonance. There does not seem to be a clear pattern as to why sometimes the effect is found and sometimes it is not. The part of this thesis that examines cognitive measurement methods (Chapter 4) will try to address some of the potential issues that are prevalent in the literature.

The first problem we aim to address, not only in chapter 4, but across all the word-based experiments in this thesis, is how potentially important lexical covariates are controlled for in the stimulus selection process and at the analysis stage. There is usually at least some degree of higher cognitive effort associated with speaking in one's L2, and it is debatable whether this can ever be fully disentangled from other responses measured. We strongly believe that controlling for lexical confounds that may affect the ease of word processing (Such as word length and frequency, see: Ferrand et al., 2011) is important when testing bilingual speakers. While most studies account for at least some of the lexical covariates, this is usually not done in a systematic way. This thesis will advocate for controlling for lexical covariates both during the design and the analysis stages of research.

The second issue, which is partially related to not properly controlling for the lexical covariates, is the use of translation equivalents between L1 and L2. Translation equivalents are often used to compare participants' L1 and L2. Admittedly, there is a lack of validated databases to obtain lexical norms (especially norms other than frequency) for languages other than English. However, we argue that translation equivalents should not be used; effectively, if the study is within participants, stimuli word meaning will be repeated to the participants, creating additional dependencies in the data. Also, the lexical covariates of these translation equivalents are usually not controlled, which creates further problems in interpreting the results in a meaningful way (for example, see: Grabovac & Pléh, 2014; Opitz & Degner, 2012). On the other hand, Fan et al., (2016), did control for lexical covariates in the translation equivalents they used.

Cognitive measurement will be discussed in more detail in the introduction of chapter 4.

1.6.6 Metacognitive measurement

The third measurement strand of reduced emotional resonance is metacognitive assessment of emotion. This is done by asking the participants to rate words on different lexical dimensions, typically on emotional arousal, and in some cases also valence. Metacognitive assessment has not been used as frequently as the other two methods, and in many studies, it has only been used as a control measure rather than an experimental one.

Affective ratings have often not detected differences between bilinguals and monolinguals (for example: Iacozza et al., 2017; Ong, Hussain, Chow, & Thompson, 2017; Winskel, 2013). In a recent study Garrido and Prada (2018) found no main effect of test language, but the crucial interaction of word type and test language was found – L1 words were rates as more extreme in valence, as opposed to L2 words. However, this effect was only found for taboo words.

Dewaele (2016; 2018) found contrasting evidence to this in his rating studies. Instead of arousal or valence, the participants were asked to rate English terms on offensiveness. Bilingual participants (English as their L2) rated all the offensive terms as *more* offensive than L1 speakers did, except for the word “cunt”, the offensiveness of which was underestimated by the bilingual speakers. Dewaele (2016) suggested this may be due to the bilingual speakers overcompensating for the reduced emotional resonance they are experiencing when using these words. Perhaps the bilingual speakers are trying to avoid misusing the terms (Dewaele, 2016).

If bilinguals experience their L2 as being less emotional, why is this not reflected on an explicit measure of the emotional dimensions of language? The inconsistent findings with affective ratings may suggest that ratings of single words, as opposed to how a person feels in a language, tap into different sources of cognitive judgments. It is possible that rating single words isolated out of context and assessed consciously does not tap onto the concept-level of the words.

Puntoni and colleagues (2009) asked their participants to rate the emotional intensity of advertising slogans in L1 and L2. Interestingly, they found an effect of language on the emotionality ratings, showing that slogans in L1 were rated as more emotional than those presented in L2. This may show further evidence for the need of contextual cues in explicit rating measures.

Metacognitive measurement will be discussed in further detail in chapter 5.

1.7 Implications

The third focus of this thesis is on the implications of reduced emotional resonance. Implications of reduced emotional resonance have been studied in several different contexts, such as consumer behaviour and marketing (Klesse, Levav, & Goukens, 2015; Puntoni et al., 2009), superstitious beliefs (Hadjichristidis, Geipel, & Surian, 2019), causality judgments (Diaz-Lago & Matute, 2018) and illusory correlations, such as the hot-hand fallacy (Gao, Zika, Rogers, & Thierry, 2015).

The most substantial focus has been on how L2 might mitigate different cognitive biases in the context of decision-making. Speaking in an L2 has been found to make the speaker less prone to decision-making biases. This is called the Foreign Language Effect (FLE) (Costa, Foucart, Hayakawa, et al., 2014). The FLE has been typically studied with classic decision-making dilemmas, such as The Asian disease problem, where participants are asked to decide which medicine to choose to save lives (Keysar, Hayakawa, & An, 2012), and the footbridge (Costa et al., 2014) and the trolley dilemmas (Cipolletti, McFarlane, & Weissglass, 2015), in which the participant must decide whether they would sacrifice a single person to save the lives of several others.

There are two main domains in which the FLE has been found to affect speakers' decisions: loss aversion and utilitarian decisions (Costa, Vives, & Corey, 2017). When doing the experiment in their L2, participants are typically less risk averse, and less prone to framing effects in the dilemmas presented (Keysar et al., 2012). L2 also seems to prompt more utilitarian judgments (Costa et al., 2014).

Typically, the FLE is discussed in the context of the dual-system theories of decision making (Costa et al., 2017), positing that there are two routes of decision-making: one intuitive, and one more deliberative. It has been suggested that the increased cognitive effort associated with using one's L2 activates the deliberative route, which in turn is less prone to biases. This is called the "heightened utilitarianism" account (Hayakawa, Tannenbaum, Costa, Corey, & Keysar, 2017). The second possible underlying mechanism is the "blunted deontology" account. This account suggests that speaking in L2 and the reduced emotional response associated with it deactivates the emotional and heuristic processes associated with the intuitive route and, as a result, decisions in L2 are less prone to biases (Hayakawa et al., 2017). In their systematic study of the two accounts, Haykawa and colleagues (2017) concluded that across six experiments, the participant responses were more aligned with the blunted deontology account, suggesting that the origins of the FLE are in reduced emotion rather than increased deliberation.

This is consistent with findings suggesting that the FLE is mediated by emotionality, i.e. a causal role of emotion in decision-making (Corey et al., 2017). It has been found that FLE is present only in the footbridge dilemma, which involves an individual actively pushing another person off a bridge to save more lives, but not in the switch or trolley dilemma, where the person presses a switch instead of actively harming the individual (Cipolletti et al., 2015; Corey et al., 2017; Geipel, Hadjichristidis, & Surian, 2015a). This suggests that the FLE emerges only in scenarios which are highly emotional and morally compromising (Corey et al., 2017).

These findings seem to support a link between decision-making and reduced emotional resonance, but with the expanding literature on the underlying mechanisms of the FLE, contrasting evidence has also been found. Geipel and colleagues (2015a) tested participants on both the trolley and footbridge dilemma, and found that L2 reduced emotions in both dilemmas, but the interaction of test language and moral judgment was not mediated by reduced emotion. They suggest the FLE may be due to reduced access to moral norms in L2, as opposed to reduced emotion. This aligns with their previous findings suggesting that moral transgressions with no tangible consequences are judged less harshly in L2 (Geipel, Hadjichristidis, & Surian, 2015b).

There is also some evidence suggesting that FLE may emerge because of reduced mental imagery in L2 (Hayakawa & Keysar, 2018). In this study it was found that mental imagery in general was less vivid in L2, but also that it mediated the FLE in moral judgments that the participants made in the footbridge dilemma. These findings together expand the literature beyond the idea of intuitive vs. deliberative thinking and ties it back to the concept of L2 language embodiment on a wider scale.

The section of this thesis looking into implications of reduced emotional resonance in L2 (Chapter 6) aims to expand the FLE literature beyond moral judgments, to attributions. Attributions are another domain, which is notoriously susceptible to different cognitive biases, which in turn may mean that attribution judgments are less biased in L2.

Chapter 2 : Exploring the underlying reasons of reduced emotional resonance with pupillometry

2.1 Introduction

Reduced emotional resonance of L2 has been studied through various different methodologies, for example, online questionnaires (Dewaele, 2010a), psychophysiological measurements (Caldwell-Harris et al., 2010; Harris, 2004; Toivo & Scheepers, 2019), and more anecdotally from bilingual authors' memoirs (Besemeres, 2006). While numerous studies support the existence of the phenomenon, there is little research that systematically explores the underlying reasons for *why* bilingual speakers experience reduced emotional resonance in their L2.

One of the core assumptions for reduced emotional resonance is that languages acquired early in life have established stronger emotional associations than those acquired later. Harris et al. (2004; 2006), compared late and early bilinguals' skin-conductance responses to different types of emotionally arousing words, such as childhood reprimands. They found that the difference in skin conductance response to emotionally arousing words in participants' L1 and L2 was smaller in early bilinguals than in late bilinguals. This suggests that early acquisition facilitates physiological emotional reactions to L2 words much like to those in L1. Early acquisition as an indicator of greater emotional responsiveness is also grounded in the theory of reduced emotional resonance. The brain maturation approach posits that languages learnt in childhood are more emotional, as the brain areas associated with emotion processing, such as amygdala, are developing in conjunction with learning the language (Harris et al., 2006). Consequently, the emotional meaning of words is more deeply encoded in the brain (Pavlenko 2005).

It is unlikely that brain maturation is the only factor predicting reduced emotional resonance in L2 - some studies have found only weak, if any, support for age of acquisition (AoA) as a viable predictor. For example, Eilola and colleagues (2007) found no significant difference in the interference effects between highly proficient late bilinguals' languages in an emotional Stroop task. While using cognitive testing to measure emotional resonance of a language is controversial, their findings may suggest that AoA may not be a reliable predictor of reduced emotional resonance, and that it is possible for late bilinguals' two languages to bear similar emotional resonance. Further, in the Bilingualism and Emotions Questionnaire (Dewaele, 2010) AoA was a relatively weak predictor of perceived emotionality of a language. Given the variation in individual language backgrounds, it seems almost self-evident that the emotional context in which a bilingual's languages are framed should also affect the emotional resonance of a language.

Two existing theories account for the emotional context of a language. The theory of language embodiment suggests that emotional weight of a language is based on operant conditioning. Words become the conditioned stimuli, which then cause an emotional reaction as a conditioned response. According to the theory of language embodiment, the grounds for how emotional a language becomes are autobiographical. This theory argues that it is rare for a bilinguals' L2 to become embodied, because there is typically less experience of emotional situations in an L2, and it is often learnt in a less naturalistic setting than L1 (Pavlenko, 2005). The theory takes a wider and more situational approach than the brain maturation accounts, but also explains why languages learnt in childhood are often more emotional.

The emotional context of learning theory (Harris et al., 2006) is based on similar assumptions about language and its associations as the theory of language embodiment - in fact the differences between the two theories have not been extensively discussed in the literature. The emotional context of learning theory posits that L1 often has stronger emotional resonance as a result of the context of learning. Further, the theory argues "language is stored with its context of occurrence" (Harris et al., 2006). This notion is based on the associative nature of human learning; exposure to multiple occurrences of language across a variety of different contexts facilitates stronger emotional connotations. Both theories discussed above emphasise the role of emotional framing of a language across multiple contexts. Consequently, emotional resonance of a language is not a predetermined concept but rather a *dynamic one*.

The Bilingualism and Emotions Questionnaire, is an internet based self-rating study and perhaps the single most comprehensive research project yet conducted on bilingual emotions (Dewaele & Pavlenko, 2001-2003). The questionnaire gathered nearly 1600 bilinguals' responses. The study found participants' perceived language emotionality was mediated by the following five factors: Age of Acquisition, perceived language dominance (the higher the language dominance in a language, the higher the reported emotionality), Context of Acquisition (CoA), order of acquisition and frequency of use. However, as briefly discussed above, the effect of AoA was not very strong (Dewaele, 2010).

Skin-conductance findings align with the predictors established in the Bilingualism and emotions questionnaire. Harris and colleagues found that the earlier the acquisition, or the more naturalistic the learning context, the more likely L2 was to have equal emotional resonance when compared to L1 (Harris, 2004; Harris et al., 2006). Pavlenko (2012) divided the predictors discussed in these studies (order of acquisition, the AoA, the CoA, frequency of language use, and language dominance) into two superordinate groups. The order of acquisition and AoA together are combined into age effects, whereas CoA, frequency of use and language dominance are context effects.

There is also further support for frequency of language use as a predictor of emotional resonance. For example, Degner and colleagues (2011) found that frequency of language use affects emotional resonance; using an affective priming task, it was found that the affective priming effect in L2 was only prevalent in participants with high L2 use frequency and high level of immersion in the L2. The frequency effect was also found in a study that compared multilingual speakers' L2 and L3; self-ratings of language emotionality in L2 and L3 were mediated by the (self-rated) frequency of everyday exposure to the language (Räsänen & Pine, 2012).

While the Bilingualism and Emotions Questionnaire is very extensive, the data are based on self-reports of the participants and have been used as a database for a number of papers and book chapters (for example: Dewaele, 2004, 2008; Dewaele, 2010a; Pavlenko, 2006). Arguably, the field would benefit from a fresh dataset and a novel data collection approach - perceived emotionality, emotion conceptualisation and physical emotional reactions are distinct from one another, and all three should be studied in order to best understand how emotional resonance of a language works. No study to date has systematically inspected relationships between, and the relative importance of the predictors of language emotionality outlined above. Hence, the present study aims to explore which factors in an individual's language history and current language use predict greater emotional resonance in L2, using single-word stimuli.

This chapter will examine emotional resonance in L2 through physiological reactions to emotionally arousing words. We will use the paradigm established in Toivo and Scheepers (2019), in which the participants are shown words that are of high or low in emotional arousal, and their pupillary responses are measured. In the original study, the difference between pupillary response to non-arousing and arousing words was larger in bilinguals' L1, as opposed to L2. This interaction, found both when comparing bilinguals vs. monolinguals, and bilinguals' L1 and L2, was interpreted as reduced emotional resonance in L2.

Schmidtke (2014) has previously established that bilinguals' pupillary responses are sensitive to cognitive effort. This, in turn, may be problematic for the interpretation of any bilingual pupillometry findings; how should we disambiguate the effect of emotional arousal from the effect of cognitive effort? We will account for this confound by strictly controlling the lexical covariates in our stimulus set, and by including those covariates in the analyses. We will also include a word recognition task after each trial and ask the participants to complete a short English proficiency test.

The approach of the present study is exploratory. Our aim is to examine potential underlying factors of emotional resonance of a language and their relative importance through physiological responses in a bilingual sample. To achieve this, bilingual speakers with varying language backgrounds will be recruited.

The predictive factors will be collected through a questionnaire, the questions of which are based on the potential predictor variables discussed earlier: Age of Acquisition, Context of Acquisition, frequency of language use, order of acquisition and language dominance. Given that many of these predictors will be correlated (for example, we can assume that learning L2 from an early age will be associated with learning L2 at home), we aim to address this by reducing the number of questionnaire items into uncorrelated Principal Components. This approach will address the issue of variable collinearity in modelling, but also help to reduce the number of predictors in a given model (although some variance will inevitably be lost).

With a new English word stimulus set, we also expected to replicate the interaction found in Toivo & Scheepers (2019), by comparing a group of monolingual English speakers with bilingual speakers.

2.2 Method

2.2.1 Materials and Stimuli

A modified pen and paper version of the language history questionnaire (Li, Zhang, Tsai, & Puls, 2013) was used to identify participants' language background (see Appendix A for full questionnaire and coding scheme). Questions measuring national identification, accent, language learning skills and general proficiency were removed (participants' proficiency was measured with LexTALE proficiency test instead, see below under Procedure). Further situations of language use/context of acquisition were added (such as using a language with a partner or with flatmates), and some questions were shortened.

The stimuli for the eye-tracking part of the experiment consisted of 240 English words (80 high arousing, 80 low arousing and 80 neutral distractor words). The full list of stimuli words can be found in Appendix B. The candidate words were selected from the Warriner, Kuperman & Brysbaert (2013) affective lexicon database. The rating scale for arousal in this database ranges from 1-9. Words rated 6 or higher (67% or more arousal) were selected for the candidate pool as high arousing candidates (554 words). Words rated 2.9 or below (32% or less arousal) were selected for the low arousal category (724 words). Words rated 4.5-4.62 (~50% arousal) were selected for the neutral category (708 words).

Arousal, valence and dominance ratings were obtained from the Warriner et al. (2013) database. Further norms (length, frequency, logarithmic frequency, number of orthographic neighbours and bigram frequency) were obtained from the English lexicon project website (Balota et al., 2007). One hundred and seventy-six words of the initial candidate list did not appear in the English lexicon project database and were consequently removed. In the final step concreteness ratings were obtained from Brysbaert, Kuperman & Warriner (2014) concreteness ratings database. Thirsty-five of the words did not appear in the concreteness database and were removed, leaving a final candidate pool of 1775 words.

From this candidate pool the final stimuli were selected algorithmically. The words were selected as "triplets" – one high arousing, one low arousing and one neutral word were matched on the above-mentioned lexical covariates as closely as possible, via minimizing Euclidean distances in the multivariate space of lexical covariates considered.

Table 1 Means (SDs) per variable in each word category (N = 80 per category)

	HA words (e.g. *detonate*)	LA words (e.g. *parental*)	Neutral (e.g. *maverick*)
Arousal	6.36 (0.32)	2.68 (0.20)	4.55 (0.04)
Valence	4.59 (1.92)	5.45 (1.02)	4.95 (1.45)
Dominance	4.79 (1.37)	5.28 (0.82)	5.11 (0.98)
Number of Letters	7.51 (2.38)	7.31 (2.22)	7.45 (2.23)
Number of Syllables	2.49 (0.99)	2.45 (0.98)	2.49 (0.99)
Orthographic Neighbours	0.05 (1.00)	0.07 (1.00)	0.03 (1.02)
Frequency per million (log, HAL)	7.66 (1.74)	7.77 (1.82)	7.58 (1.55)
Frequency per million (log, SUBTLEX)	5.47 (1.62)	5.36 (1.57)	5.30 (1.52)
Bigram frequency	8.12 (0.52)	8.13 (0.53)	8.14 (0.50)
Concreteness	3.24 (0.88)	3.32 (0.95)	3.25 (0.97)

Next, these covariates were entered into a Principal Component Analysis (Varimax rotation with Kaiser Normalisation) to reduce the number of lexical covariates in the analysis stage. Five orthogonal factors were extracted, together explaining ~90% of the total variance across the 9 input variables. Table 2 below shows the component loadings of the 5 extracted components.

Table 2 Factor loadings after Varimax rotation.

	Principal Component				
	Length & Orthographic neighbours	Lexical frequency	Valence & Dominance	Bigram frequency	Concreteness
Valence	.003	.091	.925	−.030	.104
Dominance	−.061	.086	.927	.040	−.069
Number of letters	.739	−.202	.026	.544	−.065
Number of syllables	.777	−.135	−.004	.449	−.072
Orthographic neighbours	−.904	.197	.080	.001	.044
Lexical frequency (HAL)	−.121	.951	.107	−.051	.027
Lexical Frequency (SUBTLEX)	−.263	.900	.090	−.167	.059
Bigram frequency	.256	−.126	.007	.905	−.120
Concreteness	−.084	.015	.029	−.114	.987

Absolute loadings greater than .4 are highlighted via lighter shading, and absolute loadings greater .7 via darker shading.

2.2.2 Participants

A total of 116 participants were recruited (32 male, 84 female). Participants were aged 18-53 years and the mean age was 22.94 years (SD=5.89). Participants were from 35 different countries of origin (a full list can be found in Appendix C). Ninety-two of the participants classified themselves as bilingual and 24 were a control group of monolingual English speakers. The bilingual participants were from a variety of different language backgrounds, including both early and late bilinguals. The average length of stay of the bilingual participants in an English-speaking country (cumulative) was 5 years 8 months (SD= 93.21 months), ranging from 4 months to 47 years. All bilingual participants had started learning English between 0 and 12 years. Mode age for English exposure for the bilingual participants (starting point for English acquisition) was 0 years, with 21 participants having learnt English from birth. Mean age for English acquisition was 5 years. Participants were paid or awarded course credits for their participation.

2.2.3 Apparatus

An SR EyeLink II head-mounted eye-tracker was used to record pupil size at 250 Hz data sampling rate. Only the participant's dominant eye was tracked and viewing of the stimuli was binocular. The display screen resolution was 1024*768 pixels and refresh rate 100Hz.

2.2.4 Procedure

Participants were first asked to fill out the language questionnaire. After completing the questionnaire, participants were instructed to the Miles test to determine their eye-dominance, seated in front of a computer screen and the eye-tracker was set up. Calibrations were conducted after the first 10 trials and subsequently every 15-40 trials. The experiment had 240 trials in English (stimuli described above in more detail), presented in a randomised order. The first ten trials were all neutral distractor items to allow the participant a practice block. Each trial consisted of a fixation dot, then a mask of X's for 500 ms, the word presentation (the length of presentation was a function of the number of letters in the word (t=50ms + 26ms * length of word in characters), then the mask was shown again for 1900 ms, followed by a question mark. Participants were holding a game pad and were instructed to wait for the appearance of the question mark and then press the left key if they did not recognise the word and right key if they did, respectively. The eye-tracker recording for each trial started from the experimenter-initiated onset of the trial and ended to participants' trigger response making the typical length of a recording period for each trial 3000ms. After the eye-tracking experiment participants were asked to complete the LexTALE proficiency test (Lemhöfer & Broersma, 2012) on a computer.

2.3 Results

2.3.1 LexTALE

Participants' LexTALE English proficiency test results were recorded as one score, as returned by the LexTALE website. This score shows the percentage of correct responses, corrected for the unequal proportion of words and non-words in the test by averaging the percentages correct for these two item types. It is calculated as follows: (number of words correct/40*100) + (number of nonwords correct/20*100) / 2 (see Lemhöfer & Broersma, 2012 for more information). The mean LexTALE score across all participants was 84.8% (SD=11.9%). Table 3 (below) summarises the means, standard deviations and score ranges between the bilingual and monolingual English-speaking participants.

Table 3 LexTALE scores

Group	Mean	SD	Min score	Max score
Bilinguals	83.2%	12.1%	47.5%	100%
Monolinguals	90.9%	9.0%	72.5%	100%

2.3.2 Language History Questionnaire data and missing values

The full coding scheme for the Language History Questionnaire data can be found in Appendix A.

Question 10 and its subsections are designed to measure age of acquisition across different contexts. This produced a large number of missing values, for example, for participants who have not learnt English at home. Hence, responses to question 10 were transformed into "percentage of lifetime"-scores (a more thorough explanation of the questionnaire coding can be found in Appendix A).

Other missing values, due to participants' skipping questions, were replaced by column means (missing values accounted for 2% of the observations when question 10 was excluded). Missing values were almost exclusively in the responses to subsections of question 16, measuring language preference across different situations as a difference score between the languages. These missing values are explained by participants who are not currently working (23 missing values) or do not have a partner (29 missing values). These missing values were replaced with a column mean.

2.3.3 PCA

A principal components analysis with Varimax rotation was performed on the bilinguals' language history questionnaire data combined with their LexTALE proficiency scores. The number of extracted components (eleven principal components) was chosen to capture at least 80% of the variation in the data, based on the cumulative variance.

2.3.4 PCA bootstrapping

Non-parametric percentile bootstrapping of 60 000 resamples was performed to determine 99.5% CIs per factor loading. This was done in order to assess primarily the *robustness* of the obtained component loadings. Indeed, it turned out that some of the observed component loadings were strong, but not very robust against resampling, while others were moderate yet highly reliable. The reason for this is the degree of clustering in the observed data distributions (see Figure 1 below for some examples), which has implications for whether observed components/component loadings are generalizable to new participant samples.

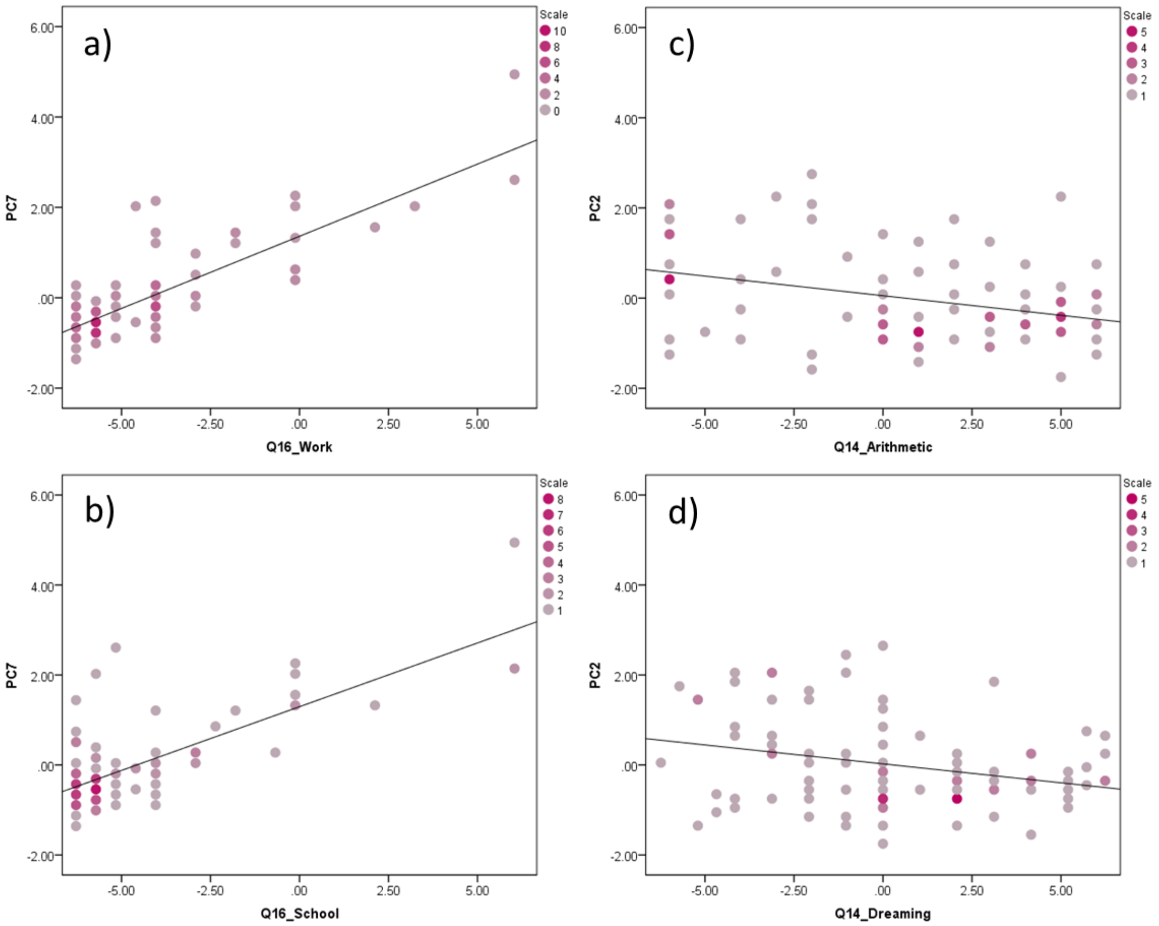

Figure 1 Strength vs. robustness of factor loadings.

Figure 1 show the robustness and strength of the factor loadings. The panels on the left show that both *L1 dominance at work* (a) and *L1 dominance at school* (b) load strongly positively on PC7 (principal component representing *L1 prevalence at work/school*). However, the relevant data are highly clustered (Note: different colour shades represent different numbers of observations in each X/Y-bin) because most participants reported that they typically use English at work or school (negative values on the x-axes). The factor loadings in panels (a) and (b) are therefore not very robust against resampling. In contrast, the panels on the right show weak negative correlations between *L1 dominance: arithmetic* (c) respectively *L1 dominance: dreaming* (d) and PC2 (principal component representing *early acquisition of English / English at home*), indicating that *early acquisition of English / English at home* is weakly associated with a lower reported dominance of L1 in arithmetic tasks or when dreaming, respectively. Since the data in (c) and (d) are more evenly spread, these factor loadings – albeit weak – are rather robust against resampling.

Table 4 Varimax-rotated item loadings (and their robustness over 60k resamples) for the 11 extracted principal components

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Q4Q7: Length of stay	-0.421	0.081	-0.551	0.022	0.203	0.246	-0.170	-0.121	-0.096	0.146	0.225
Q8: AOE	0.206	-0.798	0.163	0.064	-0.053	-0.232	0.113	0.107	0.016	-0.046	-0.146
Q8: Years learning	-0.488	0.311	-0.100	0.207	0.020	0.496	0.033	0.086	0.159	0.281	0.170
Q9: School instruction	-0.427	0.210	-0.149	-0.047	0.589	0.078	0.053	-0.371	0.046	-0.105	-0.164
Q10: English at home %	-0.157	0.829	0.009	0.058	0.119	0.062	-0.009	0.030	0.006	-0.228	0.077
Q10: School %	0.064	-0.005	-0.040	-0.143	0.123	0.900	-0.070	-0.019	0.008	-0.015	-0.019
Q10: Immigration %	-0.050	-0.117	-0.275	0.378	0.686	0.072	-0.089	0.047	-0.056	0.065	0.350
Q10: Informal %	-0.145	0.423	-0.204	0.049	0.580	0.353	-0.038	0.219	-0.122	-0.073	0.022
Q10: Games %	0.201	-0.112	0.088	-0.078	0.010	0.010	0.009	0.900	0.011	0.028	-0.053
Q11: Hours	0.004	-0.103	0.001	0.868	0.138	-0.114	-0.017	0.048	0.139	-0.101	-0.045
Q12: Hours	-0.166	0.238	-0.086	0.801	-0.011	-0.003	0.052	-0.174	-0.208	0.139	0.008
Q13: Order	-0.159	0.770	-0.247	0.070	-0.090	-0.253	0.063	-0.152	-0.134	0.085	0.016
Q14: Thinking	0.755	-0.125	0.261	-0.019	-0.027	0.037	0.002	0.122	0.212	0.080	-0.004
Q14: Talking	0.778	-0.098	0.163	-0.033	0.099	-0.089	0.068	0.081	0.115	0.268	-0.173
Q14: Swearing	0.420	-0.051	0.181	0.009	-0.026	0.035	0.162	0.037	0.023	0.784	-0.114
Q14: Emotions	0.777	-0.139	0.056	-0.029	-0.098	-0.041	0.121	0.171	0.150	0.248	-0.116
Q14: Dreaming	0.793	-0.275	0.179	-0.133	-0.167	-0.058	0.052	0.056	0.176	0.021	0.037
Q14: Arithmetic	0.280	-0.348	0.736	0.003	-0.195	-0.014	-0.008	0.118	0.029	0.310	0.054
Q14: Numbers	0.427	-0.293	0.741	-0.115	-0.161	-0.016	0.011	0.000	-0.104	0.069	-0.025
Q15: Dominance	-0.670	0.279	-0.174	-0.008	0.240	-0.240	-0.228	0.042	0.011	0.081	0.018
Q16: Family	0.480	-0.594	0.247	-0.043	-0.153	0.008	-0.047	0.093	-0.016	-0.124	-0.197
Q16: Friends	0.472	0.009	0.383	-0.159	-0.078	0.025	0.273	0.244	0.440	0.123	-0.040
Q16: Partner	0.332	-0.077	-0.044	0.009	-0.041	0.030	0.083	-0.029	0.835	0.009	-0.140
Q16: Work	0.070	-0.227	-0.119	0.074	-0.150	-0.010	0.809	0.120	0.005	0.192	-0.051
Q16: School	0.193	0.238	0.234	-0.056	0.127	-0.094	0.751	-0.139	0.144	-0.056	-0.001
LexTALE	-0.148	0.240	-0.031	-0.044	0.072	0.008	-0.037	-0.045	-0.131	-0.094	0.876

modest (|loading| < .7) but robust (*p < .005)

strong (|loading| > .7) and robust (*p < .005)

strong (|loading| > .7) but not robust (*p > .005)

2.3.5 Interpretation of the Loading Matrix

Component 1 - L1 Dominance: Self. Variance explained: 32.9%. A positive score on this component means stronger L1 dominance in self-related domains, such as thinking and talking to self. The component has strong positive loadings from L1 dominance in Thinking/Talking to Self/Expressing Emotions/Dreaming; moderate positive loadings from L1 dominance in communication with Family/Friends and L1 dominance in memorizing numbers. It also has moderate negative loadings from Length of Stay in English-speaking country and English dominance (percentage across the different situations as captured by the questionnaire)

Component 2 - Early English Acquisition / English at Home. Variance explained: 8.53%. A positive score on this component means earlier English acquisition and more English use at home. The component has strong positive loadings from lifetime percentage of English at Home and having acquired English first before other languages. The component also has strong negative loading from Age of acquiring English (AOE); moderate negative loadings from L1 dominance in Dreaming, L1 dominance in communicating with Family, and L1 dominance in arithmetic tasks.

Component 3 - L1 Dominance: Maths. Variance explained: 6.94%. A positive score on this component means higher dominance of L1 in mathematical tasks. The component has strong positive loadings from L1 dominance in arithmetic and L1 dominance in remembering numbers, and moderate negative loading from length of stay in an English-speaking country.

Component 4 - Current English use per day. Variance explained: 6.84%. A positive score on this component means more daily English use, as estimated by the participant. The component has strong positive loadings from hours of English per day across different contexts (such as chatting online or watching TV), and hours of English per day with different people (such as with family or friends)

Component 5 - Lifetime spent in English Environment. Variance explained: 4.95%. A positive score on the component means a higher percentage of the participant’s lifetime spent in English-speaking environment. The component has moderate positive loadings from lifetime percentage since immigration to English-speaking country and lifetime percentage of informal English use and moderate (but unreliable) positive loading from English School Instruction (i.e. whether school classes were given in English).

Component 6 - English at School. Variance explained: 4.37%. A positive score on this component means a higher lifetime percentage of English schooling. This is mostly driven by strong and robust positive loading from lifetime percentage of English at school, plus moderate (but unreliable) loading from years spent learning English.

Component 7 - L1 Dominance: School/Work. Variance explained: 3.81%. A positive score on this component means higher dominance of L1 at work/school. However, the relevant loadings are not very robust (see Figure 1).

Component 8 - Lifetime English Online Gaming. Variance explained: 3.64%. A positive score on this component means higher percentage lifetime since acquiring English via online gaming.

Component 9 - L1 Dominance: Partner. Variance explained: 3.30%. A positive score on this component means higher L1 dominance in communicating with partner. This component has moderate (yet unreliable) loading from L1 dominance in communicating with friends as well.

Component 10 - L1 Dominance: Swearing. Variance explained: 2.91%. A more positive score means higher L1 dominance for swearing.

Component 11: LexTALE Proficiency. Variance explained: 2.57%. A more positive score means higher English proficiency according to LexTALE. Interestingly, LexTALE loaded on ‘its own’ component, suggesting that English proficiency, as measured by this test, it is not very strongly related to any of the language background variables measured in our sample of bilinguals.

2.3.6 Button press: Between-group comparisons

The mean accuracy of the word recognition task (button press after each trial) was ~92%. The bilingual group's mean accuracy was ~93% and the monolingual group's mean accuracy was ~90%. The bilingual group's recognition accuracy ranged from ~48% to 100% and the monolingual group ranged from ~66% to ~99% recognition accuracy.

Word recognition probability was analysed in the statistical software R, using Generalized Linear Mixed Effect Models (GLMEM), as implemented in the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). The model was specified a binary logistic model family (predicting "word recognised" button responses) and was fitted by maximum likelihood (Laplace Approximation). Full syntax of all the models used can be found in Appendix D.

The fixed effect structure of the model included word type (coded as 0=low arousing, 1=high arousing), participant group (0=monolingual, 1=bilingual), the stimulus-related lexical covariates (lexical frequency, bigram frequency, concreteness and length & orthographic neighbours) and LexTALE proficiency scores. It should be noted that the Principal Component of Valence and Dominance was not included in these models (or any others in this thesis), as valence and arousal have a U-shaped relationship, which would interfere with model fit (Kuperman et al., 2014). Both word type and participant group were deviation coded, and the lexical covariates and LexTALE scores were mean-centered. Further, 2-way interactions between the stimulus-related variables and participant group were entered as fixed effects, as well as the word type:group 2-way interaction. The participant specific PCA scores were not included in the button response models, as this task was simply included to monitor participant language proficiency, and we did not have this information about the monolingual participants. Button press in the bilingual participants only, as predicted from the person-specific PCA components can be found in section "Analysis for bilingual participants only" below.

Random effect structure of the model included by subjects and by-item random intercepts, as well as by-subject random slopes on each of the item related predictor variables (all of which were within-subjects but between-items), and by-item random slopes on each of the participant-related predictor variables (all of which were between subjects but within-items). This approach appropriately accounts for repeated-measures dependencies in the design, and hence the results are generalisable across subjects and items. Random correlations were included. Table 5 below summarises the predictors, interaction terms and their significance.

Table 5 Button response model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
Intercept	4.61	0.21	22.22	<0.001***	4.20	5.01
Word Type	0.44	0.22	2.05	0.04*	0.02	0.87
Group	0.45	0.46	0.98	0.33	-0.45	1.35
Lexical frequency	1.23	0.12	9.90	<0.001***	0.99	1.48
Bigram frequency	-0.03	0.11	-0.26	0.80	-0.25	0.19
Concreteness	0.16	0.11	1.50	0.13	-0.05	0.37
Length& Orthographic neighbours	-0.39	0.13	-3.11	0.002***	-0.63	-0.14
LexTALE	-0.002	0.01	-0.23	0.82	-0.03	0.02
Group: Word Type	-0.10	0.20	-0.49	0.63	-0.49	0.29
Group: Lexical frequency	-0.10	0.15	-0.64	0.52	-0.39	0.20
Group: Bigram frequency	0.007	0.11	0.06	0.95	-0.20	0.22
Group:Concreteness	-0.006	0.09	-0.06	0.95	-0.19	0.18
Group:Length& orthographic neighbours	0.03	0.15	0.18	0.86	-0.26	0.31
Group:LexTALE	0.02	0.04	0.49	0.62	-0.06	0.09

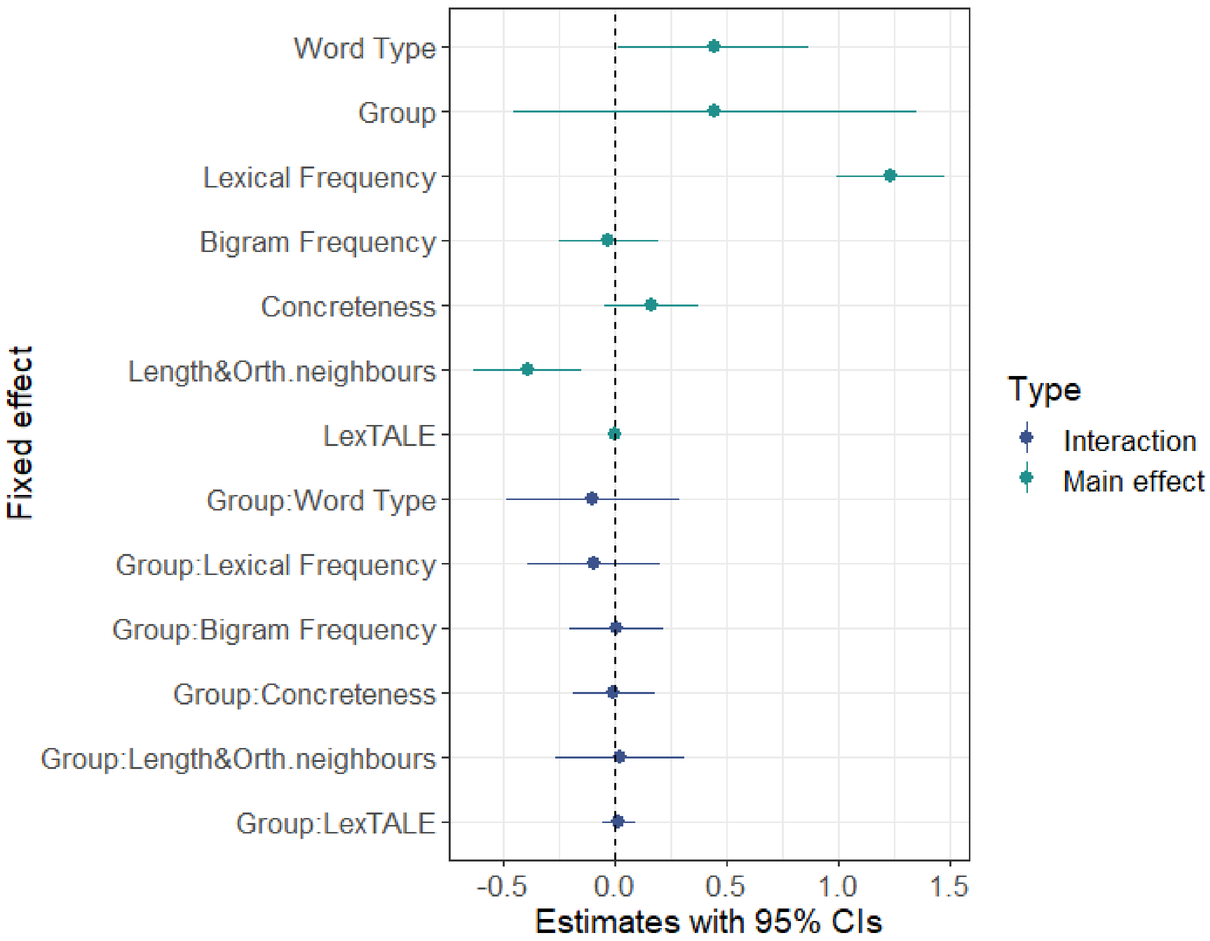

Figure 2 Button response task model estimates with 95% Confidence Intervals

Table 5 and blobbogram (Figure 2) above indicate that there was no significant difference between the monolingual and bilingual group in terms of word recognition likelihood, nor did any of the item-specific variables interact with the group variable in predicting word recognition. This can be taken as further evidence of the bilingual group’s high overall level in English proficiency. As expected, longer words with more orthographic neighbours were less likely to be recognised (negative coefficient), words with higher frequency were recognised more likely than lower-frequency words, and high arousing words were recognised more likely than low arousing words (positive coefficients). Interestingly, LexTALE (English proficiency) scores did not predict word recognition likelihood.

2.3.7 Pupil data pre-processing

The initial sampling rate of data was 250Hz. This was down-sampled to 10Hz. For each participant and each trial, pupil size and eye-position data were extracted for a time period starting from 150 ms before the onset of the critical word presentation and ending at 1900 ms after the onset of the word presentation. The baseline time bin before word onset was 150 ms, and the subsequent 20 time bins were 100 ms each (hence the data consisted of twenty one time bins per trial in total). For each of the time bins per trial the average pupil size (in numbers of pixels per eye-camera sample) and eye-position (average X- and Y-coordinates in pixels) from the eye-tracker output were extracted. This was done using SR-Research Data Viewer software (Version 2.1.1). The pupil size data were converted into decimal logarithms for further analysis. To remove noise resulting from small eye-movements or drifts, multiple regression analysis with X- and Y-position of the eye as orthogonal predictors of log pupil size was performed. This was done separately for each participant. All the data available (21 [time bins] × 240 [experimental and neutral distractor trials] = 5040 data points per participant) were used for this. The predicted log pupil sizes from these regression analyses were then subtracted from the actual log pupil sizes per trial and time bin, i.e. subsequent analyses were all based on position-adjusted residual log pupil size data.

The baseline log pupil size (at t = -150) was then subtracted from all pupil sizes per trials and pupil data were converted back into 'proportion baseline' pupil sizes (10^(pupil size in timestamp - baseline)). For the final analysis, the pupil data were converted into 'area under the curve' estimates from 600 - 1900 ms after word onset (see Toivo & Scheepers, 2019 for further discussion). Blink-related gaps (typically 100-200 ms in duration) were replaced with linear interpolation estimates over adjacent time bins. A total of 114 trials (out of 27 840) from 30 participants were removed altogether as they contained a large number of blinks and/or fixations outside the screen area, resulting in missing data for five or more consecutive time bins per trial.

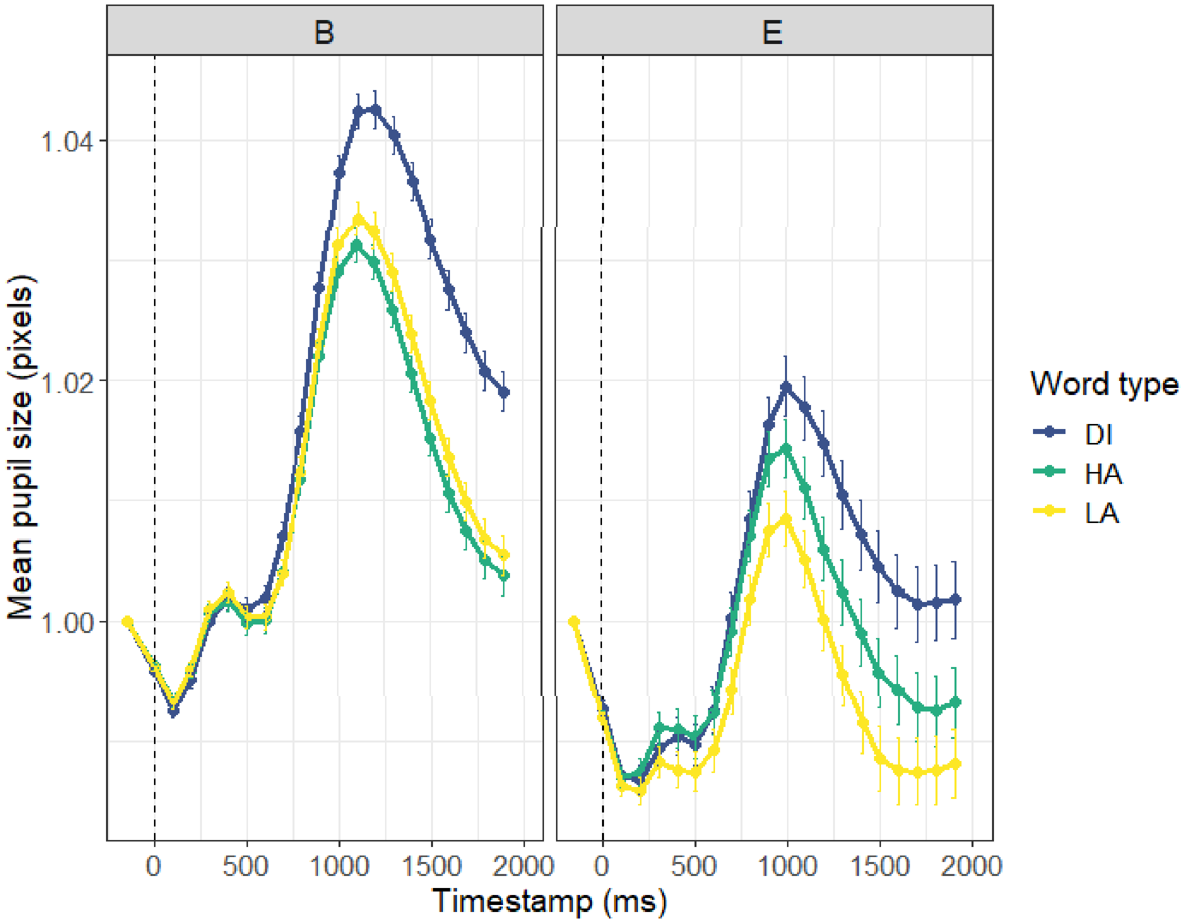

Figure 3 Average pupil size across time with standard error, split by group and word type

Figure 3 shows pupil size changes over time relative to the by-trial baseline (as explained earlier, the baseline time bin always assumed a pupil-size value of 1). Different word types are indicated by different colour shades (Green = HA, high arousing words, yellow=LA, low arousing words and blue=DI, neutral distractor words). The graph on the left shows data for the bilingual participant group (N=92) and the graph on the right the corresponding data from the monolingual English group (N=24). As expected, bilinguals seemed to have a larger overall pupil response, regardless of word type. This is possibly due to the cognitive effort effect (e.g., Schmidtke, 2014). Most importantly, consistent with Toivo and Scheepers (2019), English monolingual speakers showed a clear word type effect such that pupil responses to HA words were more positive than to LA words; by contrast, the HA-LA contrast was much smaller for bilingual participants and actually in the opposite direction.

Interestingly, the neutral words seem to have elicited strong pupillary responses across both participant groups. This may be due to surprise (pupil size is known to be sensitive to inconsistencies in stimuli, see: Scheepers et al., 2013).

2.3.8 Pupil area prediction models

Participants' pupillary responses were predicted using GLMEMs (lme4 package, as mentioned above). The first model was run to examine the differences in pupillary reactions between the groups and as a response to the word characteristics variables. The model was fitted by maximum likelihood (Laplace Approximation) and the Gamma distribution combined with identity link was used, given that the pupil area data were strongly positively skewed, as can be seen in Figure 4 below.

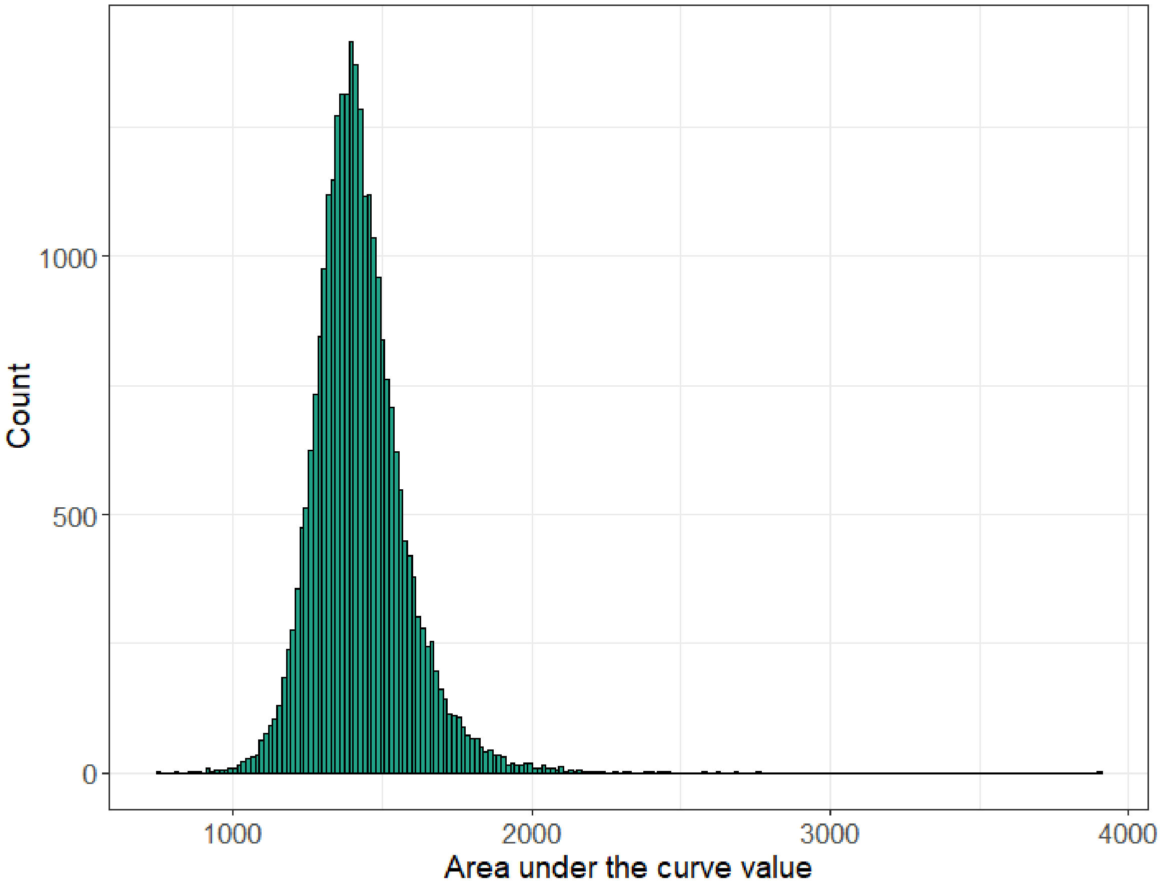

Figure 4 Distribution of the area under the curve values

Neutral words were excluded from this part of the analysis in the effort of keeping the models concise. Trials where the participants did not recognise the word were also excluded from the analysis. The fixed effects structure included participant Group and Word type (both deviation coded), LexTALE scores, and the following lexical covariates: Length & Orthographic neighbours, Lexical frequency, Bigram frequency, and Concreteness. LexTALE scores and the lexical covariates were all mean-centered. The random effects structure, as above, included by-item random intercepts and slopes for all the participant-related predictors (Group) and by-subject random intercepts and slopes for all the item-related predictors (Lexical covariates). The full model structure can be found in Appendix D.

Table 6 Pupil response model summary (between groups)

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
Intercept	1420.21	2.80	507.89	<0.001***	1414.73	1425.69
Word Type	0.11	3.15	0.04	0.97	-6.06	6.29
Group	26.42	7.14	3.70	<0.001***	12.42	40.41
Lexical frequency	-11.99	1.80	-6.67	<0.001***	-15.51	-8.49
Bigram frequency	-1.73	1.64	-1.05	0.29	-4.95	1.49
Concreteness	0.68	1.73	0.39	0.70	-2.72	4.07
Length & Orthographic neighbours	-2.02	1.76	-1.15	0.25	-5.47	1.42
LexTALE	-0.07	0.21	-0.31	0.75	-0.49	0.35
Group: Word Type	-11.72	5.88	-1.99	0.046*	-23.24	-0.19
Group:Lexical frequency	-8.16	3.68	-2.22	0.03*	-15.36	-0.95
Group:Bigram frequency	0.06	3.19	0.018	0.99	-6.20	6.31
Group:Concreteness	-1.20	3.41	-0.35	0.72	-7.90	5.49
Group:Length& orthographic neighbours	-0.95	3.52	-0.27	0.78	-7.86	5.95
Group:LexTALE	-0.21	0.63	-0.34	0.73	-1.45	1.02

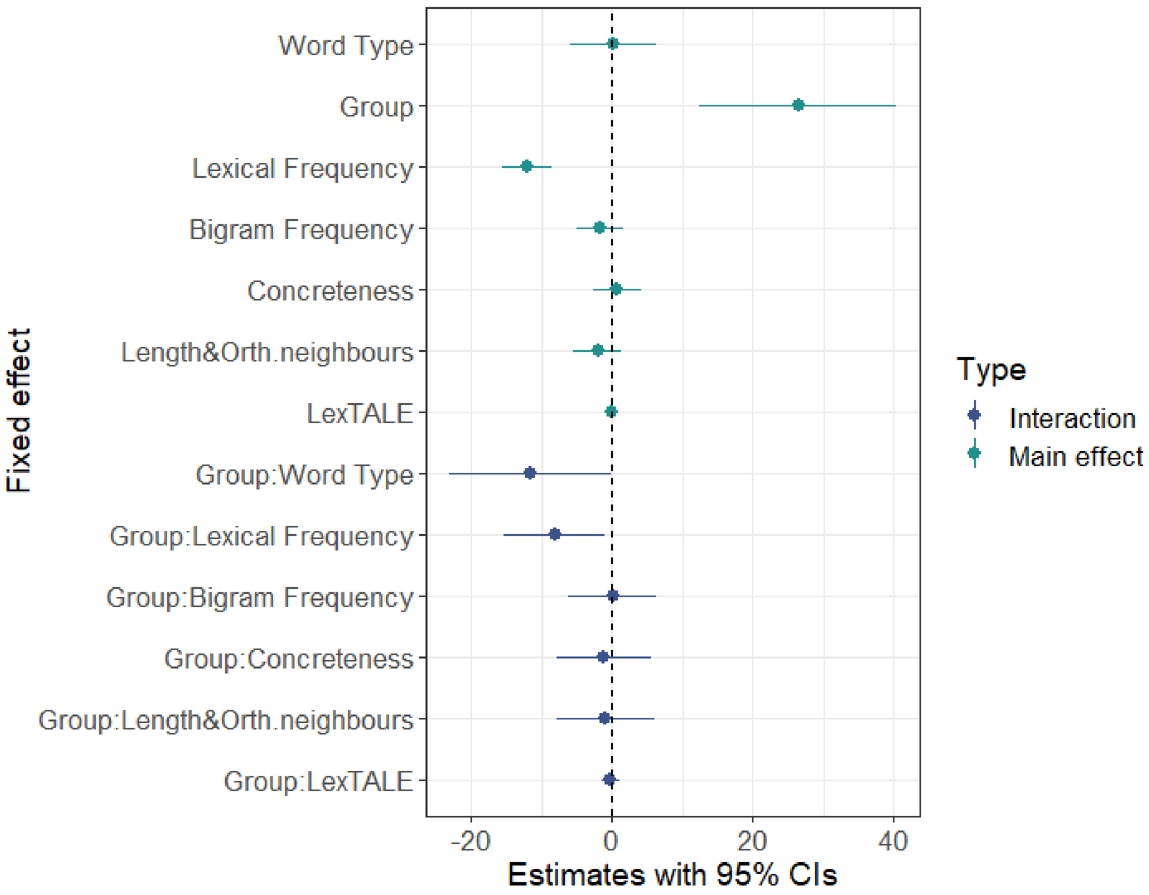

Figure 5 Pupil response estimates with 95% Confidence Intervals

The above table (Table 6) and blobbogram (Figure 5) summarise the pupil area model results comparing bilinguals and monolinguals' pupillary responses. The main effect of group was significant, suggesting bilinguals overall had a larger pupillary response. Words with smaller lexical frequency elicited stronger pupillary response in both groups. The crucial interaction of word type and group is also significant (although marginally). When decomposed into simple effects (using dummy coding of the group variable), the results indicate no reliable word type simple effect for the bilingual group (β = -1.97, SE = 3.06, t = -0.6, p = 0.52) but a clear HA-LA contrast for the English monolingual group (β = 8.09, SE = 1.71, t = 4.7, p < 0.001). This supports the idea of reduced emotional resonance – the emotionally arousing words in participants' L2 did not yield stronger pupil responses in comparison to low emotional arousal words. The interaction of Lexical frequency and Group is also significant, and when decomposed into simple effects, higher lexical frequency predicts higher pupillary response in bilinguals (β=8.05, SE=3.67, t=2.19, p=0.03), whereas in monolinguals lower frequency words predict higher pupillary response (ß=-8.01, SE=1.81, t=-4.44, p<0.001)

2.3.9 Analysis for bilingual participants only

The purpose of the following analyses was to find out how the 11 person-specific principal components affected word recognition performance and pupil-size changes in the sample of the bilingual participant group (N=92). Again, occurrences of positive word recognition responses at the end of each trial were analysed in terms of a binary logistic approach and pupil-change area under the curve data were modelled via Gamma regression.

All predictors (*word type*, the 11 subject-specific principal components, and the 4 item-specific control components) were mean-centered and entered into the models, which not only included the main effects of each of the 16 predictors, but also 2-way interactions between *word type* and each of the 11 subject-specific principal components. In the analyses of the pupil size data, these interactions were of major theoretical interest, as they point to potential subject-specific variables that may enhance or reduce emotional resonance in bilinguals (as revealed in more dilated pupils when processing HA words).

One problem that became obvious after a few test runs was that because of the large number of fixed effects considered (16 main effects plus 11 interactions), it was not possible to test all these effects simultaneously within a model that would also take design-specific subject- and item-related measurement dependencies into account (recall that all item-related variables were within-subjects and all subject-related variables within-items). That is, complete models with design-appropriate ‘maximal’ random effect structures resulted in severe convergence problems.

Because of this, the following approach was taken. In a first set of analyses, all effects were considered simultaneously, but without consideration of any random effects. This approach was based on GLM and employed the following model structure:

```
DV ~ wtype + # main effect of word type
     pc1 + pc2 + ... + pc11 + # main effects of participant components
     wtype:pc1 + wtype:pc2 + ... + wtype:pc11 + # 2-way interactions
     len_ortho_n + lexfreq + bgfreq + concrete # item covariates
```

The GLMs should give a rough idea of effect patterns when all fixed effect predictors are considered simultaneously, but without guaranteeing generalizability across subjects and items, because this approach does not account for measurement dependencies within the data.

The second set of analyses was based on Generalised Linear Mixed Effect Models with design-appropriate maximal random effect structures, but considering only one of the 11 subject-specific components at a time (thus, for every DV, 11 such models were tested). The GLMEM model structures were as follows:

```
DV ~ wtype + pc? + wtype:pc? + # ? ranging from 1 to 11
     len_ortho_n + lexfreq + bgfreq + concrete # item covariates
     # by-subject random effects (including correlations):
     (1 + wtype + len_ortho_n + lexfreq + bgfreq + concrete | subj) +
     # by-item random effects (including correlations):
     (1 + pc? | item)
```

Word type and all item-related covariates received by-subject random slopes in the GLMEMs (these variables varied within-subjects but between-items), whereas the participants-specific principal components were accompanied with by-item random slopes (these components varied between-subjects but within-items).

As will become clear from the following results sections, the GLM and GLMEM approaches yielded very similar effect parameter estimates (both in magnitude and direction), but many of the effects suggested by the GLMs were not generalizable across participants and items (as revealed by the fact that the GLMEMs were generally more conservative). Because of the latter, the GLMEM results are given higher inferential credibility for theoretical interpretation.

2.3.10 Word Recognition Responses

Table 7 below summarises the estimates, SEs, Z, and p-values on the left are from an omnibus GLM considering all predictor variables simultaneously, but not taking subject- or item dependencies into account. The corresponding statistics on the right are from Generalized Linear Mixed Models with maximal (by-subject/item) random effect structures, but considering only *one* participant-specific principal component at a time. Since there were 11 such GLMEMs, each also including the main effect of *word type* (top row) and the item-specific control variables (bottom four rows), statistics for the latter are given as ranges (min, max) across these 11 GLMEM analyses. Significant effect parameters (p ≤ .05) are shaded.

Table 7 Binary logistic regression results for word recognition responses at the end of each trial

	GLM (omnibus)				GLMEM (one-by-one)			
Effect	*Estim*	*SE*	*Z*	*P*	*Estim*	*SE*	*Z*	*P*
wtype[1]	0.257	0.082	3.148	0.002	0.38, 0.41	0.22, 0.23	1.68, 1.83	0.07, 0.09
PC1	-0.085	0.040	-2.156	0.031	-0.041	0.187	-0.222	0.825
PC2	-0.110	0.038	-2.904	0.004	-0.174	0.177	-0.986	0.324
PC3	-0.315	0.040	-7.923	0.000	-0.431	0.184	-2.339	0.019
PC4	0.067	0.039	1.715	0.086	0.163	0.185	0.880	0.379
PC5	-0.349	0.035	-9.928	0.000	-0.346	0.175	-1.981	0.048
PC6	0.113	0.039	2.883	0.004	0.232	0.179	1.299	0.194
PC7	0.128	0.042	3.029	0.002	0.104	0.182	0.571	0.568
PC8	-0.155	0.035	-4.459	0.000	-0.234	0.173	-1.350	0.177
PC9	0.174	0.038	4.620	0.000	0.285	0.176	1.619	0.105
PC10	-0.117	0.040	-2.935	0.003	-0.079	0.185	-0.425	0.671
PC11	0.000	0.037	0.005	0.996	0.017	0.179	0.092	0.926
wtype:PC1	0.137	0.079	1.734	0.083	0.127	0.099	1.284	0.199
wtype:PC2	-0.020	0.076	-0.266	0.791	-0.019	0.097	-0.197	0.844
wtype:PC3	-0.233	0.079	-2.938	0.003	-0.292	0.104	-2.803	0.005
wtype:PC4	-0.030	0.078	-0.380	0.704	-0.137	0.111	-1.236	0.217
wtype:PC5	-0.032	0.070	-0.461	0.645	0.103	0.098	1.051	0.293
wtype:PC6	0.003	0.078	0.035	0.972	0.059	0.100	0.587	0.558
wtype:PC7	0.078	0.084	0.931	0.352	0.070	0.106	0.659	0.510
wtype:PC8	-0.061	0.069	-0.873	0.383	-0.092	0.090	-1.023	0.306
wtype:PC9	0.086	0.075	1.137	0.256	0.089	0.096	0.922	0.357
wtype:PC10	-0.034	0.080	-0.425	0.671	0.055	0.112	0.487	0.626
wtype:PC11	-0.014	0.075	-0.193	0.847	0.019	0.095	0.203	0.839
Length & orth. neighbours	-0.299	0.044	-6.816	0.000	-0.36, -0.31	0.13, 0.13	-2.70, -2.32	0.01, 0.02
Lexical frequency	0.834	0.041	20.211	0.000	1.15, 1.20	0.13, 0.13	8.67, 9.14	0.00, 0.00
Bigram frequency	-0.005	0.041	-0.117	0.907	0.03, 0.07	0.12, 0.12	0.27, 0.57	0.56, 0.79
Concreteness	0.098	0.036	2.746	0.006	0.10, 0.14	0.11, 0.11	0.93, 1.23	0.22, 0.35

[1] dummy-coded (0 = LA word, 1 = HA word) and then mean-centred; positive estimates indicate better recognition of HA than LA words

Table 7 above shows the binary logistic modelling results for affirmative word recognition responses at the end of each trial.

There was a marginal main effect of *word type* (suggesting that HA words were more likely to be recognised than LA words) which mirrors the findings from the previous monolingual-bilingual comparisons. Further, there were negative main effects of *PC3* and *PC5*. The former appears plausible: Bilingual participants with a *stronger L1 preference for mathematical tasks* (*PC3*) were less likely to report that they have recognised the words presented in English. The negative influence of *PC5*, however, is probably less straightforward to interpret: Bilinguals that have spent more of their *lifetime in an English-speaking environment* (*PC5*) appeared to be less likely to report that they have recognised the English words (It should be noted that this effect was just below the significance threshold and actually not very strong).

The only reliable interaction effect was between *PC3* and *word type*: Bilingual participants with a stronger *L1 preference for mathematical tasks* (*PC3*) showed a weaker effect of *word type* on word recognition performance, which makes sense considering that *PC3* appeared to be associated with reduced word recognition in general.

Finally, there was a reliably negative influence of *length and orthogonal neighbours* on word recognition (longer words with fewer orthographic neighbours were less likely to be recognised) and a reliably positive influence of *lexical frequency* (more frequent words were generally more likely to be recognized).

Indeed, considering absolute parameter magnitudes, it appears that *lexical frequency* had by far the strongest effect on word recognition performance, whereas all the other effects were rather small at best (log odds < 0.5, suggesting Cohen's *ds* < 0.3).

Perhaps also worth noting is that bilinguals' *proficiency in English* (*PC11*) did not appear to have any influence on word recognition performance, echoing the previous results from the monolingual-bilingual comparisons.

2.3.11 Pupil-Change Area Under the Curve Data

Table 8 below summarises estimates, SEs, Z, and p-values on the left are from an omnibus GLM considering all predictor variables simultaneously, but not taking subject- or item dependencies into account. The corresponding statistics on the right are from Generalized Linear Mixed Effects Models with maximal (by-subject/item) random effect structures but considering only *one* participant-specific principal component at a time. Since there were 11 such GLMEMs, each also including the main effect of *word type* (top row) and the item-specific control variables (bottom four rows), statistics for the latter are given as ranges (min, max) across these 11 GLMEM analyses. Significant effect parameters ($p \leq .05$) are shaded.

Table 8 Gamma regression results for the pupil-change area under the curve data (all trials included).

	GLM (omnibus)				GLMEM (one-by-one)			
Effect	*Estimate*	*SE*	*Z*	*P*	*Estimate*	*SE*	*t*	*P*
wtype[1]	-1.793	2.463	-0.728	0.467	-2.33, -2.07	1.95, 3.64	-1.14, -0.59	0.25, 0.55
PC1	0.233	1.244	0.187	0.852	-1.536	4.060	-0.379	0.705
PC2	-9.222	1.226	-7.520	0.000	-8.012	2.828	-2.833	0.005
PC3	2.001	1.238	1.616	0.106	-0.300	2.289	-0.131	0.896
PC4	-7.168	1.226	-5.848	0.000	-4.600	2.822	-1.631	0.103
PC5	1.144	1.231	0.929	0.353	2.159	2.822	0.765	0.444
PC6	2.183	1.233	1.769	0.077	1.516	2.170	0.698	0.485
PC7	-7.064	1.225	-5.766	0.000	-6.931	2.814	-2.463	0.014
PC8	-7.088	1.232	-5.753	0.000	-6.201	2.816	-2.202	0.028
PC9	9.493	1.239	7.662	0.000	7.743	1.791	4.326	0.000
PC10	-0.784	1.243	-0.631	0.528	-0.913	2.618	-0.349	0.727
PC11	-2.089	1.239	-1.686	0.092	-0.415	4.051	-0.103	0.918
wtype:PC1	-0.479	2.487	-0.193	0.847	0.060	1.984	0.030	0.976
wtype:PC2	1.746	2.453	0.712	0.477	1.413	2.184	0.647	0.518
wtype:PC3	-3.605	2.476	-1.456	0.145	-2.741	2.497	-1.099	0.272
wtype:PC4	1.839	2.452	0.750	0.453	1.373	3.044	0.451	0.652
wtype:PC5	1.013	2.462	0.411	0.681	0.060	2.936	0.020	0.984
wtype:PC6	-0.078	2.467	-0.032	0.975	-0.275	2.149	-0.128	0.898
wtype:PC7	-1.440	2.450	-0.588	0.557	-1.553	2.755	-0.564	0.573
wtype:PC8	-0.022	2.464	-0.009	0.993	0.011	3.042	0.004	0.997
wtype:PC9	2.211	2.478	0.892	0.372	2.445	2.033	1.203	0.229
wtype:PC10	-4.878	2.486	-1.962	0.050	-4.809	2.032	-2.366	0.018
wtype:PC11	-0.359	2.478	-0.145	0.885	-0.164	2.627	-0.063	0.950
Length &orth. neighbours	-1.787	1.225	-1.459	0.145	-1.90, -1.16	1.61, 2.61	-1.12, -0.61	0.26, 0.54
Lexical frequency	-12.52	1.201	-10.43	0.000	-13.1, -12.5	1.91, 2.18	-6.80, -5.87	0.00, 0.00
Bigram frequency	-1.605	1.205	-1.332	0.183	-1.76, -1.18	1.74, 2.82	-0.99, -0.56	0.32, 0.58
Concreteness	0.717	1.246	0.575	0.565	0.13, 0.75	1.76, 2.14	0.07, 0.35	0.72, 0.94

[1] dummy-coded (0 = LA word, 1 = HA word) and then mean-centred; positive estimates indicate more dilated pupils for HA than LA words

As Table 8 shows, there were reliably negative main effects of *PC2* (*earlier English acquisition / more English at home* was associated with less dilated pupils in response to the English words), *PC7* (which probably should not be interpreted given the less robust nature of this component – see bootstrapping PCA results above), and *PC8* (more *lifetime spent on English on-line gaming* was, again, associated with less dilated pupils in response to the English words). *Lexical frequency* also had a strong negative effect (more frequent words were associated with less dilated pupils). Finally, *PC9* had a reasonably strong *positive* influence on pupil-size changes, meaning that a *stronger L1 dominance in communication with one's partner* was associated with more dilated pupils in response to the English words).

The only significant interaction was between *PC10* and *word type*, and its negative coefficient indicates that a greater *L1 dominance for swearing* (*PC10*) was associated with a more negative effect of *word type* on pupil-change area under the curve. This interaction effect appears to be the only (vague) pointer to a possible explanation for the previously established monolingual vs. bilingual contrast in pupillary responses to emotionally charged words, as all other components measured by the Language History Questionnaire did not show an interaction with *word type* in predicting bilinguals' pupil responses. There was no indication of a mediating influence of *English proficiency* (*PC11*).

2.3.12 Pupil-Change Area Under the Curve Data ('recognised' trials only)

Table 9 below summarises the estimates, SEs, Z, and p-values on the left are from an omnibus GLM considering all predictor variables simultaneously, but not taking subject- or item dependencies into account. The corresponding statistics on the right are from Generalized Linear Mixed Models with maximal (by-subject/item) random effect structures but considering only *one* participant-specific principal component at a time. Since there were 11 such GLMEMs, each also including the main effect of *word type* (top row) and the lexical covariate variables (bottom four rows), statistics for the latter are given as ranges (min, max) across these 11 GLMEM analyses. Significant effect parameters ($p \leq .05$) are shaded.

Table 9 Gamma regression results for the pupil size area under the curve data (only ‘recognised’ trials included).

	GLM (omnibus)				GLMEM (one-by-one)			
Effect	*Estim*	*SE*	*Z*	*P*	*Estim*	*SE*	*t*	*P*
wtype[1]	-1.746	2.517	-0.693	0.488	-3.00, -2.32	1.93, 11.47	-1.47, -0.25	0.14, 0.80
PC1	-0.229	1.273	-0.180	0.857	-1.716	2.878	-0.600	0.551
PC2	-8.888	1.249	-7.114	0.000	-7.100	2.159	-3.300	0.001
PC3	0.867	1.267	0.684	0.494	-0.036	2.875	0.000	0.990
PC4	-6.958	1.245	-5.591	0.000	-4.736	4.441	-1.100	0.286
PC5	0.606	1.274	0.476	0.634	1.592	10.376	0.150	0.878
PC6	2.169	1.253	1.731	0.083	1.779	2.387	0.700	0.456
PC7	-6.582	1.239	-5.313	0.000	-6.793	2.485	-2.700	0.006
PC8	-7.134	1.263	-5.650	0.000	-6.836	5.145	-1.300	0.184
PC9	9.714	1.267	7.668	0.000	8.104	2.844	2.800	0.004
PC10	-0.627	1.267	-0.495	0.621	-1.666	2.933	-0.600	0.570
PC11	-1.979	1.265	-1.564	0.118	-0.922	2.937	-0.300	0.754
wtype:PC1	0.093	2.545	0.036	0.971	0.686	2.943	0.200	0.816
wtype:PC2	1.507	2.499	0.603	0.547	1.493	2.968	0.500	0.615
wtype:PC3	-3.557	2.535	-1.403	0.161	-3.841	3.221	-1.200	0.233
wtype:PC4	1.450	2.489	0.583	0.560	1.118	3.198	0.300	0.727
wtype:PC5	0.007	2.549	0.003	0.998	-1.301	5.906	-0.220	0.826
wtype:PC6	0.228	2.506	0.091	0.928	0.284	1.918	0.100	0.883
wtype:PC7	-2.230	2.478	-0.900	0.368	-2.242	2.146	-1.000	0.296
wtype:PC8	0.301	2.525	0.119	0.905	0.777	2.429	0.300	0.749
wtype:PC9	3.172	2.534	1.252	0.211	2.917	2.916	1.000	0.317
wtype:PC10	-5.738	2.534	-2.265	0.024	-5.143	3.353	-1.500	0.125
wtype:PC11	-0.772	2.531	-0.305	0.760	-0.255	3.245	-0.100	0.937
Length & orth. neighbours	-2.731	1.243	-2.198	0.028	-2.37, -1.73	1.72, 12.4	-1.29, - 0.19	0.20, 0.85
Lexical frequency	-11.88	1.235	-9.614	0.000	-14.0, -13.3	1.90, 5.63	-7.23, -2.43	0.00, 0.02
Bigram frequency	-1.636	1.222	-1.338	0.181	-1.91, -1.19	1.73, 7.66	-1.08, -0.24	0.28, 0.81
Concreteness	0.777	1.272	0.611	0.541	-0.13, 0.49	1.90, 6.78	-0.06, 0.26	0.80, 0.97

[1] dummy-coded (0 = LA word, 1 = HA word) and then mean-centred; positive estimates indicate more dilated pupils for HA than LA words

Table 9 shows modelling results for pupil-change area under the curve data, this time only considering trials with affirmative word recognition responses (ca. 6% data loss compared to the previous analysis). As can be seen, results did not change much, but the main effect of *PC8* and the interaction between *PC10* and *word type* now failed to reach significance in the GLMM analysis. This could be due to a reduction in power as a result of fewer (and less balanced) observations when only trials with affirmative word recognition responses are considered.

2.4 Discussion

The aim of Experiment 1 was to replicate our pupillometry paradigm (Toivo & Scheepers, 2019) with new items, and to explore which factors in bilinguals' language background predict reduced emotional resonance in L2. Reduced emotional resonance is measured by the difference between pupillary response to high arousing and low arousing stimuli words (namely, no difference between word types indicates reduced emotional resonance of the language, since both emotional and not emotional words are processed in a similar manner).

Replicating the effect found in various skin-conductance studies (Caldwell-Harris et al., 2010; Harris, 2004; Harris et al., 2003; Harris et al., 2006), and our first pupil experiment comparing L1 and L2 of Finnish and German speakers (Toivo & Scheepers, 2019), we found reduced emotional resonance in our bilingual group, when comparing them to a monolingual group (see Figures 3 and 5). For the models measuring general pupillary responses transformed into area under the curve values, there was a significant interaction of Group and Word type. Decomposed into simple effects, the results indicate no reliable word type simple effect for the bilingual group, but a clear HA-LA contrast for the English monolingual group.

There was no reliable main effect of word type. However, and rather unexpectedly, the distractor words elicited the strongest pupillary response in both participant groups (it should be noted that this finding is descriptive, as neutral items were excluded from the GLMEMs). Without systematic testing we can only speculate, but it is possible this is due to an effect of surprise. Pupil responses have previously been found to be sensitive to surprise (Scheepers et al., 2013). Perhaps the participants detect pattern of both HA and LA words and start anticipating these semantic categories, which constitute two thirds of the stimuli, whilst the last third of neutral words are unexpected.

When having a closer look at the word type-participant group interaction (see Figure 3), it appears that in the bilingual group LA words elicit stronger pupil responses in comparison to HA words. This effect was not significant in the LMEMs, but is worth discussing briefly. This issue was also prevalent in the first pupillometry study where the same paradigm was used (Toivo & Scheepers, 2019). In the previous study, there was no suitable baseline condition, which means we cannot say with certainty whether HA words drive pupil sizes up, or LA words drive pupil sizes down in bilinguals. In similar vein, in the present findings we observed a main effect of participant group, showing that bilinguals overall had stronger pupillary response. Hence, from the present data set we cannot conclusively tell whether the LA effect is due to a drop in the HA responses, or a heightened response to the LA words. It is not possible to disambiguate the effect based on these two sets of findings.

Words low in arousal generally have slower and poorer word recognition, suggesting slower lexical access (Kousta et al., 2009) - this was also found in our button press task. It is possible that LA words elicit a stronger cognitive effort effect in the bilingual group, which in turn drives the pupil response up. This effort may be something integral to the LA category, and not captured by the other lexical dimensions we controlled for. The stimuli set is not imbalanced on the known lexical dimensions, but it is possible there is variation that occurs uncontrolled.

It is also possible the pupillary response to words is a combination of emotion embodiment and word processing effort, and how these two concepts are intertwined. Emotion embodiment in HA words is reduced in L2, which reduces the physiological response to them. On the other hand, the general processing advantage may also reduce the bilingual cognitive effort effect on the HA words, while in LA words the effect remains. These two effects combined would then cause the reduced difference between pupillary responses to HA vs. LA words.

This does not necessarily mean the effect found is only due to cognitive effort and should not be interpreted as reduced emotional resonance at all. Some element of the pupillary response is captured by the arousal dimension, given that the HA words indeed do elicit the strongest pupillary response in the monolingual group. Further work is required to clarify the origins of this effect fully - understanding the LA-HA mechanism would help us understand the origins of reduced emotional resonance better.

The models predicting reduced emotional resonance from participant language background showed no significant effect with any of the variables previously outlined in theory and the Bilingualism and Emotions Questionnaire (see section 1.5). These findings do not align with the contextual learning theory, or the theory of language embodiment. Interestingly, we did not find a significant effect of Age or Context of Acquisition, which generally are considered the most significant predictors of reduced emotional resonance in literature. Further, although a substantial number of the participants were classified as early bilinguals and the mode age for English exposure in fact was 0 years of age, we still found evidence for reduced emotional resonance in the bilingual sample.

It is possible there was not enough variation within the pupil responses, or in the participants' language background, and too many predictors for the models to run adequately. We chose the PCA approach for this specific reason; it reduces the number of predictor elements to uncorrelated items. Perhaps it is not possible to predict physiological responses from bilinguals' complicated language background information, or this would have to be done in a multi-lab study format; the GLMEM comparisons between all trials vs. only recognised trials suggest there might be issues with statistical power.

The lack of predictive value in any of the questionnaire items may suggests that reduced emotional resonance is driven by factors that have not been previously established in the literature and are not captured by the language background questionnaire. This is a mere speculation, but it highlights the importance of obtaining more experimental evidence. The theories remain largely untested, and whilst they make intuitive sense, it is not enough to make such strong claims based on very few empirical studies systematically investigating the background factors.

Another very interesting point about the present findings is that the only significant predictor of reduced emotional resonance was participants' preference to swear in L1 (this should be interpreted with the caveat that it was not significant in the recognised only GLMEMs). This suggests there is a link between the use of specific type of high arousing language and the physiological response this type of language elicits; swearing is one of the factors bilinguals often mention in anecdotal evidence about "feeling less" in their L2.

This finding also aligns with some previous work that has been done on swearing specifically: Dewaele (Dewaele, 2010a; Dewaele, 2010b, 2011, 2018) has found that bilinguals often prefer to swear in their L1 and tend to overestimate the offensiveness of swearwords in L2 (Dewaele, 2016). This was explained by the bilinguals overcompensating for the reduced emotional resonance the swearwords carry and trying to avoid socially inappropriate use of swearwords in their L2 (Dewaele, 2016). To establish if swearing indeed has a special role in the bilingual emotion processing, how this may relate to the existing theory and whether this effect will replicate, more experimental work is needed.

Another noteworthy point about the results is how participant proficiency, as measured by the LexTALE proficiency test for advanced English learners had no significant interaction with word type. The LexTALE scores of the present sample are very high overall, which is not surprising given that we tested university students who are studying in their L2 and are fully immersed in the language environment. The absence of any predictive value of the LexTALE test may be due to ceiling effects; variation within the scores is small. This suggests two things: the present sample is highly proficient in their L2, but also if we want to pinpoint specific factors in language skill that may interfere with physiological effects, we need more intricate proficiency testing. This, on the other hand will be more time-consuming for the participants and for the researcher.

Participant proficiency, however, must be considered in all bilingual research. This is especially important with pupillometry experiments, as there is evidence for heightened pupillary response due to increased cognitive load when processing one's L2 (e.g., Schmidtke, 2014). This effect can be seen in the present data as well; there is a main effect of participant group, showing that bilinguals overall had a stronger pupillary response. This is most likely due to cognitive effort. However, we do not think the effect renders all findings meaningless, as have tried to account for this confound at all stages of the experiment.

The stimulus words are controlled for variables that are known to increase cognitive load (for example, word length and frequency), and these variables are also added as covariates in the analysis. We have also added a word recognition task after each trial, and words not recognised/known by the participant were removed from the final analysis. We found no differences between the groups in their word recognition accuracy, and as stated above, the LexTALE proficiency scores are very high in both groups.

A final idea stemming from the present findings is the possibility that there is a difference between physiological responses and how bilinguals perceive their feelings about the language – further research should tap onto this difference between explicit, perceived emotion and physiological responses to emotional arousal. We will do this in study 4 by looking at the differences between rating data and physiological response to the same items.

The present experiment could have benefited from adding explicit measures to the questionnaire to gauge participants' self-reports about using their L2 and emotional situations, and how they *feel* in the L2. Reduced emotional resonance, however, has rarely been straightforwardly operationalised in the literature, and is a complicated concept. There is a difference in how bilinguals may perceive the emotions between their languages, and how the emotion concepts are embodied and what the physiological response related to those words is (Pavlenko, 2005). At present, the exact relationship between these concepts is unclear. To unravel it, it is important to do research in naturalistic, more applied settings as well as simplified, word-to-word level to identify the underlying structures in full.

Chapter 3 Comparing SCR measurement and pupillometry

3.1 Introduction

This chapter will focus on physiological measurement of emotion and compare two physiological measurement methods: pupillometry and skin-conductance response (SCR).

Physiological measurement techniques have provided relatively consistent research findings when measuring reduced emotional resonance of bilinguals' L2 (e.g. Caldwell-Harris & Aycicegi-Dinn, 2009; Caldwell-Harris et al., 2010; Iacozza et al., 2017; Toivo & Scheepers, 2019). Based on the activation of the autonomous nervous system, physiological responses to emotionally arousing stimuli are uncontrolled and automatic. These measurement methods probe into physiological reactions of participants, as elicited by an emotional stimulus, rather than their perceived emotional resonance. This may give a certain advantage over questionnaire-based approaches or affective ratings - participants are unable to exert conscious control over their physiological responses and there is no need for the experimenter to conceal the purpose of the experiment.

Thus far the most prominent physiological technique of measuring reduced emotional resonance in bilinguals is skin-conductance measurement (SCR), pioneered by Catherine Caldwell-Harris' group (Caldwell-Harris et al., 2010; Harris, 2004; Harris et al., 2006). Their experiments have explored a multitude of scenarios and different speaker groups (Spanish-English, Turkish-English and Mandarin-English), finding evidence for reduced emotional resonance over several experiments, and particularly in the case of late bilinguals.

Eilola and Havelka (2010) have also used skin-conductance measurement in an experiment exploring reduced emotional resonance in L2. Comparing an emotional/taboo Stroop Task and a physiological measure in Greek L1 speakers (English L2), they did not detect reduced emotional resonance with the Stroop task – their participants responded much like an L1 English control group which they were compared to. However, in the SCR task they found evidence for reduced emotional resonance. The observed skin conductance response to negative and taboo words was stronger in the monolingual English control group, as opposed to the bilingual group. Crucially, the difference between neutral/positive words and negative/taboo words was larger in the monolingual speaker group.

These findings together suggest that SCR can be used as a measurement method of reduced emotional resonance, and it has in fact provided some of the most compelling evidence to date. This chapter attempts to address a prevalent methodological issue around SCR measurement in bilinguals, namely the appropriate use of lexical covariates in material selection and statistical analyses. For a more thorough account of lexical covariates, see section on stimulus selection in the overall introduction in Chapter 1.

Caldwell-Harris and colleagues have typically split their stimuli into categories of different types of emotional language. These categories include taboo phrases like “She’s a slut”, insults (e.g. “You’re so fat”), childhood reprimands (e.g. “That’s not nice”) and endearments (e.g. “I love you”), aversive words and positive words. While these experiments generally detected reduced emotional resonance, there were differences between the word categories and speaker groups. For example, in the experiments with late Spanish-English and Turkish-English bilingual participants, endearments and taboo words yielded the stronger SCRs in L2 than in participants’ L1 (Harris et al., 2003; Harris et al., 2006). In the Mandarin-English experiment, on the other hand, bilinguals showed no difference in emotional SCRs between languages, except for L2 endearments, which elicited a higher skin-conductance response than their L1 counterparts. These findings may be due to different cultural norms in emotion expression.

This approach of splitting the stimuli into different groups of full sentences and utterances, played to the participants via audio makes the language setting more naturalistic. It also gives room for examining the effect of different stimulus types and how emotion may be embodied differently, depending on the context of learning and current use. However, there is no real control of the lexical variables affecting physiological reactions to the stimuli. This is problematic, as the autonomous nervous system also responds to tasks that require cognitive effort (Schmidtke, 2014). For example, lower word frequency typically predicts larger pupillary reactions, indicating more effortful processing of rare words (for example, check Experiment 1, section 2.3.8). In experiments where this is not fully controlled for, it may be possible that the differences between the stimulus categories are in fact due to lexical confounds.

No previous research has directly compared pupillometry with skin-conductance measurement in bilinguals. Pupillary responses are faster than skin-conductance responses: they typically peak and re-set within 2 seconds, whereas skin-conductance responses may take up to 8-10 seconds. Both measures are relatively noisy and require elaborate pre-processing. The two methods have been shown to follow similar patterns when measuring emotional arousal (Bradley et al., 2008).

Skin-conductance measurement is often done with images (e.g. Bernat, Cadwallader, Seo, Vizueta, & Patrick, 2011; Bradley, Codispoti, Cuthbert, & Lang, 2001) or auditory stimuli. Visually presented linguistic stimuli are used somewhat rarely (Boucsein, 2012). For pupillary measurement, the single-word stimuli approach has worked on our previous experiments (see chapter 1 and (Toivo & Scheepers, 2019). In fact, as the gaze should be as steady as possible for pupillary measurement (eye-movements add noise to the measurement of pupil size), the single-word approach may in some cases be better than presenting full sentences. Further, we have not observed strong habituation effects ("flat-lining") with pupil measurement, which contrasts to some extent with skin-conductance measurement (Boucsein, 2012). In fear-conditioning paradigms, it has been found that pupil responses are less prone to habituation effects than SCR (Leuchs, Schneider, & Spoormaker, 2019).

Eilola and Havelka's (2010) experiment is the only one to date using a single-word paradigm with skin-conductance measurement to explore reduced emotional resonance of L2. Their participants completed an emotional Stroop task while their skin-conductance levels (as opposed to skin conductance response) were measured. It should be noted that the word stimuli used were split into categories (neutral, positive, negative and taboo words). These words were controlled for frequency, length and familiarity, offensiveness and emotional arousal. However, the controlling was achieved via one-way ANOVAs with Dunn-Bonferroni corrections showing no word-category differences at item level, but these differences were not included analyses of the experimental data. Scheepers (2014) has argued that such control predictors should be included in the analysis stage as well to account for repetition in the items and to avoid making spurious inferences.

The first aim of this experiment is to explore whether reduced emotional resonance in bilinguals can be detected with a single-word paradigm using SCR measurement under strict control for lexical covariates. The second purpose of this experiment is to see whether we can replicate our pupillometry findings from Experiment 1 with an SCR paradigm.

3.2 Method

3.2.1 Materials and stimuli

A short pen-and-paper language background questionnaire was administered to identify participants' language background (Appendix A).

The stimulus materials for the skin-conductance experiment were taken from Experiment 1 (check section 2.2.1). The stimuli consisted of 240 English words (80 high arousing, 80 low arousing and 80 neutral distractor words). The full list can be found in Appendix B.

3.2.2 Participants

A total of 79 participants were recruited (27 male, 59 female). Participants were aged 18-25 years and the mean age was 18.9 years (SD=1.2 years). Participants were from 14 countries (a full list can be found in Appendix C). Fifty-seven of the participants were bilingual and 22 were monolingual English speakers. The bilingual participants were from a variety of different language backgrounds, including both early and late bilinguals. Average stay of participants in an English-speaking country (cumulative) was 8 years 5 months (SD= 96.3 months). The length of stay in an English-speaking country ranged from 3 months to 20 years. All bilingual participants had started learning English between 0 and 13 years. Bilingual group's mode age for English exposure (starting point for English acquisition) was 5 years, with 8 participants having learnt English from birth. Mean age for English acquisition was 4.6 years (bilingual participants). All participants were taking Introduction to psychology class at Boston University and were awarded course credits for their participation.

3.2.3 Apparatus

A Davicon C2A Custom Skin Conductance Monitor (NeuroDyne Medical Corporation) was used to record the skin-conductance response and skin-conductance level of the participants. The stimuli words were presented on a separate screen through ePrime 2.0.

3.2.4 Procedure

The procedure was kept as similar to Experiment 1 as possible given the limits imposed by the software, and different paradigm requirements for eye-tracking versus skin-conductance measurement.

Participants were first asked to fill out the short language questionnaire and complete the LexTALE test (Lemhöfer & Broersma, 2012) on a computer. After completing the questionnaire and the LexTALE test, two skin-conductance sensors were attached to participants' non-dominant hand (index and middle finger), and a 30-second window of measurement was taken before starting the experiment to allow for the signal to stabilise. The experiment had 240 trials in English, presented in a randomised order. Each trial consisted of a fixation dot for 1000ms, then a mask of X's for 1000 ms, the word presentation for 3000ms, and finally the second mask was shown for 5000 ms, followed by a question mark. Total length of each trial was 10 seconds (this was deliberately made longer than trials in experiment 1, as SCRs are generally slower than pupillary responses). Participants were instructed to wait for the appearance of the question mark and then press letter S on a keyboard if they did not recognise the word, and letter A if they did, respectively. The skin-conductance recording for each trial was controlled by the experimenter, and participants were given three breaks throughout the experiment (every 60 trials).

3.3 Results

3.3.1 LexTALE

Participants' LexTALE proficiency scores are summarised in the Table 10 by participant group (bilingual/monolingual English speaker).

Table 10 Mean LexTALE scores and SDs by participant group

Group	Mean	SD	Min score	Max Score
Bilinguals	80.74%	11.24%	56.25%	98.75%
Monolinguals	92.34%	7.36%	71.25%	100.%

3.3.2 Data exclusion

A number of data points were excluded for technical reasons. Eprime stopped working during five sessions, and consequently the SCR data from those five sessions could not be used. This resulted in a final sample 74 participants, of which 53 were bilinguals and 21 were monolinguals. 337 trials were removed (from two participants) because Eprime temporarily froze during the experiment. A further 23 trials (from one participant) were removed because there was no SCR signal.

During the experiment the experimenters were noting down any trials where the participant was moving, causing the SCR live curve on Neurodyne to visibly increase and not return to baseline during the following trial. A total of 984 trials from 50 participants were removed because of movement. Overall, these two procedures resulted in 5% of the data being excluded for movement noise and 2% of the data being excluded for technical issues.

3.3.3 SCR pre-processing

The Davicon C2A subtracts the base point from the maximum score during each 10-second recording interval, and returns a numeric value measured in micromhos. This signifies the amplitude of the phasic SCR, and is the measurement unit we are transforming into area under the curve values.

The data were downsampled into 100-millisecond time bins and the length of each trial was set to 10 seconds. Some trials were longer than this due to manual synchronisation of Eprime and Neurodyne measurement, and any time bins over 10 seconds were excluded from the analysis. For the same reason, a number of trials were shorter, and the missing values in those trials were replaced with a mean SCR value of that specific trial across all time bins. Once interpolated, the data were baseline corrected in a similar manner to Experiment 1 (see page 54) where the baseline SCR value (time bin 100ms) was extracted from all time bins resulting in a *proportional* change of SCR for every trial. The data were then converted into area under the curve values (time bins >2500ms summed up * 100). The 2500 ms time bin was chosen as a cut-off for area under the curve values, as SCRs are generally rather sluggish, and from Figure 7 (see below) it can be seen that the effect starts to emerge only after 2500ms.

3.3.4 Button responses

The overall mean likelihood of word recognition was ~92%. The bilingual group's mean likelihood was ~93% and the monolingual group's mean likelihood was ~90%. The bilingual group's recognition likelihood ranged from ~48% to 100% and the monolingual group ranged from ~66% to ~99%.

We ran similar analysis to Experiment 1 (see section 2.3.6); first word recognition was predicted from participant group, word type and the lexical covariates. To do this, participant group and word type were deviation coded, and lexical covariates and LexTALE scores were mean-centered. Binomial mixed effects models with the L-BFGS-B optimiser from the optimx package were run to examine the relative effect of each of the predictor variables.

For the fixed effects structure, group, word type, LexTALE score and the lexical covariates (see table 11 below) were entered as fixed main effects, and word type, LexTALE score and the covariates were entered as 2-way interactions with group. The random effects structure included subject and item random intercepts, and the following random slope terms: word type and the covariates were entered as by-subject random slope effects (all of these variables were within-subjects/between items), and group was entered as a by-item random slope effect. Full model syntax can be found on Appendix D

From the estimates table (Table 11) and blobbogram (Figure 6) below it can be seen that there is no difference in the word recognition likelihood between the experimental groups. Interestingly, there is also no effect of word type, perhaps due to the longer viewing time. As expected, higher lexical frequency predicts higher word recognition likelihood. Higher LexTALE scores predict higher word recognition likelihood. The interaction of group and word type is also significant. Decomposed into simple effects, higher lexical frequency predicts lower word recognition likelihood in bilinguals (ß = -0.49, SE = 0.24, t = -2.02, p = 0.043*). This is somewhat unexpected. In monolinguals, higher lexical frequency predicts higher word recognition likelihood (ß=0.49, SE=0.25, t=2.03, p=0.041*)

Table 11 Button response model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
(Intercept)	5.26	0.30	17.48	<0.001***	4.67	5.85
Word Type	-0.15	0.31	-0.47	0.63	-0.76	0.47
Group	0.33	0.71	0.46	0.65	-1.06	1.72
Lexical frequency	1.33	0.17	7.86	<0.001***	1.00	1.66
Bigram frequency	-0.03	0.17	-0.16	0.87	-0.35	0.30
Concreteness	-0.04	0.16	-0.22	0.82	-0.34	0.27
Length& orthographic neighbours	-0.16	0.16	-1.02	0.31	-0.48	0.15
LexTALE	0.11	0.02	4.62	<0.001***	0.06	0.15
Group:Word Type	0.03	0.43	0.08	0.94	-0.80	0.87
Group:Lexical frequency	-0.49	0.24	-2.03	0.04*	-0.97	-0.02
Group: Bigram frequency	0.35	0.22	1.58	0.11	-0.08	0.78
Group:Concreteness	-0.10	0.21	-0.48	0.63	-0.52	0.32
Group:Length& orthographic neighbours	-0.17	0.23	-0.73	0.47	-0.63	0.29
Group:LexTALE	0.01	0.06	0.23	0.82	-0.11	0.13

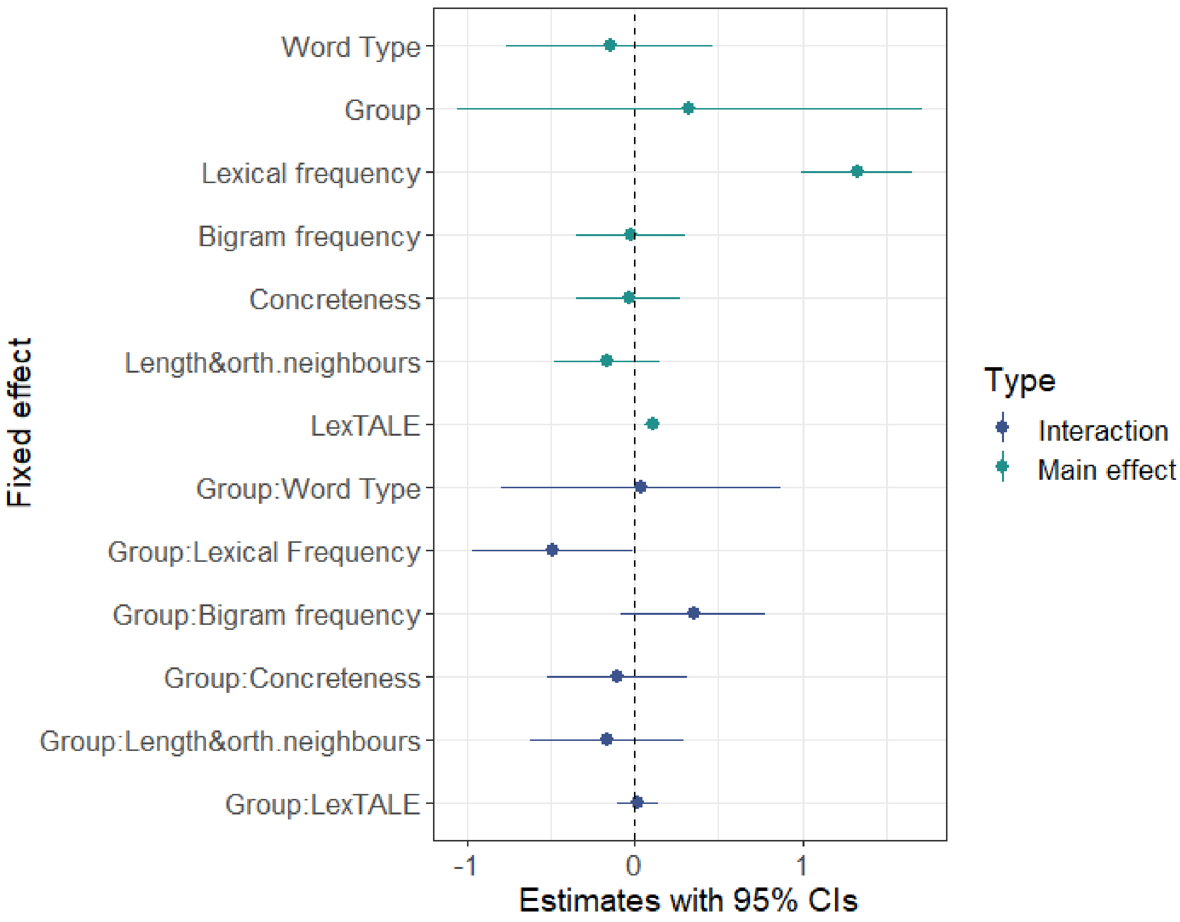

Figure 6 Button response task: Fixed effect estimates with 95% Confidence Intervals

3.3.5 SCR data and linear models

Figure 7 (below) shows skin-conductance responses averaged for each time bin, word type and participant group (B=Bilinguals, E=Monolingual English speakers) and standard error bars. Interestingly, the pattern seems to be opposite to what we have observed in previous pupil experiments: in the bilingual groups (left panel) the difference between HA (high arousing) and LA (low arousing) words seems to be larger than in the monolingual group (right panel). Also, LA words seem to elicit higher SCRs than HA words in the monolingual group. The overall SCRs seem higher in the monolingual group, which is a pattern we have typically observed in the pupil responses of bilinguals and have previously attributed this to increased cognitive effort. Responses to DI (neutral distractor words), marked in blue, seem to fall between the two other word types, which is also different to our pupil findings where DI words typically elicit the strongest pupil responses.

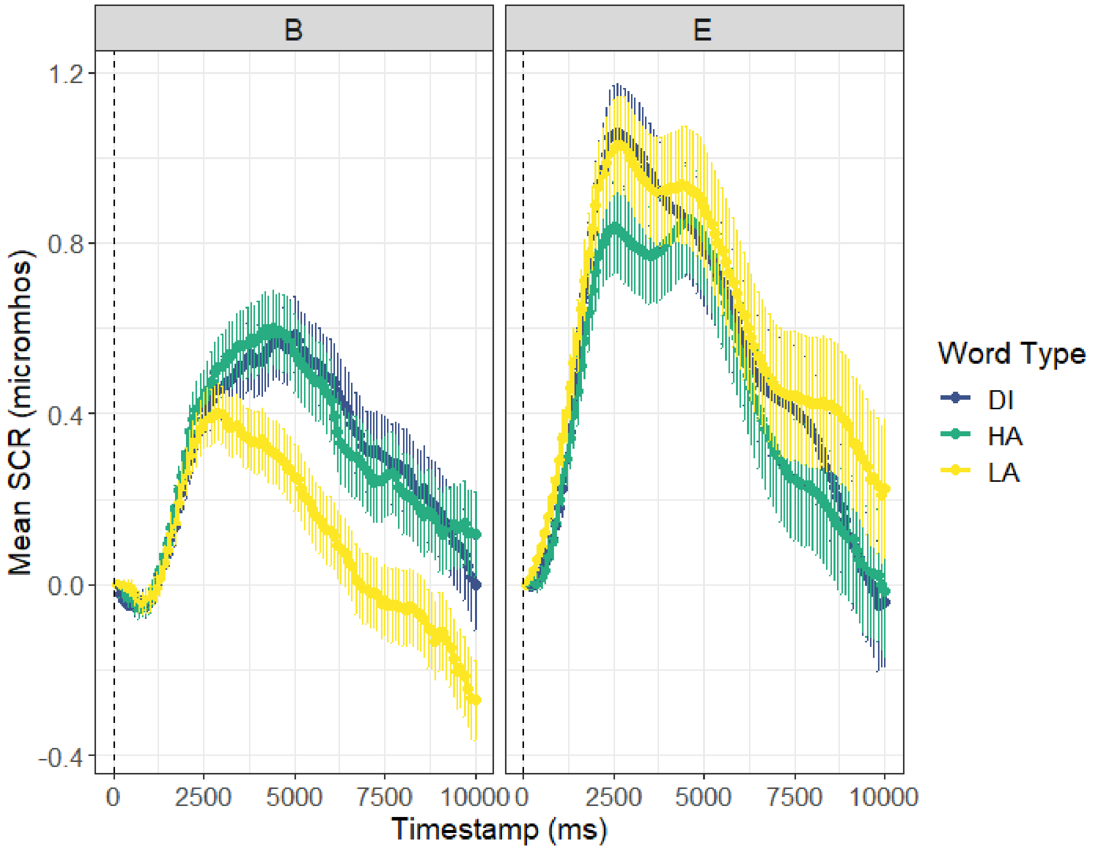

Figure 7 SCR (micromhos) averaged across time, split by participant group and word type

Next, linear mixed models predicting area under the curve values were run. It should be noted the data have extreme kurtosis (30), but the skewness is fairly small (1.5) – see boxplot (Figure 8) below. It has been shown that kurtosis negatively affects test power rather than inflating Type 1 error rate (Khan & Rayner, 2003), which is why a linear model was chosen. The findings of this experiment should be interpreted with this caveat.

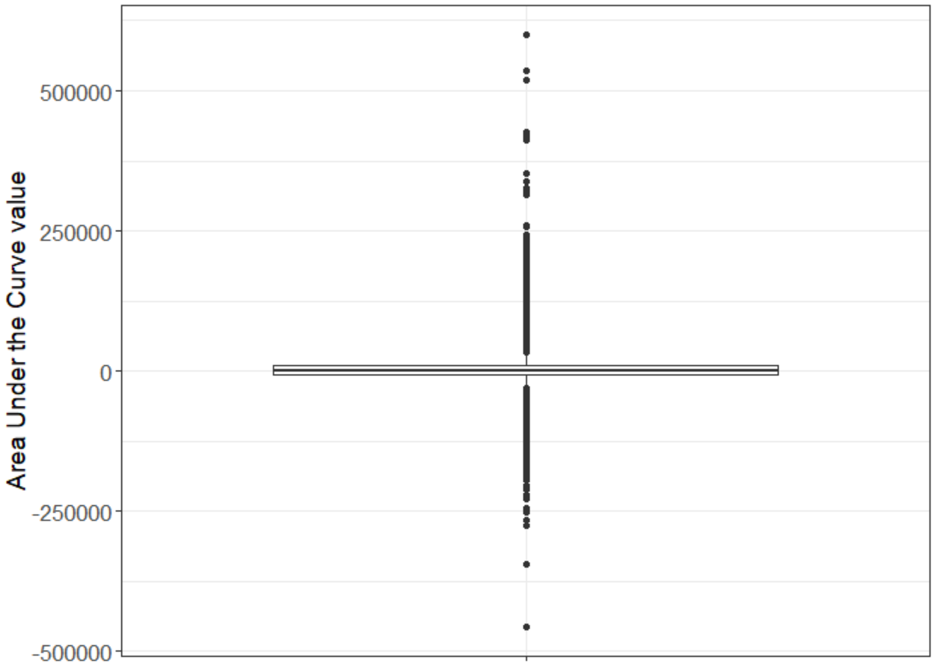

Figure 8 Distribution of the SCR area under the curve values

For all the linear models, the L-BFGS-B optimiser from the optimx package was used (see Appendix D for full model syntax). The model structure detailed above for the button press task was used for the area under the curve models as well.

The blobbogram (Figure 9) and summary table (Table 12) show that most of the lexical covariates' estimates are relatively small. This is possibly due to the longer viewing time of the items. There is a main effect of group, showing that the monolingual speakers overall had a higher SCR, which contrasts our previous findings. Interestingly, there is not effect of word type. There is a trend suggesting an interaction between group and Word Type in the opposite direction expected, but this is not significant.

Table 12. SCR Area under the curve model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
Intercept	3082.88	548.50	5.62	<0.001***	2007.84	4157.93
Word Type	1141.30	821.23	1.39	0.16	-468.29	2750.88
Group	3999.46	1415.43	2.83	<0.001***	1225.28	6773.64
Lexical frequency	299.16	355.11	0.84	0.40	-396.84	995.16
Bigram Frequency	40.37	364.06	0.11	0.91	-673.18	753.92
Concreteness	-52.60	368.60	-0.14	0.89	-775.04	669.84
Length& Orthographic neighbours	-266.88	381.27	-0.70	0.48	-1014.15	480.38
LexTALE	-72.60	47.31	-1.53	0.12	-165.34	20.13
Group: Word Type	-3463.10	1877.76	-1.84	0.07	-7143.45	217.24
Group: Lexical frequency	443.94	819.65	0.54	0.59	-1162.54	2050.43
Group: Bigram frequency	565.38	845.46	0.67	0.50	-1091.69	2222.44
Group: Concreteness	410.41	851.19	0.48	0.63	-1257.89	2078.71
Group: Length & Orthographic neighbours	-1162.71	879.80	-1.32	0.19	-2887.08	561.66
Group: LexTALE	-191.67	124.03	-1.55	0.12	-434.75	51.42

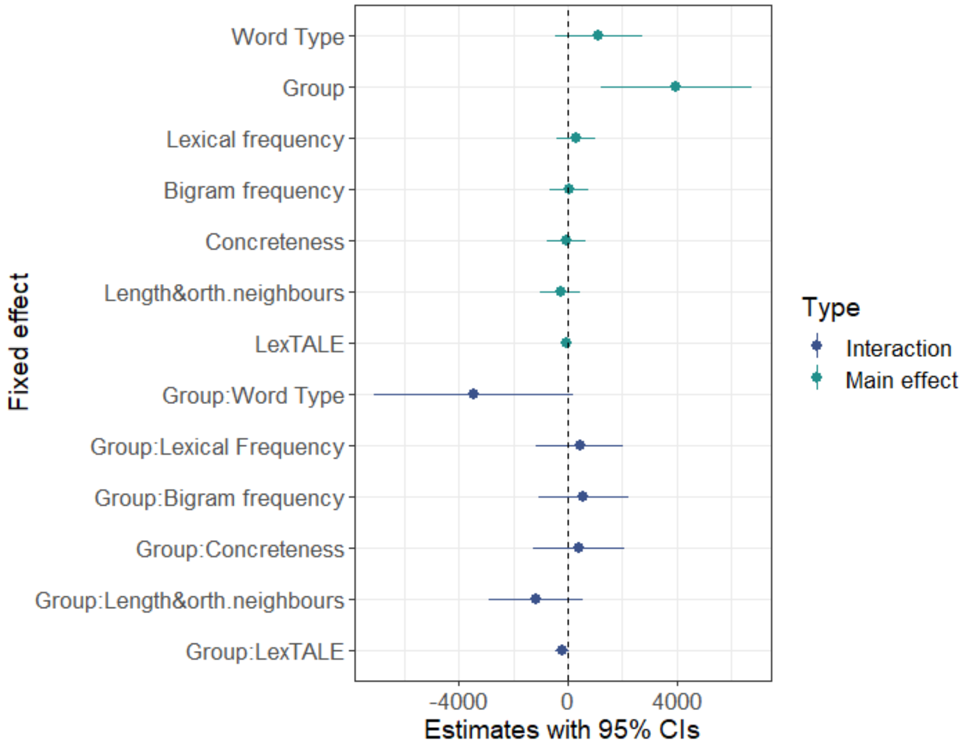

Figure 9 SCR model estimates with 95% Confidence Intervals

3.3.6 By-item correlations with Experiment 1

By-item correlations were explored with Experiment 1 data. First, a mean area under the curve values were calculated for each item (all word types included) for both Experiment 1 and Experiment 2 data. These values were then compared using Spearman correlation. In the full data set (without splitting the data into participant groups), we found no correlation between the area under the curve values (rs(236)=-0.02, p=0.81). Next, the data were split into participant groups. In the monolinguals, there was no significant correlation in Experiment 1 and 2 area under the curve values (rs(236)=-0.04, p=0.56), and the same was true for the bilinguals (rs(236)=-0.04, p=0.58).

3.4 Discussion

The aim of this experiment was to examine whether skin-conductance responses to emotional language can be detected with single-word stimuli. Our second aim was to see whether we can replicate out previous pupillometry findings using the same stimuli in an SCR paradigm.

There was a main effect of group, but to the opposite direction we have observed in Experiment 1; the monolingual group had larger overall SCRs than the bilingual group. Further, when broken down into word types, the monolingual group seemed to have the opposite pattern in how they responded to different word types (see Figure 7), but the Group:Word Type interaction was not significant. Hence, the experiment did not detect a word type effect on the SCRs with single-word stimuli. Interestingly, the area under the curve values calculated by item were not correlated to those found in Experiment 1, which further supports the reversed pattern observed with bilinguals vs. monolinguals.

As can be seen in figure 8, the data had extreme kurtosis and were mostly centered around 0. This, in turn, suggests that the participants were habituated to the paradigm, as speculated in the introduction; perhaps single-word approach is not suitable for SCR measurement, or there were simply too many trials, making habituation more likely. The only other experiment using single-word paradigm to detect reduced emotional resonance (Eilola & Havelka, 2010) had a significantly smaller number of trials (80 as opposed to 240). However, reducing the number of trials becomes problematic in terms of statistical power, when the lexical covariates are included in the analysis. A potential solution to this is to embed carefully controlled target words into sentences in an SCR paradigm, as has been done previously with pupillometry work (Iacozza et al., 2017).

We cannot draw firm conclusions about reduced emotional resonance in this sample. In fact, these findings are in contrast with our previous pupillometry findings: the pattern observed is the opposite to what we have consistently found with pupil experiments ((Toivo & Scheepers, 2019), Experiment 1, Experiment 4). In the present SCR experiment, bilinguals seemed to have higher response to high arousing words as opposed to low arousing words, and the opposite difference was observed in the monolingual group. However, it should be noted that these are merely descriptive differences, and the effect was not significant in the linear model.

Due to the different samples tested at different locations, there is not much sense in speculating whether the two methods tap onto different effects; the present data are not conclusive in this regard. Ideally, we would have conducted pupillometry and SCR-measurement simultaneously, or at least on the same participants. Unfortunately, this was not possible because of software and hardware availability.

One potential conclusion is that the findings are due to Experiment 2 bilingual sample behaving more like monolingual speakers. However, this does not explain why monolinguals exhibited such a different pattern to what we have previously observed with pupillometry. When comparing the bilingual samples from Experiment 1 and 2, there are two striking differences that make any interpretation difficult.

Firstly, Experiment 1 had a higher proportion of “balanced” early bilinguals who have acquired both languages from birth. This was not observed in Experiment 2 - most bilinguals had acquired English around the age of 5. The mean age of acquisition for Experiments 1 and 2 was very similar across both experiments, 5 years and 4.6 years, respectively. The small difference between these two means also suggests that Experiment 1 also had more bilinguals who had learnt their L2 from slightly older age. Hence, judging from the age of acquisition within and between these two samples, it cannot be concluded that either of the samples would be “more L1-like” in their language processing. When assessing the AoA as a potential explanation for the present findings, an important point to note is that we did not find age of acquisition to predict word type differences in pupillary response in Experiment 1.

Secondly, the bilinguals in Experiment 2 had typically stayed in an English-speaking country for a longer time than bilinguals in Experiment 1 (8 years 5 months and 5 years 8 months, respectively). These differences can potentially be explained by differences in immigration structure in Europe and the USA. A number of the early bilingual speakers in Experiment 1 had an English-speaking parent, but they had lived in multiple countries and their length of stay in an English-speaking country was more commonly split between a number of countries. On the other hand, participants in Experiment 2 had typically moved to the USA at a very early age, or lived their whole life there, but spoken only their indicated L1 for the first few years (presumably at home with their parents). These participants started acquiring English slightly later, at the age of 3-5. This suggests the two samples are different in terms of their context of learning English.

Another possible explanation for the discrepancies between Experiment 1 and 2 would be language proficiency differences between the groups of participants. However, this does not seem to be feasible either, as bilinguals in both Experiment 1 and 2 had very high scores on the LexTALE English proficiency test. Score differences between the two groups seemed marginal at best, with the participants in Experiment 1 having slightly higher scores (mean scores of 83% and 80%, respectively). These three factors combined illustrate that classifying bilinguals based on a limited set of language background questions is difficult and lends support to viewing bilingualism as a continuum rather than a theoretical domain with clearly separable subtypes.

The monolingual effects are equally, if not more, perplexing. Based on the linear mixed effects model, the monolinguals’ SCRs were overall higher than the bilinguals’, as reflected in a significant main effect of participant group. This is an effect we have typically observed with the *bilingual groups* in our pupillometry experiments (see Experiment 1 and 4, and Toivo & Scheepers, 2019). We have interpreted this as the effect of cognitive effort associated with speaking in one's L2 (see: Schmidtke, 2014).

SCRs are also sensitive to cognitive effort such as lying (the common use as a polygraph, but also (Caldwell-Harris & Aycicegi-Dinn, 2009; Engström, Johansson, & Östlund, 2005; Nourbakhsh, Wang, Chen, & Calvo, 2012), which is why the present SCR findings are particularly surprising. There are two potential explanations to the reversed effect. Firstly, it has been suggested SCRs are not as sensitive to cognitive effort as pupil response is (Haapalainen, Kim, Forlizzi, & Dey, 2010). Assuming that the participants are all high-fluency L2 speakers immersed in an L2 environment, perhaps the effect of cognitive effort was too small to detect with SCR.

It is also possible that there was no effect of cognitive effort associated with the task at all - the words were shown to the participants for a longer time, so it is possible there was no real effort of reading. To date, there is no systematic research into how effects of lexical covariates may interact with prolonged viewing time, and how this might affect word recognition – maybe the effects of covariates related to cognitive effort (such as word length and frequency) are diminished if viewing time is longer. It would also be interesting to explore this with pupillometry. Increasing viewing time of the words, provided the task is not based on reaction times, might help eliminate the effect of cognitive effort, which is problematic for all bilingualism research.

Another interesting finding is the effect of neutral distractor words (it should be noted neutral words were not included in the linear mode, which means these findings are merely descriptive). Contrary to the pupil experiments we have carried out, the distractors seemed to elicit effects similar to what we originally expected them to, and the magnitude of SCR response fell between the high arousing and low arousing words. Previously, we have interpreted the high response to neutral words as an effect of surprise. While SCR's are known to be sensitive to surprise (Nikula, 1991), it seems we did not detect similar surprise effect as we have with the neutral words across our pupillometry experiments. This is potentially due to the measurement sensitivity discussed above, or the longer viewing time of the items. Alternatively, we may have misinterpreted the distractor word effect in our pupil experiments (this point is discussed more in detail in Experiment 3, see section 4.4).

The present SCR experiment is mostly inconclusive. It seems that our participants showed a high degree of habituation in their SCRs. This may suggest that single-word stimuli paradigms with a large number of items are not very effective, as we did not detect any word type differences. We also failed to replicate previous pupillometry findings, as the interaction of word type and participant group was to the opposite direction expected. Further research should look into conducting pupillometry and SCR measurement concurrently (see: Bradley et al., 2008).

Chapter 4 Comparing cognitive and physiological measurement

4.1 Introduction

Reduced emotional resonance in bilinguals is often measured with cognitive paradigms, such as the Stroop Task (Dudschig et al., 2014; Eilola et al., 2007; Fan et al., 2018) and the Lexical Decision Task (Conrad et al., 2011; Kazanas & Altarriba, 2016; Ponari et al., 2015). Using cognitive paradigms is quick and simple, but the findings are not very conclusive to date. Some of the experiments have detected reduced emotional resonance of L2 using a number of different cognitive paradigms (for example: Altarriba & Canary, 2004; Fan et al., 2018; Fan et al., 2016; Ivaz, Costa, & Dunabeitia, 2016; Segalowitz, Trofimovich, Gatbonton, & Sokolovskaya, 2008; Winskel, 2013), while others have not detected the effect, or have detected it with only some of the measures (for example: Dudschig et al., 2014; Eilola & Havelka, 2010; Eilola et al., 2007; Kazanas & Altarriba, 2016; Sutton, Altarriba, Gianico, & Basnight-Brown, 2007).

Given how many experiments are using these paradigms, and the large variance in their findings, it is surprising there has not been any systematic methodological investigation. Strong conclusions are drawn about the extent and origins of the reduced emotional resonance effect, but the studies rarely consider how the chosen methodology may have affected their findings. Whilst these paradigms are well established in other psycholinguistics research areas, using them to measure bilingual emotion can be complicated for multiple reasons. The purpose of this chapter is to address the methodological problems associated with cognitive paradigms and compare cognitive and physiological measurement of emotion.

Cognitive paradigms are, rather obviously, based on tasks that require at least some cognitive effort. This is potentially problematic when studying bilinguals, as opposed to monolingual speakers. There is a long ongoing debate about the potential cognitive benefits of speaking two or more languages (for a review, see: Adesope, Lavin, Thompson, & Ungerleider, 2010). On the other hand, there typically is some degree of increased cognitive effort when operating in one's

L2. These factors bring noise that is very difficult to remove, given that they are assumed to be fundamental characteristics of the target population. The cognitive effort effect can be controlled to some extent with the inclusion of lexical covariates, but it is unclear whether it can ever be fully disentangled.

Most of the studies that use cognitive paradigms are not systematic in controlling the lexical covariates of their word stimuli. This introduces a number of possible confounding variables – for example, word processing and recognition are mediated by several lexical characteristics of the words, such as length, frequency, number of orthographic neighbours (New, Ferrand, Pallier, & Brysbaert, 2006), concreteness (Brysbaert et al., 2014), valence and arousal (Kuperman, Estes, Brysbaert, & Warriner, 2014; Scott et al., 2014). Most of the studies that use cognitive paradigms control for length and frequency, but fail to account for a number of other variables listed above.

Many of the studies use translation equivalents when testing in both L1 and L2 (e.g. Fan et al., 2016; Grabovac & Pléh, 2014; Winskel, 2013). This is problematic for two reasons. Firstly, if the experiment is designed so that participants see stimulus words in both (or all) test languages, and all the stimulus words are translated, the participants will effectively be presented the same meaning twice (or more times, given the design might have items repeating). Secondly, very few of the studies control for the lexical characteristics of the translation equivalents they use. The original word stimuli (typically English) may be controlled well, but stimulus words in the other language are simply translated from the English originals and not normed in the other language. While some of characteristics related to semantics, such as arousal and valence, may be similar across languages, length and frequency certainly will not be. This creates additional dependencies within the data and introduces new confounds. To adequately control for the lexical features of the stimuli, the stimuli words should not be translation equivalents. They should also be normed separately by language, and then matched on each of the lexical covariates for the final stimulus set.

Measuring emotion processing through cognitive paradigms assumes automaticity of emotional response. Depending on the task, to demonstrate reduced emotional resonance of L2, bilingual participants should either have an L1 advantage effect, or an L2 advantage effect. This depends on the type of the test and is based on the assumption that emotion processing in L1 is more automatic. The L1 advantage effect means an increased emotional congruence in a task in participants' L1 (Pavlenko, 2012). For example, in an Implicit Association test, this should lead to more pronounced associations of L1 words to the emotion categories presented in the experiment (e.g. positive and negative). In a Lexical Decision Task, emotional words in L1 should have a larger facilitation effect than in L2. Alternatively, an L2 advantage effect means there is a smaller interference effect of the emotional aspects of the stimuli. For example, in an emotional Stroop tasks, the interference of emotional words should be smaller in L2.

Hence, cognitive paradigms do not measure a direct emotional response, but rather how the emotional response facilitates or interferes with other processes (typically reaction time or accuracy). In comparison to physiological measurement of emotion, this adds a further dimension of measurement. This may partially explain why cognitive paradigms have been so inconclusive in measuring reduced emotional resonance.

One way of measuring whether cognitive paradigms can detect reduced emotional resonance is to compare them with a physiological measure. Thus far, only one experiment has done this; Eilola and Havelka (2010) used skin-conductance measurement alongside an emotional Stroop task. In the Stroop task, they found identical interference effects between the monolingual speaker group and the bilingual group. When measuring the participants' skin-conductance responses, the emotional stimuli (only presented in one language) elicited larger responses in the monolingual speaker group. This suggests the two methodologies may tap onto different facets of language processing; cognitive paradigms are based on *meaning* of the words, whereas physiological measures are based on *feeling* of the words. Eilola and Havelka (2010) pointed at the distinction between denotative and connotative meaning of words. Highly proficient bilinguals may understand the meaning perfectly well, but still lack the feeling and connections behind the word. This will be discussed in more detail in the discussion part of this chapter (see section 4.4).

Most experiments measuring reduced emotional resonance with a cognitive paradigm have used either a Stroop task or a Lexical Decision task. Emotional Stroop task has been criticised for the semantic incompatibility between the stimuli words and colour information. There is a lack of direct semantic conflict/agreement, which Stoop Tasks are essentially based on (Fan et al., 2018). For this reason, the present experiment will use a Lexical Decision Task; the task itself is directly related to the words and their meaning. We will conduct the LDT with a new, algorithmically controlled stimulus set across two languages (English and German) with no translation equivalents. Lexical covariates will also be included in the analysis stage, and we will compare the LDT findings to pupillometry data gathered from the same participants.

This experiment will also explore a possible positivity bias in bilingual emotion processing. This has previously been observed in bilingual reading times (Sheikh & Titone, 2016) – positive words had faster first-pass reading times in L2, but this facilitation effect was not found with negative words. The authors suggested this is due to emotional disembodiment in negative L2 words. Caldwell-Harris and colleagues (2010) provide further evidence, although tentative, for positivity bias in L2. They found that listening to endearments in English (participants' L2) elicited stronger SCRs than endearments in Chinese (L1). The pattern was reversed for insults. This could potentially signify that positive phrases in L2 are more strongly embodied than negative phrases. However, it should be noted that this study did not use any other types of positive high arousing words as stimuli, and the stimuli words were not normed on valence, which means we cannot draw strong conclusions.

Reading time measures provide contextual cues (Sheikh & Titone, 2016), and thus may be more likely to tap onto the concept of a word rather than just the denotative meaning of an isolated word. However, based on our previous work on strictly controlled and normed stimulus sets, we do not believe there will be a positivity bias in bilinguals' *physiological response* to emotional language. Pupillary response has been previously found to be driven by arousal rather than valence (Kuchinke, Vo, Hofmann, & Jacobs, 2007; Partala & Surakka, 2003). We will investigate the positivity bias by splitting the word stimuli into categories in a different way in comparison to our previous experiments. Here, we will include high valence (positive) and low valence (negative) words as separate categories (both high in arousal), and compare them to neutral valence, low arousal words. We do not expect there to be a difference in pupil responses between the negative and positive stimuli words.

The experiment has three aims:

1. To conduct a lexical decision experiment with two well controlled stimulus sets, one in participants' L1 and one in L2, and investigate whether the paradigm will detect reduced emotional resonance in bilingual participants' L2. If the lexical decision task will show reduced emotional resonance of L2, we expect the facilitation effects of high arousing words to be higher in L1 as opposed to L2 (i.e. there will be a smaller difference in reaction times between high and low arousing words in L2, in comparison to L1).

2. To compare findings from the lexical decision task with a pupillometry task on the same participants, using the same stimuli. We expect to replicate out previous pupillometry findings and observe reduced emotional resonance in our bilingual group's L2, as suggested by smaller word type (high arousing versus low arousing) differences in participants L2 in comparison to L1.

3. To investigate whether we will find a positivity bias in the bilinguals' pupillary responses. We suggest pupillary response is mainly driven by arousal rather than valence. Hence, we have split the stimuli based on arousal *and* valence. We hypothesise there we will not be a positivity bias in participants' pupillary responses, and both positive and negative words will elicit a comparable pupillary response.

4.2 Method

4.2.1 Materials and Stimuli

4.2.1.1 English stimuli candidate selection

English words were picked from (Warriner et al., 2013) database for affective lexicon. The words were grouped into three categories based on the database's valence ratings on a scale ranging from 1-9, and all the words in the database falling onto these categories were picked for the initial candidate list (5086 words); positive (average valence rating 6.5 and above), neutral (4-4.8) and negative (less than 3). Acronyms and two-part words were removed, and it was ensured all the words included were at least three letters long.

In the next step, further word norms were collected from other databases. Concreteness ratings were collected from (Brysbaert et al., 2014) database, and word frequency (HAL) was collected from the English Lexicon project website (Balota et al., 2007), transformed into frequency per million and logarithmic frequency (log10). This procedure eliminated 450 words as the databases are not identical, leaving the candidate sample of 4636 English words that were normed on length, valence, concreteness, lexical frequency per million transformed into logarithmic frequency, arousal, and dominance.

4.2.1.2 German stimuli candidate selection

German words were selected from the Berlin Affective Word List Reloaded (BAWL-R) database (Vo et al., 2009). The database is the largest German database available to date, and consists of 2902 words, which is less substantial than any of the English databases available – hence, in the initial candidate selection phase, the words were not categorised as described above, but all the words were used. Frequency per million was also taken from BAWL-R. The German candidate words were also normed on all the 7 lexical properties listed above.

4.2.1.3 Comparing the languages

In the next step, English cognates were removed from the German candidate pool, leaving a final candidate set of 160 German words. Finally, translation equivalents were removed. Since the English candidate pool was much larger, the translation equivalents (107 words) were removed from that set, rather than decreasing the already rather small number of German candidates.

Using this method, the 160 word German candidate pool was not sufficiently large to run an algorithm that would produce a balanced set of stimuli. To extend the German candidate set, every 5th word from the English candidates were selected, and translated into German (906 words). To avoid including translation equivalents in the final stimulus items, these words were removed from the English stimulus candidate pool. Before further norming the new candidates in German, 97 of them were removed because they were cognates and 25 because the translation equivalent for German was not unique (appeared more than once in the candidate pool), or the translation had two or more words in it, leaving the additional German pool at 783 words. After this, the word translations were checked by two L1 German speakers, and 27 words were removed due to translation ambiguities.

Frequency ratings and number of syllables for the translated German candidate words were taken from the CELEX database (Mannheim frequency per million). This left a candidate pool of 647 new German words in addition to the 160 taken from German databases.

Table 13 Number of candidate items by language

	English	German (from BAWL)	German translated	German TOTAL
Negative	734	49	129	178
Positive	1655	41	291	332
Neutral	1234	70	227	297
Total	**3623**	160	647	**807**

4.2.1.4 Converting the scales

The five rating databases used for the stimuli all differ in their rating scales. To account for any differences, all the scales were transformed to range from 0 to 1. This was done by adding to, or subtracting from the smallest point of the scale a constant to make the scale starting point zero. Then, we divided the result by the number of scale points minus 1. So, for example, for the BAWL valence rating scale ranging from -3-3, to convert each observation we first added 3 and then divided the result by 6. The German LAN concreteness ratings were laid out opposite to the Brysbaert et al. (2014) concreteness database, i.e. high ratings in this database indicate a very concrete word, while high ratings in LAN indicate a very abstract word. After scaling the concreteness ratings to range from 0-1, the LAN ratings for German words were transformed to correspond the English ratings with a simple 1-scale proportion calculation.

4.2.1.5 Selecting the final material set

The final stimuli set was selected using an algorithm, which would pseudo-randomly select 40 word triplets (each with a POS, a NEG, and a NEU word) of 'maximally similar' words, where similarity was defined in terms of Euclidean distance in a 4-vector space (corresponding to standardized versions of the 4 control variables logfreq, syllables, letters, and concreteness). The algorithm ensured that cross-language differences in any of the 7 variables were as small as possible. Tables 14 and 15 below show the item norms for selected items.

Table 14 Item norms (means and SDs) for the selected German materials (40 positive high-arousal, 40 negative high-arousal, and 40 neutral low-arousal words).

	POS (HA)		NEG (HA)		NEU (LA)	
	mean	SD	mean	SD	mean	SD
Logfreq	0.896	0.991	0.815	0.885	0.770	0.861
Syllables	2.30	0.91	2.28	0.85	2.33	0.92
Letters	7.10	2.25	7.18	2.35	7.28	2.22
Valence	0.781	0.068	0.175	0.041	0.446	0.049
Dominance	0.621	0.093	0.368	0.091	0.504	0.086
Arousal	0.613	0.064	0.611	0.077	0.291	0.043
Concreteness	0.470	0.254	0.473	0.219	0.490	0.232

Note: apart from the intended differences in Valence, Dominance, and Arousal across the three word types, differences in the other control predictors were negligible (max. Cohen’s d = 0.136).

Table 15 Item norms (means and SDs) for the selected English materials (40 positive high-arousal, 40 negative high-arousal, and 40 neutral low-arousal words).

	POS (HA)		NEG (HA)		NEU (LA)	
	mean	SD	mean	SD	mean	SD
Logfreq	1.058	0.671	1.072	0.634	1.035	0.644
Syllables	2.33	0.97	2.33	0.97	2.33	0.97
Letters	7.23	2.12	7.23	2.12	7.13	1.92
Valence	0.776	0.046	0.195	0.038	0.433	0.027
Dominance	0.632	0.082	0.349	0.093	0.488	0.085
Arousal	0.607	0.061	0.584	0.057	0.298	0.035
Concreteness	0.483	0.242	0.493	0.225	0.479	0.228

Note: apart from the intended differences in Valence, Dominance, and Arousal across the three word types, differences in the other control predictors were negligible (max. Cohen’s d = 0.061).

Note on lexical frequency differences between German and English: This could be because of differently-sized corpora (more *missing entries* in German, which were simply replaced with an estimate of 0.5 per million word counts). The full list of chosen stimuli words can be found on Appendix B.

4.2.1.6 Principal Component Analysis

The selected German and English materials were pooled and the seven variables were entered into a principal component analysis. Five principal components were extracted, which together explained 95.3% of the original variance. Factor loadings are shown below in Table 16.

Table 16 PCA factor loadings

	Component				
	PC1: Length	PC2: Valence & Dominance	PC3: Arousal	PC4: Frequency	PC5: Concreteness
Logarithmic frequency	-.090	.024	.039	.992	-.070
Syllables	.953	.020	.003	-.072	-.164
Letters	.965	-.003	-.005	-.045	-.102
Concreteness	-.215	-.030	-.031	-.074	.973
Dominance	.039	.944	-.021	-.011	-.041
Arousal	-.002	.039	.998	.038	-.029
Valence	-.023	.941	.072	.041	.005

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

Rotation converged in 5 iterations. The cells shaded in purple show which covariates loaded highly on each of the Principal Components.

PC1 was strongly positively associated with length (in syllables and characters), PC2 with valence and dominance, PC3 with arousal, PC4 with lexical frequency, and PC5 with concreteness.

Anderson-Rubin factor scores for PC1 (length), PC4 (frequency) and PC5 (concreteness) were used as continuous control predictors in later analyses (using these principal components as covariates ensure zero predictor collinearity in the analysis models).

Figure 10 below illustrates how the selected items load on the two manipulated components (PC3 Arousal and PC2 Valence and Dominance). The high and low valence items are both high in arousal, and the neutral valence items are low in arousal. This separation should allow us to examine whether any effects found in the analysis stage are driven by arousal or valence.

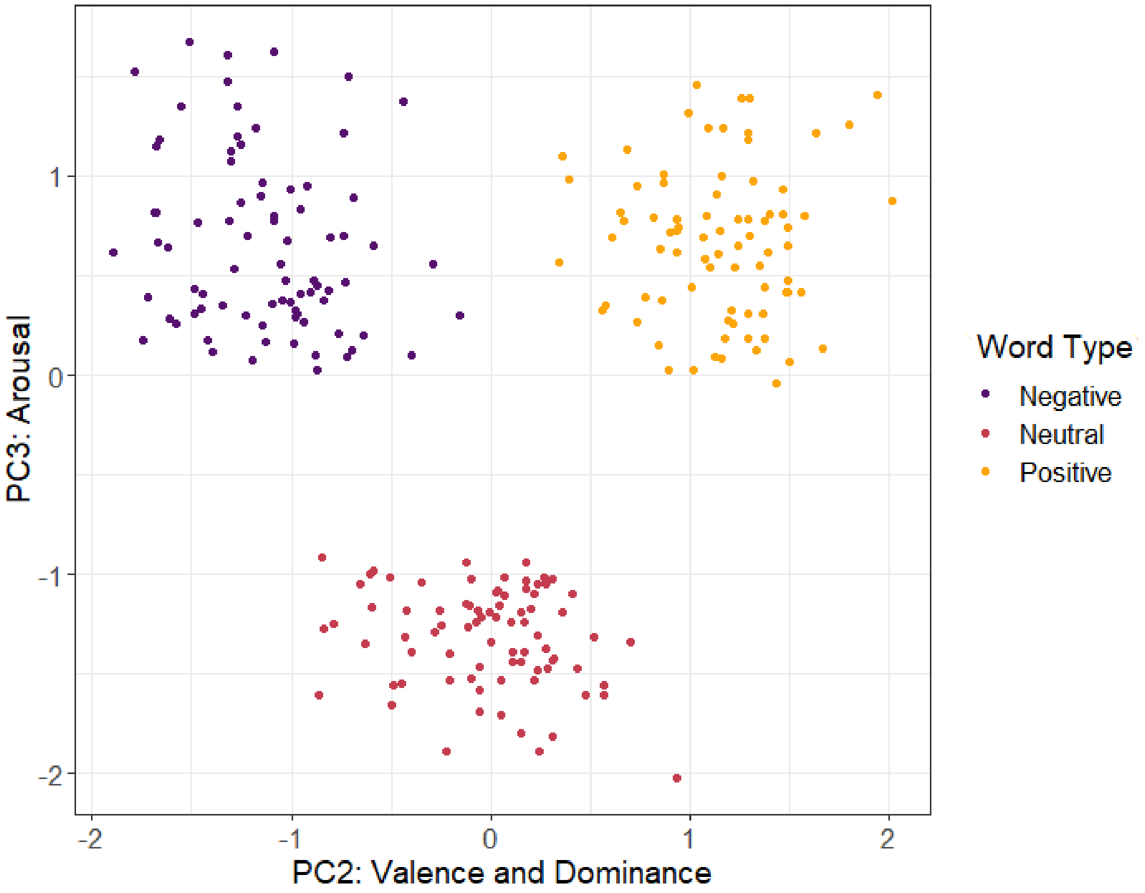

Figure 10 Selected items Arousal and Valence & Dominance PCA scores

4.2.1.7 Pseudowords

Non-words for both languages were generated using the Wuggy pseudoword generator (Keuleers & Brysbaert, 2010) and matched on length to the stimuli words. One-hundred and twenty pseudowords were generated for each language. Pseudowords were only used in the Lexical Decision Task, and not in the pupillometry part of the experiment.

4.2.2 Participants

A total of 60 participants were recruited (11 male, 49 female). Participants were between 18-48 years, and the mean age was 24.94 years (SD=6.27). Thirty of the participants were L1 German speakers and 30 were a monolingual control group of L1 English speakers. The majority of the German speakers were from Germany, apart from one participant from Switzerland and one from Luxembourg. Two of the participants listed another native language as well as German (Lithuanian and Hungarian). Average stay of participants in an English-speaking country (cumulative) was 3 years 4 months (SD= 33.87 months). The length of stay in an English-speaking country ranged from 6 months to 10.5 years. All participants had started learning English between 4 and 12 years. Mean age for English exposure was 9 years. Participants were paid or awarded course credits for their participation.

4.2.3 Apparatus

An SR EyeLink II head-mounted eye-tracker was used to record the pupil size at 250 Hz data sampling rate. Only the participant’s dominant eye was tracked and viewing of the stimuli was binocular. The display screen resolution was 1024*768 pixels and refresh rate 100Hz.

The Lexical Decision Task was conducted on ePrime 2.0.

4.2.4 Procedure

Participants were asked to come to the laboratory twice with approximately 3-7 days between the sessions. This was done to minimise the possible effects of recall of the words and habituation. The order of the sessions was counterbalanced. Monolingual English speakers did the tasks only in English, and German speakers did the eye-tracking task and the Lexical Decision Task both in English as well as German. The order of languages was counterbalanced in both sessions.

In the Lexical Decision Task session participants were instructed to respond to strings of letters appearing on screen with mouse clicks: right click if they think the string of letters is a non-word and left click if they think the stimulus is a real word. Participants were instructed to respond as quickly and accurately as possible. The Lexical Decision Tasks for each language had 240 trials, presented in a randomised order.

In the eye-tracking session participants were instructed to the Miles test to determine their eye-dominance, seated in front of a computer screen and the eye-tracker was set up. Calibrations were conducted after the first 10 trials and subsequently every 15-40 trials. The experiment had 120 trials, presented in a randomised order. The first ten trials were all neutral items to allow the participant a practice block. Each trial consisted of a fixation dot, then a mask of X's for 500 ms, the word presentation (the length of presentation was a function of the number of letters in the word (t=50ms + 26ms * length of word in characters), then the mask was shown again for 1700 ms, followed by a question mark. Participants were holding a game pad and were instructed to wait for the appearance of the question mark and then press the left key if they did not recognise the word and right key if they did, respectively. The eye-tracker recording for each trial started from the experimenter-initiated onset of the trial and ends to participants' trigger response making the typical length of a recording period for each trial ~3000ms. After the eye-tracking experiment participants were asked to complete the LexTALE test (Lemhöfer & Broersma, 2012) on a computer.

4.3 Results

4.3.1 LexTALE

L1 English speakers scored 92.63% on average (ranging from 71.25-100%), while L1 German speakers scored 86.89% on average (ranging from 56.25-100%). Eight English speakers scored 100% and one German speaker scored 100%. Table 17 below summarises the mean LexTALE scores and standard deviations by participant group.

Table 17 Mean LexTALE scores by participant group

Group	Mean	SD	Min	Max
German speakers	86.89%	11.70%	56.25%	100%
English speakers	92.63%	7.53%	71.25%	100%

4.3.2 Lexical Decision Task

The analysis is split into four parts: by-group analysis of reaction time (both groups), by-language analysis of reaction time (Germans only), by-group analysis of word recognition accuracy (both groups) and by-language analysis of button response (Germans only). Non-word trials and incorrect responses were removed from the analysis.

The reaction time (RT) data were modelled via generalized linear mixed effects models using the Gamma (identity) model family; the latter accounts for the characteristic positive skew in the distribution of RTs while still maintaining additive relationships between predictors and RTs. Hence, we were generous in data inclusion. The below figures show reaction time distribution of correct trials split by group (Figure 11) and language (Figure 13, German participants only).

Trials where reaction time was over 3000ms were removed. This resulted in 53 trials being removed (0.5% of the correct response data), out of which 33 were from English participants and 20 from German participants (10 trials from the English experiment and 10 from the German experiment).

4.3.2.1 By-Group reaction time analysis

4.3.2.1.1 Descriptive statistics

Table 18 Summary of mean RTs (ms) and standard deviations (in brackets) per group and word type

	Neutral	Negative	Positive
German speakers	856.36 (349.01)	800.62 (300.81)	785.80 (302.15)
English speakers	815.99 (361.94)	762.83 (323.64)	751.24 (307.09)

Figure 11 below and mean reaction times suggest the Germans were slightly slower in their response overall but the difference seems marginal. Neutral distractor items seem to elicit slowest reaction times in both groups, and the difference between negative and positive words seems to be very small, suggesting the speed of lexical decision is facilitated by arousal rather than valence.

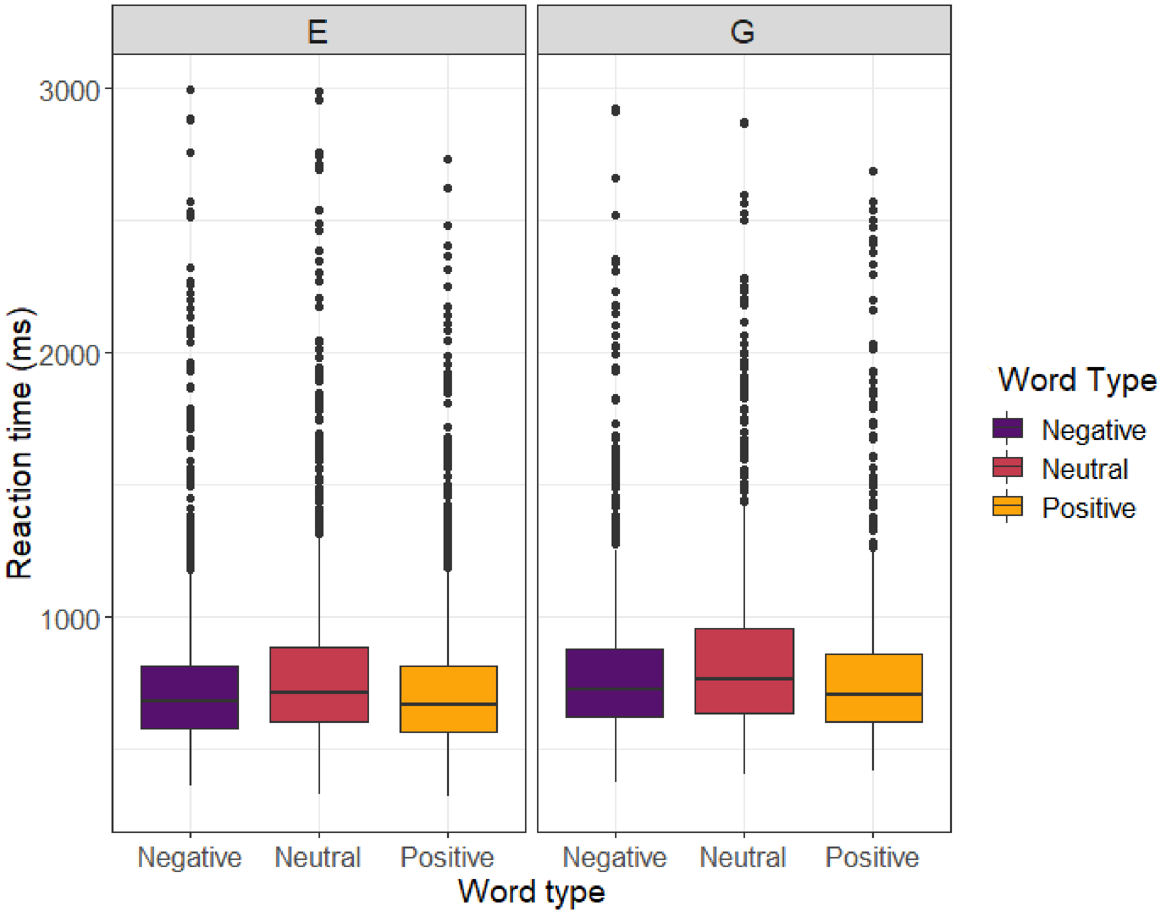

Figure 11 LDT reaction time distributions split by participant group and word type

4.3.2.1.2 By-group reaction time Linear Mixed Effects Models

By-group Lexical Decision Task data were analysed using generalised linear mixed effects models. Only the English language data were considered for this set of analyses. Participant group and word type were deviation coded for analysis. Word type was coded with neutral distractor words as the "baseline" and the deviation coding of positive and negative words were scaled to this baseline as separate variables (condition1=positive vs. neutral and condition 2=negative vs neutral). LexTALE scores, as well as the principal components used as item covariates (check method section for details) were mean centered. The by-group LDT reaction time model (see Appendix D for full syntax) included main effects of group, word type, LexTALE and the three principal components, namely Length, Frequency and Concreteness of the items. Word type and principal components were also included as fixed interactions with group. In the random structure, we included by-subject and by-item random intercepts. Word type was entered as a by-subject random slope and group and LexTALE were entered as by-item random slopes. The model was specified with Gamma family argument, and optimx package (Nash & Varadhan, 2011) L-BFGS-B optimiser

Table 19 and Figure 12 below summarise the fixed effects of the first by-group model, and their confidence intervals and estimates

Table 19 By-group LDT reaction time model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
Intercept	828.60	10.94	75.75	<0.001***	807.16	850.04
Group	40.71	19.38	2.10	0.04*	2.73	78.68
condition1 (positive)	-88.11	16.09	-5.48	<0.001 ***	-119.64	-56.58
condition2 (negative)	-72.77	16.10	-4.52	<0.001 ***	-104.32	-41.23
LexTALE	-1.57	1.11	-1.42	0.16	-3.73	0.60
Length (PC1)	32.17	6.10	5.27	<0.001 ***	20.21	44.14
Frequency (PC4)	-64.55	7.36	-8.77	<0.001 ***	-78.97	-50.13
Concreteness (PC5)	-6.66	5.84	-1.14	0.25	-18.10	4.79
Group:condition1	-8.53	21.07	-0.40	0.60	-49.82	32.76
Group:condition2	0.08	21.06	0.004	0.10	-41.21	41.37
Group:Length (PC1)	3.44	6.81	0.50	0.61	-9.92	16.79
Group:Frequency (PC4)	-15.24	7.84	-1.94	0.05.	-30.60	0.13
Group:Concreteness (PC5)	12.58	6.16	2.04	0.04*	0.50	24.65
Group:LexTALE	-1.73	2.16	-0.80	0.42	-5.95	2.50

As expected, the stimulus-related covariates affect reaction times: longer words have slower reaction times, whereas higher frequency words and concrete words facilitate faster word recognition. It should be noted these effects are not entirely reliable, given that the item-related covariates were not included in the random effect structure of the models.

It appears there is a main effect of Group on reaction times; German speakers were overall slightly slower in responding than the monolingual group. A main effect of condition (word type) can also be seen: both positive and negative words facilitate faster reaction times in comparison to neutral words. To further examine the effects of word type, second model was run with different coding of word type; in this model positive words were coded as baseline, and negative words were coded as condition 1 and neutral words were coded as condition 2, respectively. This model suggests both negative (estimate=15.21, t=0.993, p=0.32) and neutral (estimate=87.20, t=5.42, p<0.001) words have slower reaction times in comparison to the positive words. However, the difference between positive and negative words is not significant.

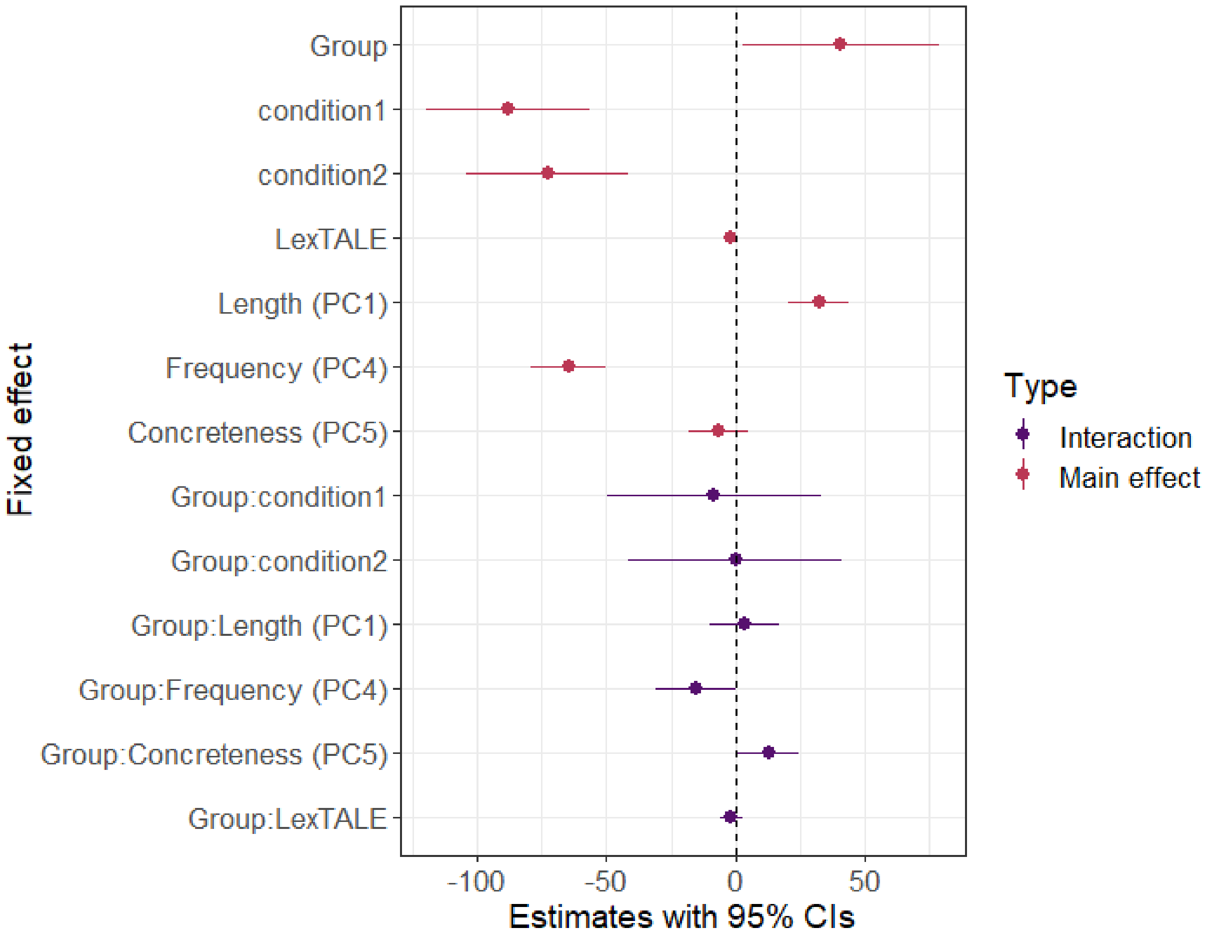

Figure 12 LDT by group reaction time model estimates with 95% Confidence Intervals

4.3.2.1.3 Model comparisons

We then ran model comparisons to examine the significance of the main effects and interactions of interest (Group, Word type and Group: Word type). In the first model comparison a null model with the same random and fixed structure as detailed above was run with the main effect of Group omitted. A Likelihood Ratio Test was run, comparing the maximal linear mixed-effect model for by-group with the null model ($x^2(1) = 1$, p = 1). It is clear there is an optimisation failure (comparing two different models, a p-value of 1 is not feasible). Hence, the interpretation of results should be based on the full model rather than the model comparison here.

The second model comparison was run to examine whether the main effect of word type is significant. A null model without main effects for condition 1 and condition 2 was run and a Likelihood Ratio Test was conducted to compare the null model to the maximal model ($x^2(2)=0$, p=1). Again, this p-value is most probably due to an optimisation failure (see above) and should not be interpreted as an actual effect.

The last model comparison was run to look at the interaction of word type and group. A null model was run with the interaction of group:word type omitted, and this was compared to the full model with a Likelihood Ratio test ($x^2(2) = 0.20$, p= 0.90). The model comparison shows there is no significant interaction of group and word type.

4.3.2.2 By-language analysis of reaction times (Germans only)

4.3.2.2.1 Descriptive statistics

Table 20 Summary of mean RTs (ms) and standard deviations (in brackets) per language and word type

	Neutral	Negative	Positive
German	788.87 (305.85)	741.96 (276.20)	731.52 (254.27)
English	856.36 (349.01)	800.62 (300.84)	785.80 (302.15)

It seems like the reaction times in German (L1) are slightly faster than in English (L2) (Figure 13). Neutral words seem to facilitate the slowest reaction times in both languages, and there does not seem to be a difference in reaction times between the neutral vs. positive/negative words between the languages.

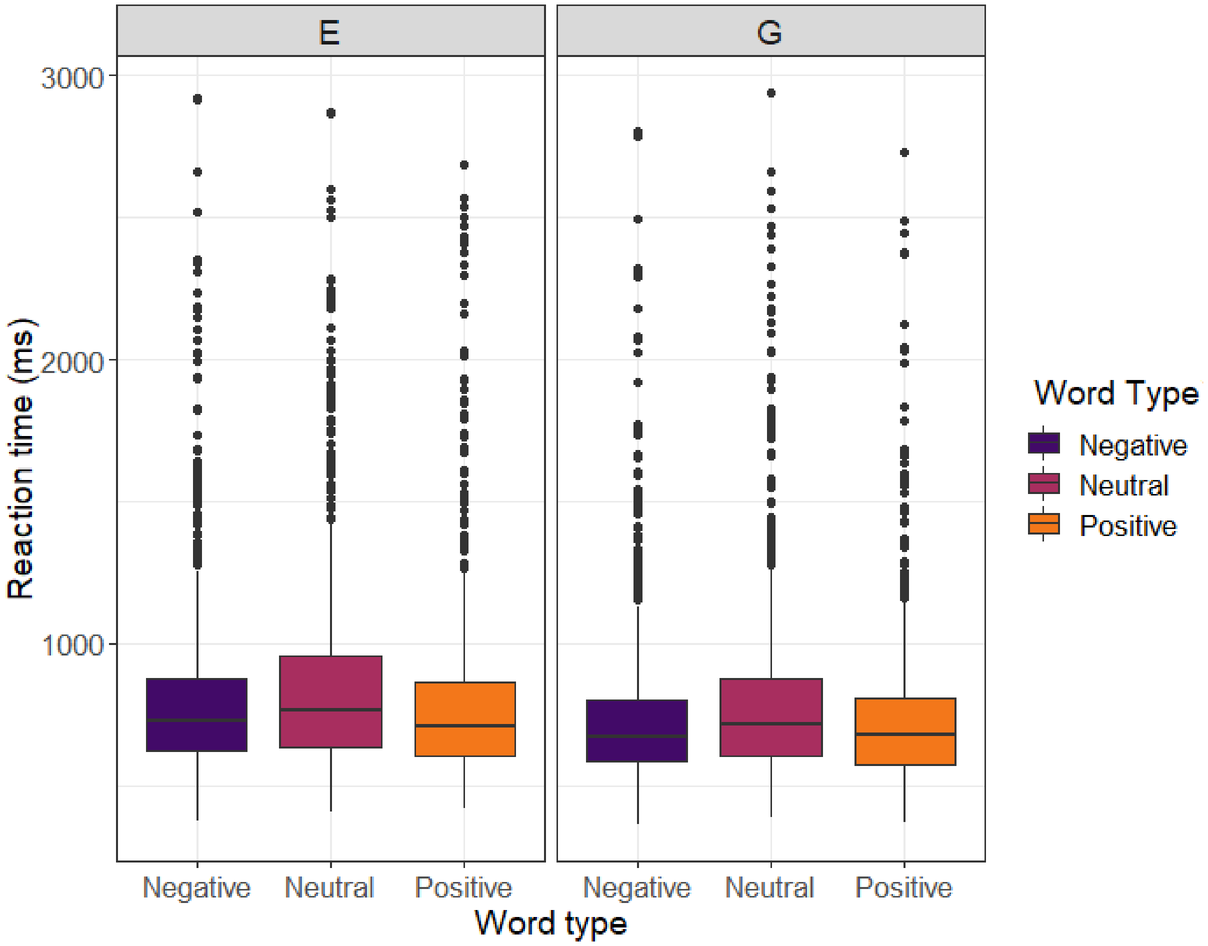

Figure 13 LDT reaction time distributions split by test language and word type

4.3.2.2.2 By-language reaction time Linear Mixed Effects Models

The by-language data were also analysed using linear mixed effects models. Test language and word type were deviation coded for analysis. Word type was coded as detailed in the above by-group section, and the covariates were centered. The first by-language model (see Appendix D for full model syntax) included main effects of test language, word type, LexTALE and the three principal components, namely Length, Frequency and Concreteness of the items. Word type and principal components were also included as fixed interactions with language. In the random structure we included by-subject and by-item random intercepts, by-subject random slopes for word type, language and the interaction of word type and language, and by-item random slope for LexTALE. The model was specified with Gamma family argument, and optimx package L-BFGS-B optimiser.

Table 21 and Figure 14 below summarise the fixed effects of the first by-language model, and their confidence intervals and estimates

Table 21 By-language LDT reaction time model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
Intercept	835.10	10.69	78.08	<0.001***	814.11	856.03
Language	-78.90	15.15	-5.21	<0.001***	-108.59	-49.20
condition1 (positive)	-75.73	13.08	-5.79	<0.001***	-101.37	-50.09
condition2 (negative)	-62.69	12.91	-4.86	<0.001***	-88.01	-37.38
LexTALE	-1.69	0.87	-1.94	0.05.	-3.40	0.02
Length (PC1)	43.58	4.93	8.85	<0.001***	33.93	53.23
Frequency (PC4)	-46.82	5.19	-9.03	<0.001***	-56.99	-36.66
Concreteness (PC5)	-1.33	4.73	-0.28	0.79	-10.61	7.94
Language:condition1	43.07	27.35	1.57	0.12	-10.54	96.69
Language:condition2	28.44	25.68	1.11	0.27	-21.89	78.76
Language:Length (PC1)	11.85	9.86	1.20	0.23	-7.48	31.17
Language:Frequency (PC4)	56.68	10.44	5.43	<0.001***	36.22	77.13
Language:Concreteness (PC5)	0.34	9.47	0.04	0.97	-18.21	18.90

Stimulus-related covariates seem to affect reaction times in a very similar manner as was observed in the by-group models. The main effect of LexTALE is verging on significance, which contrasts our previous findings – however, the effect is marginal.

The first by-language model suggests there is a main effect of language with reaction times in L2 (English) being slower. A main effect of condition (word type) can also be seen: both positive and negative words facilitate faster reaction times in comparison to neutral words, as was observed with the by-group models. As was done in the by-group analysis step, second model was run with different coding of word type; in this model positive words were coded as baseline, and negative words were coded as condition 1 and neutral words were coded as condition 2, respectively. This model suggests there is no significant difference in reaction times between negative and positive words (estimate=13.21, t=1.05, p=0.29), and confirms neutral words have slower reaction times in comparison to positive words (estimate=76.46, t=5.84, p<0.001). There is an interaction of test language and frequency, suggesting that low frequency words slow down reaction times more in participants' L2. Importantly, there is no interaction of word type and language, which suggests no effect of reduced emotional resonance can be observed in participants' L2 here.

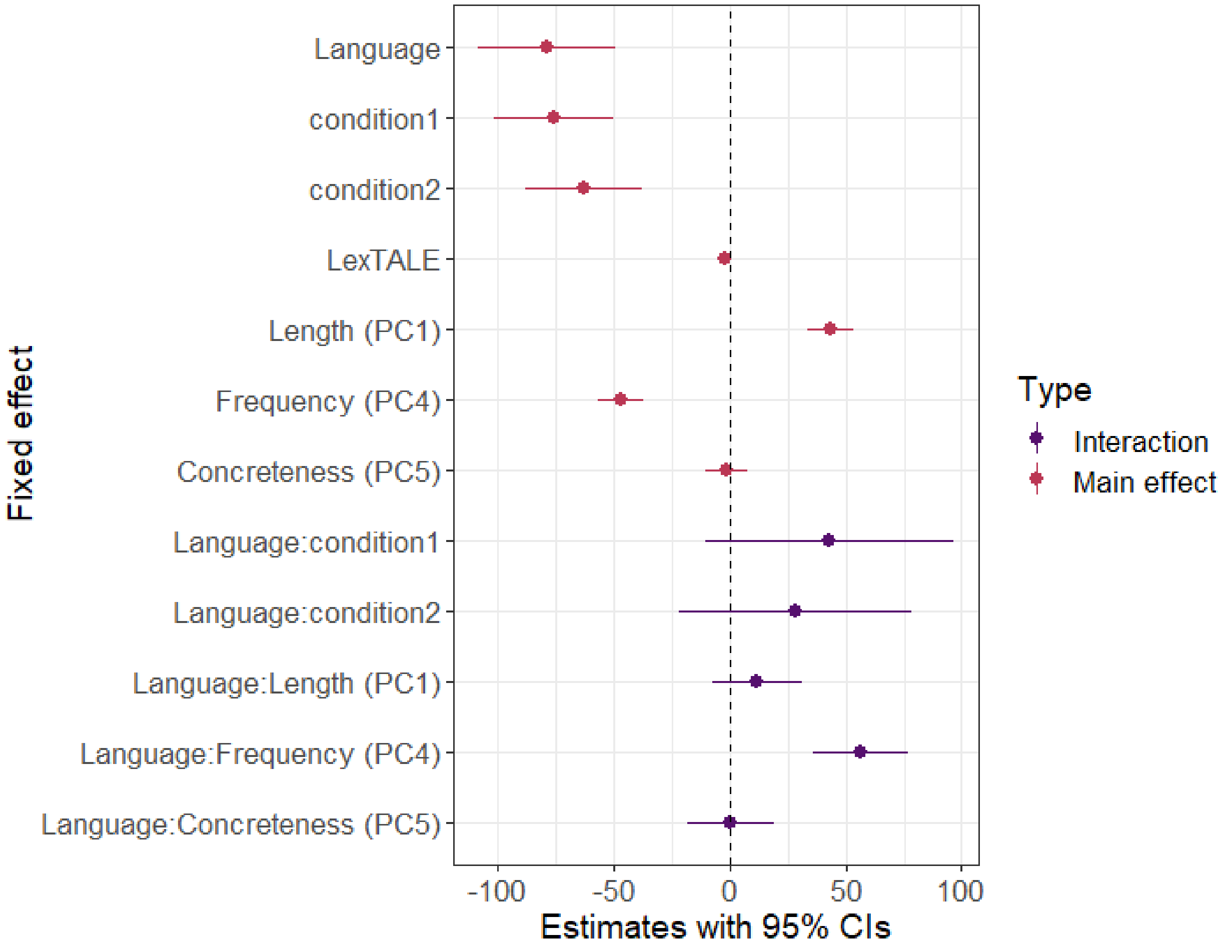

Figure 14 LDT by language reaction time model estimates with 96% Confidence Intervals

4.3.2.2.3 Model comparisons

As detailed in the by-group analysis section above, we ran model comparisons to test the main effects and interactions (language, word type and language: word type). In the first model comparison a null model was run with the main effect of language omitted. A Likelihood Ratio Test was run, comparing the maximal linear mixed-effect model for by-language with the null model (x^2 (1) = 6.45, p=0.01). The first model comparison shows a significant effect of language - German (L1) words have faster reaction times

The second model comparison was run to examine whether the main effect of condition (word type) is significant. A null model without main effects for condition 1 (positive vs. neutral) and condition 2 (negative vs. neutral) was run and a Likelihood Ratio Test was conducted to compare the null model to the full maximal model (x^2 (2)= 15.02, p<0.001). The second model comparison shows a significant effect of word type, suggesting high arousing words (both positive and negative) are processed faster than neutral words.

The third by-language model comparison was run to look at the interaction of word type and language. A null model was run with the interaction of language and word type omitted, and this was compared to the full model with a Likelihood Ratio test (x^2 (2) = 1.09, p=0.58). The model comparison shows there is no significant interaction of group and word type.

4.3.2.3 LDT word recognition accuracy analysis

Generally, participants' word recognition accuracy was very high. On average, L1 English participants were correct 96.8% of the trials, and L1 German speakers were correct 92.3% of the trials in English, and 95.5% of the trials in German. Binary logistic mixed effects models were run to examine which of the predictors in the reaction time models predict word recognition accuracy. The analysis is split into two parts: by-group and by-language. For both parts of the analysis, the binary logistic mixed effects models included the following structures: group (or language), condition (word type), and covariates, namely LexTALE and the three principal components (Length, Frequency and Concreteness) were entered as main effects in the fixed structure, and the covariates were also included as interactions with group (or language). For the by-group models random structure, by-subject and by-item random intercepts were included, word type was entered as by-subject random slope, and group and LexTALE and their interaction were entered as by-item random slopes. For the by-language models random structure, word type and language and their interactions were entered as by-subject random slopes and LexTALE was entered as by-item random slope. Full model syntax can be found on Appendix D. All models were specified with binomial (logit) family argument and the L-BFGS-B optimiser from the optimx package.

4.3.2.3.1 By-group accuracy

Table 22 and blobbogram (Figure 15) below summarise the fixed effects specified in the by-group accuracy model. Both positive and negative words are more likely to be recognised than neutral words. Interestingly, there does not seem to be an effect of group, which suggests the German speakers were highly proficient in English. The same model was run with opposite coding of condition (word type) as detailed in the reaction time by-group model section above. The second model findings suggest word recognition is driven by arousal rather than valence – no difference between negative and positive words was found (estimate= -0.38, z=-0.98, p=0.32), whereas neutral words were less likely to be recognised in comparison to positive words (estimate= -1.47, z= -3.73, p<0.001). High frequency words seem to facilitate better word recognition, but this effect should be interpreted with caution, as the item-related covariates were not included in the random structure.

Table 22 LDT by-group word recognition accuracy model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
Intercept	4.54	0.21	21.63	<0.001 ***	4.13	4.95
Group	-0.59	0.34	-1.73	0.08	-1.27	0.08
condition1 (positive)	1.47	0.37	3.94	<0.001 **	0.74	2.20
condition2 (negative)	1.10	0.36	3.03	0.002**	0.39	1.80
LexTALE	0.006	0.02	0.35	0.72	-0.03	0.04
Length (PC1)	0.18	0.15	1.22	0.22	-0.11	0.47
Frequency (PC4)	1.33	0.21	6.26	<0.001 ***	0.91	1.74
Concreteness (PC5)	-0.13	0.15	-0.91	0.36	-0.42	0.16
Group:condition1	0.37	0.44	0.82	0.41	-0.51	1.24
Group:condition2	-0.11	0.40	-0.27	0.79	-0.90	0.69
Group:Length (PC1)	0.06	0.15	0.36	0.72	-0.25	0.36
Group:Frequency (PC4)	0.46	0.25	1.82	0.07	-0.04	0.96
Group:Concreteness(PC5)	0.20	0.17	1.16	0.25	-0.14	0.54
Group: LexTALE	-0.04	0.03	-1.32	0.19	-0.11	0.02

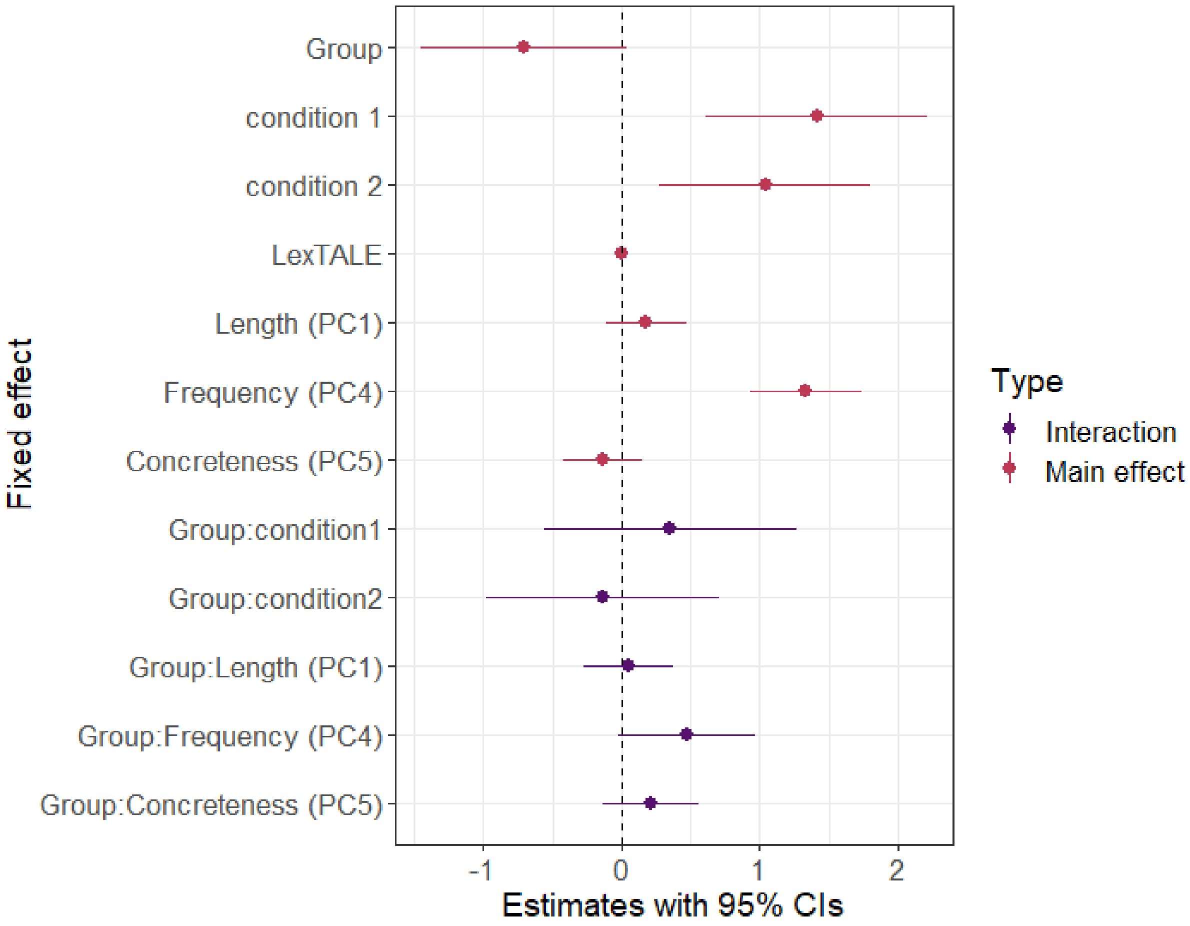

Figure 15 LDT by-group accuracy model estimates with 95% Confidence Intervals

As detailed above, model comparisons were run to examine the effect of group, condition (word type) and the group:condition interaction. The first model comparison was done with the full model as specified above, and a null model from which the main effect of group was omitted. The Likelihood Ratio test (x^2 (1) = 1.84, p=0.17) showed no significant effect of group on word recognition accuracy. Second by-group model comparison looked at the effect of condition (word type) by comparing a null model without the main effect parameters of word type with the full model. The Likelihood Ratio test found an effect of word type (x^2 (2) = 14.17, p<0.001) as suggested by the full model. The third model comparison was run to look at the interaction of group and condition (word type). The interaction terms of word type: group were removed from the null model, which was compared to the full model. No interaction was found (x^2 (2) = 1.11, p=0.57).

4.3.2.3.2 By-language accuracy

Table 23 and blobbogram (Figure 16) below summarise the fixed effects specified in the by-language accuracy model.

The first by-language accuracy model supports findings from the by-group models: both positive and negative words are recognised better than neutral words. We recoded word type as detailed above, and found no significant difference between positive and negative words (estimate =0.30, z= 0.81, p= 0.42), while positive words were recognised better than neutral words (estimate =-0.85, z= -2.46, p= 0.01). There is a main effect of language; words in L1 facilitate better word recognition accuracy. There also seems to be an interaction of condition 1 (positive words compared to neutral words) and language. The full model was run again with dummy coding of test language to decompose this interaction. The dummy coded model suggests the effect of positive words facilitating word recognition is smaller in L1, in comparison to L2 (estimate= -2.44, z= -3.47, p<0.001). More frequent and shorter words facilitate word recognition accuracy. Again, any effects with the stimuli covariates should be interpreted with caution due to not including them in the random effects structure.

Table 23 LDT by-language word recognition accuracy model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
Intercept	4.59	0.22	20.86	<0.001 ***	4.16	5.02
Language	0.86	0.36	2.41	0.02*	0.16	1.55
condition1 (positive)	0.85	0.35	2.45	0.01*	0.17	1.53
condition2 (negative)	1.17	0.36	3.22	0.001**	0.46	1.88
LexTALE	-0.02	0.01	-1.12	0.26	-0.04	0.01
Length (PC1)	0.31	0.14	2.22	0.03*	0.04	0.59
Frequency (PC4)	0.98	0.16	6.15	<0.001 ***	0.67	1.29
Concreteness (PC5)	-0.16	0.14	-1.21	0.23	-0.43	0.10
Language:condition1	-2.43	0.71	-3.44	<0.001***	-3.81	-1.04
Language:condition2	0.20	0.75	0.26	0.79	-1.27	1.66
Language:Length (PC1)	0.09	0.28	0.34	0.74	-0.46	0.65
Language:Frequency (PC4)	-1.27	0.32	-3.99	<0.001 ***	-1.89	-0.65
Language:Concreteness (PC5)	-0.23	0.27	-0.86	0.39	-0.76	0.30

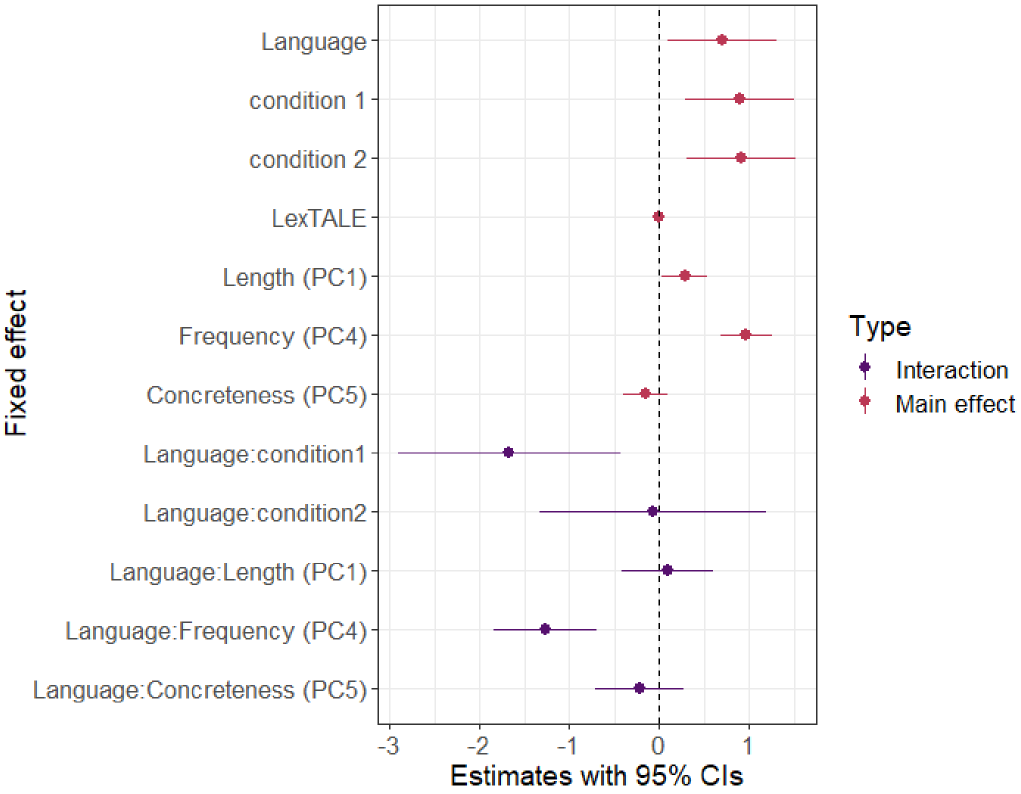

Figure 16 LDT by-language accuracy model estimates with 95% Confidence Intervals

Model comparisons were run to look at the effects of language, word type and the interaction of language and word type. The first null model that was compared to the full model had the main effect of language removed. The model comparison confirms there is an effect of language (x^2 (1)=5.10, p=0.02). The second null model had the main effect of condition (word type) removed, and the model confirms there was a main effect of word type (x^2 (2)=8.71, p=0.01). The last null model that was compared to the full model had the interaction of word type and language removed. This model showed a significant interaction of test language and word type (x^2 (2)=13.58, p=0.001).

4.3.3 Eye-tracking task

4.3.3.1 Pre-processing of pupil data

The procedure from Experiment 1 was used for pupil data pre-processing (see section 2.3.7). However, as the data appeared to have more noise than experiment 1, additional pre-processing steps were added. Time bins (within one trial and participant) where the eye-position was a Euclidean distance of larger than 68 pixels away from the centre (2 degrees of visual angle) were declared as missing values. Time bins where the absolute difference in pupil size in the time bin and the following time bin (within the same trial and participant) was larger than 0.041 were also declared as missing values. We then interpolated over these missing values to replace them with a mean of the previous and the next valid value. If the first or last time bin of a trial was a missing value, this was replaced with the next or the previous valid value, respectively. After interpolation data were transformed into area under the curve values (for more details check Experiment 1 Pupil data pre-processing section 2.3.7).

Figure 17 shows pupil change averaged across time and word type, split by speaker group (monolingual vs. bilingual speakers). From this graph it appears neutral words elicited the strongest pupil response in both groups. Interestingly, it also seems that the monolingual speaker group did not differ in their responses to negative versus positive words, whereas in the German speaker group negative words facilitate stronger pupillary response. L1 English speakers overall seem to have a smaller pupillary response, which would be consistent with the cognitive effort effect on pupil response.

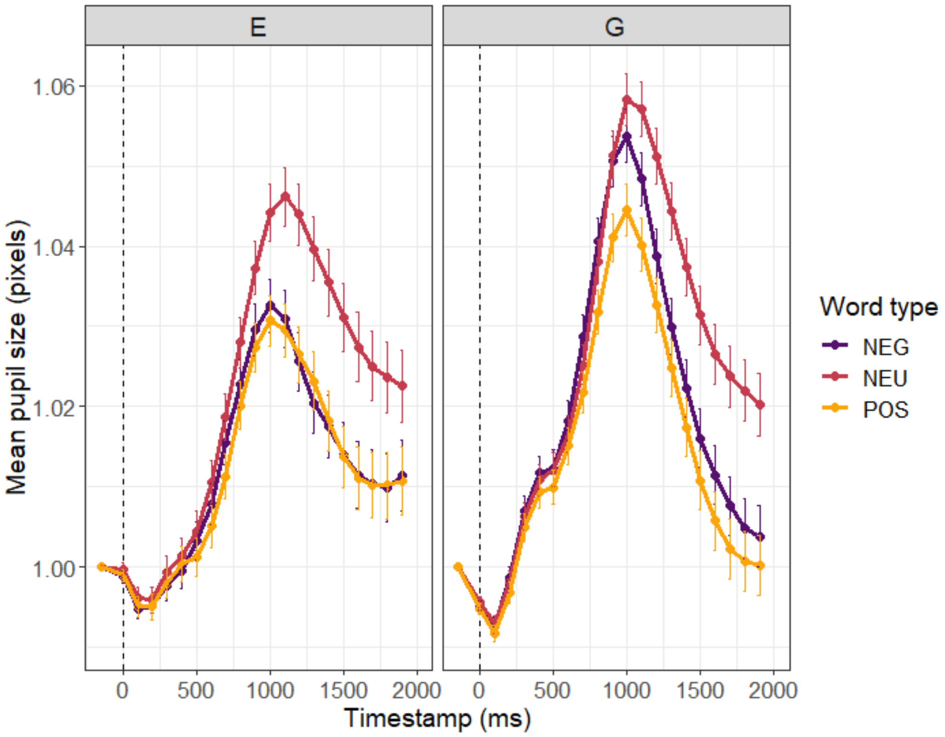

Figure 17 Pupil change across time, split by word type (by-group)

Figure 18 shows pupil change averaged across time and word type, split by test language (German speakers only, tested in English vs. tested in German). The pupil response seems stronger when the bilinguals are tested in L2 (English), which again lends support to increased cognitive effort. Here, the pattern seems identical across languages with neutral words facilitating the strongest responses and negative words facilitating stronger responses than positive words.

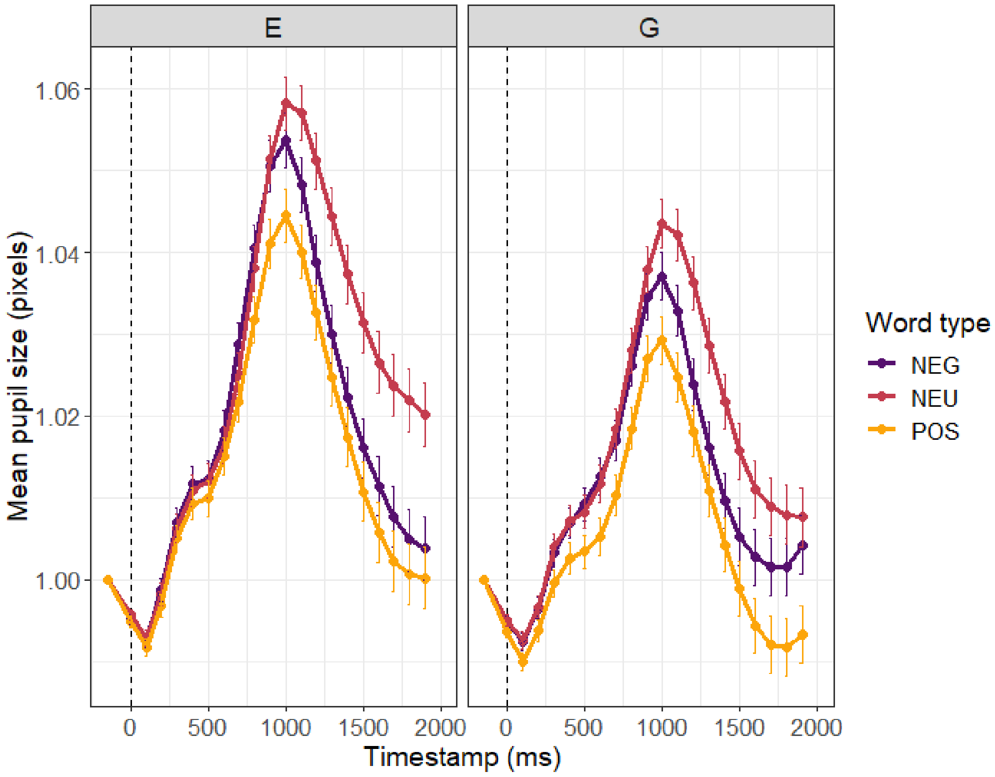

Figure 18 Pupil change across time split by word type (by-language)

4.3.3.2 Pupil data descriptive statistics

Figure 19 shows the distribution of area under the curve values split by group (monolingual speakers and bilingual speakers, English data only). Figure 20 shows the distribution split by test language. All the distributions are slightly right skewed. Due to skewness, pupillary mixed effects models will be specified with gamma family argument. These distributions only include trials where participants indicated that they recognised the word.

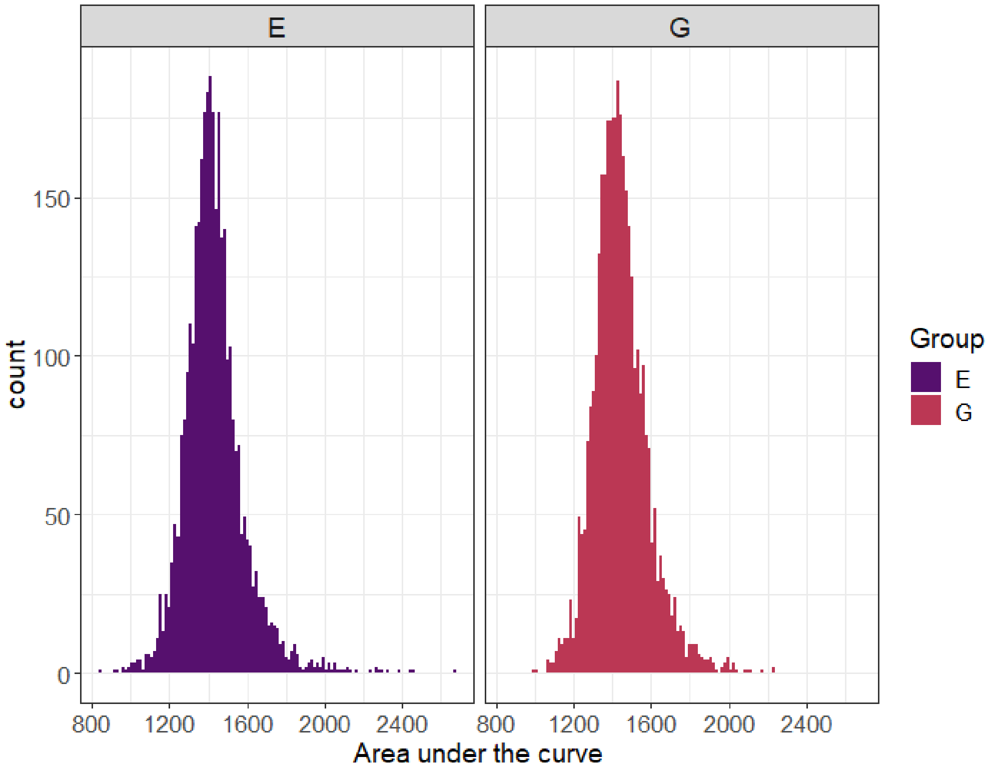

Figure 19 Area under the curve distribution by group

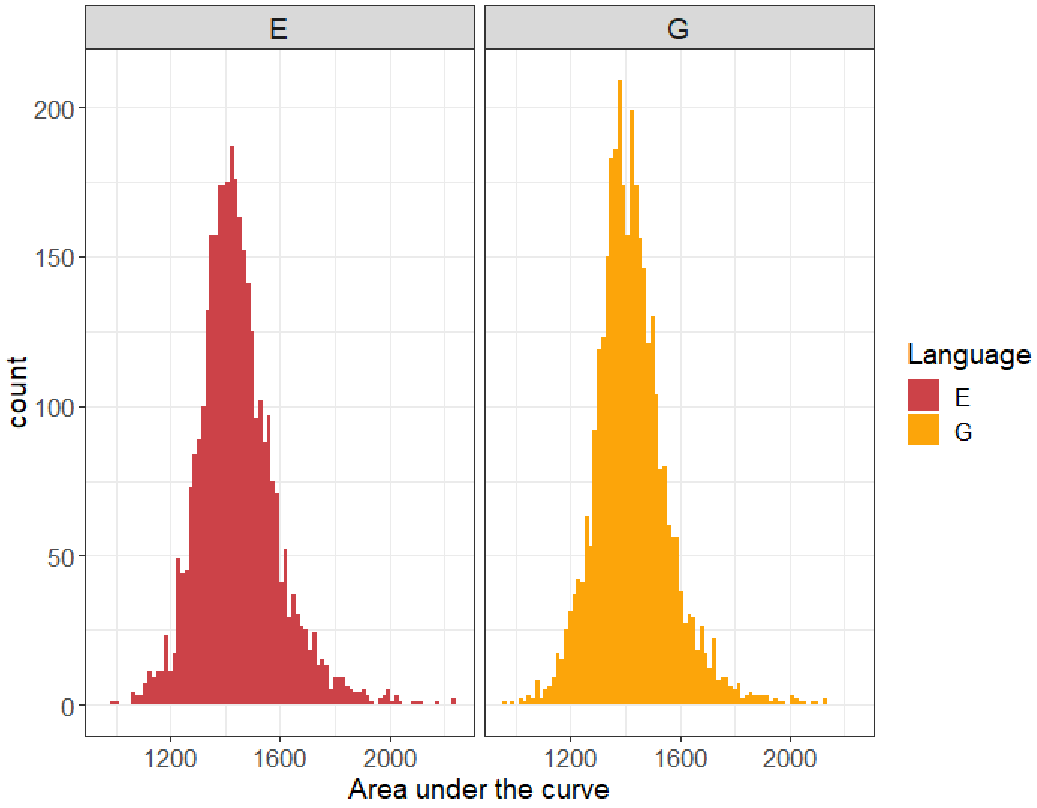

Figure 20 Area under the curve distribution by language

4.3.3.3 Analysis

The analysis is split into four parts like the Lexical Decision Task analysis: by-group analysis of pupillary response (both groups), by-language analysis of pupillary response (Germans only comparing L1 and L2), by-group analysis of button response (both groups) and by-language analysis of button response (Germans only, comparing L1 and L2). Trials where participants indicated they did not recognise the word were excluded from the pupillary response area under the curve analyses.

The data were analysed using linear mixed effects models, Gamma family argument. Due to convergence issues with the lme4 in-built optimiser functions, the L-BFGS-B optimiser from optimx package was used (specifications can be found on the model list in Appendix D).

4.3.3.3.1 By-group pupillary response Linear Mixed Effects Models

Participant group and word type were deviation coded for analysis. Word type was coded with distractor words as the “baseline” and the deviation coding of positive and negative words were scaled to this baseline as separate variables (condition1=positive vs. baseline and condition 2=negative vs. baseline). LexTALE scores, as well as the principal components used as item covariates were centered, as was done for the LDT analyses. The full by-group model (Appendix D) included main effects of group, word type, and LexTALE and Length, Frequency and Concreteness of the items as covariates. Word type and principal components were also included as fixed interactions with group. In the random structure, we included random intercepts for subject and item, word type was included as a by-subject random slope, and group and LexTALE and their interaction were included as by-item random slopes.

Table 24 and blobbogram (Figure 21) below summarise the fixed effects, their confidence intervals and estimates of the full by-group model.

Table 24 By-group pupillary response model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
Intercept	1441.79	4.39	328.62	<0.001 ***	1433.19	1450.39
Group	-2.28	8.32	-0.27	0.78	-18.60	14.03
condition1 (positive)	-17.04	6.62	-2.57	0.01*	-30.02	-4.06
condition2 (negative)	-9.91	6.31	-1.57	0.12	-22.28	2.47
LexTALE	-2.04	0.45	-4.53	<0.001 ***	-2.92	-1.15
Length (PC1)	-3.15	2.40	-1.31	0.19	-7.85	1.55
Frequency (PC4)	-10.68	2.90	-3.68	<0.001 ***	-16.37	-4.99
Concreteness (PC5)	0.29	2.31	0.13	0.90	-4.23	4.81
Group:condition1	14.12	11.38	1.24	0.21	-8.18	36.42
Group:condition2	2.73	10.67	0.26	0.80	-18.19	23.65
Group:Length (PC1)	-1.34	3.91	-0.34	0.73	-9.01	6.33
Group:Frequency (PC4)	5.70	4.70	1.21	0.22	-3.50	14.90
Group:Concreteness (PC5)	-1.97	3.71	-0.53	0.60	-9.25	5.31
Group:LexTALE	-2.27	0.91	-2.49	0.01*	-4.05	-0.49

Lower-frequency words elicited larger pupillary responses, which is consistent with cognitive effort increasing pupillary response. In contrast to the LDT findings, there is no main effect of group in pupillary response. It should be noted these effects are not entirely reliable, given that the item-related covariates were not included in the random effect structure of the models. Interestingly, there seems to be a strong main effect of LexTALE proficiency score, and an interaction of group and LexTALE score, suggesting that lower LexTALE scores in the German group predict higher pupillary response (ß= -2.43, SE=0.91, t=- 2.67, p=0.008), consistent with the cognitive effort effect previously found on pupil experiments. In the monolingual group, however, higher LexTALE scores predict higher pupillary response (ß=2.43, SE=0.91, t= -2.70, p=0.007), which is somewhat unexpected.

A main effect of condition (word type) can also be seen: both positive and negative words elicit smaller pupillary responses in comparison to neutral words, although the difference between negative and neutral words is smaller. To further examine the effects of word type, second model was run with different coding of word type; in this model positive words were coded as baseline. Negative words were coded as condition 1, and neutral words were coded as condition 2. This model suggests both negative (estimate=7.44, t=1.16, p=0.25) and neutral (estimate=17.35, t=2.62, p=0.009) words elicit larger pupillary responses than positive words. However, the difference between positive and negative words is not significant.

The blobbogram below (Figure 21) shows fixed effects estimates and their 95% confidence intervals. Condition 1 shows the difference of negative words to neutral words, and condition 2 shows the difference of positive words to neutral words. As the blobbogram shows, only word frequency, LexTALE scores and condition 1 seem to affect pupillary response.

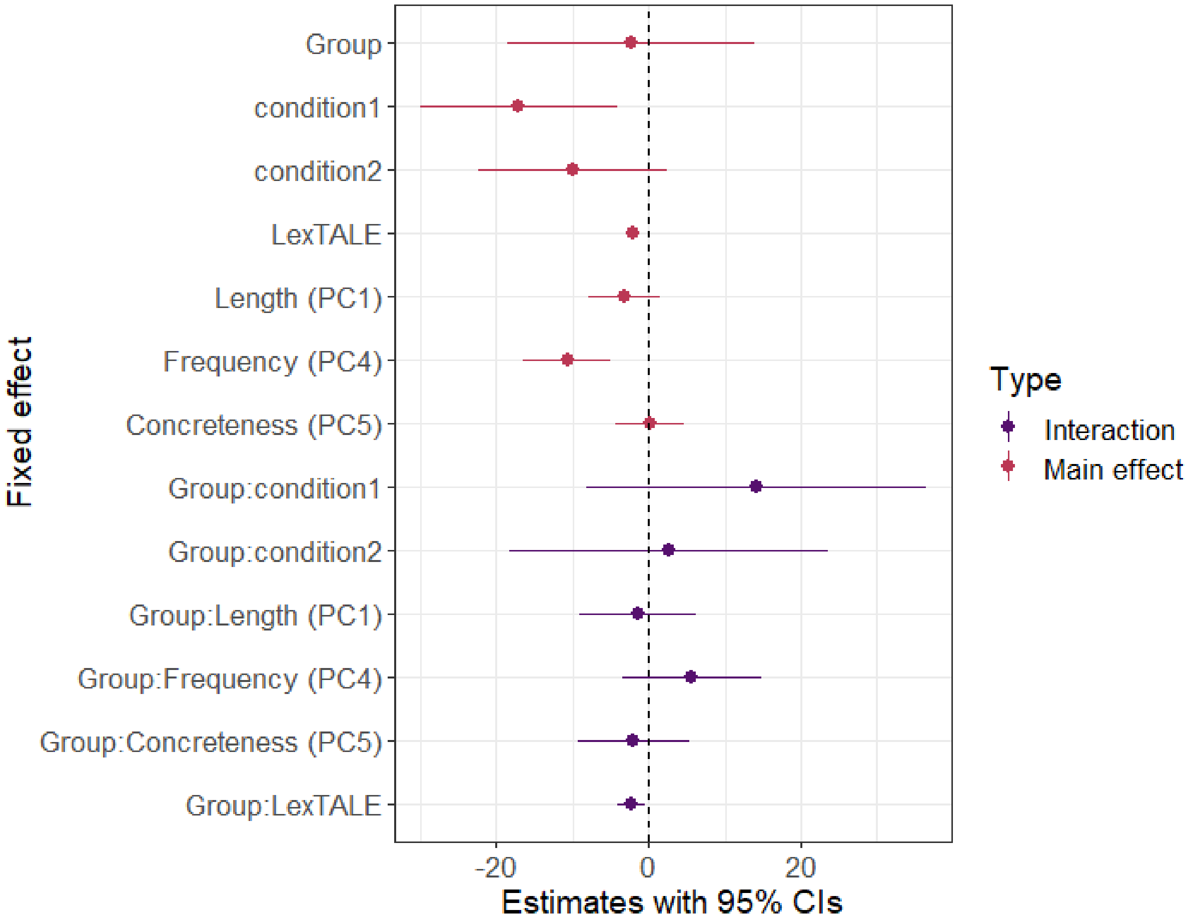

Figure 21 By-group pupillary response model estimates with 95% Confidence Intervals

4.3.3.3.2 Model comparisons

Model comparisons were ran to examine the main effects and interactions of interest (Group, word type and Group: word type). All three model comparison failed, likely due to an optimisation error, and the p-values cannot be trusted. Hence, any interpretation of results should be based on the full model instead.

In the first model comparison a null model with the same random and fixed structure as detailed above was run with the main effect of Group omitted. A Likelihood Ratio Test was run, comparing the maximal linear mixed-effect model for by-group with the null model (x^2 (1) = 0, p=1).

Next we examined the main effect of word type. A null model without main effects for condition 1 and condition 2 was run and a Likelihood Ratio Test was conducted to compare the null model to the maximal model (x^2 (2)= 0, p=1).

The last model comparison was run to look at the interaction of word type and group. A null model was run with the interaction of group and word type omitted, and this was compared to the full model with a Likelihood Ratio test (x^2 (2) = 0, p=1). The model comparison shows there is no significant interaction of group and word type.

4.3.3.3.3 By-language pupillary response Linear Mixed Effects Models

By-language pupil data were analysed using linear mixed effects models. Test language and word type were deviation coded for analysis. Word type was coded with neutral words as the “baseline” and the deviation coding of positive and negative words were scaled to this baseline as separate variables (condition1=positive and condition2=negative). LexTALE scores, as well as the principal components used as item covariates were centered. The full by-group model (Appendix D) included main effects of language, word type, and LexTALE and Length, Frequency and Concreteness of the items as covariates. Word type and the principal components were also included as fixed interactions with language. In the random structure we included random intercepts for item and subject, a by-subject random slopes for word type, language and the interaction of word type and language, and a by-item random slope for LexTALE. The model was specified with Gamma family argument.

Table 25 By-language pupillary response model summary

Term	Estimate	Std.error	z-value	p-value	95% CI low	95% CI high
Intercept	1436.10	4.35	329.91	< 0.001 ***	1427.56	1444.63
Language	-23.93	6.35	-3.77	< 0.001***	-36.39	-11.48
condition1 (positive)	-21.52	5.99	-3.59	< 0.001***	-33.28	-9.76
condition2 (negative)	-10.26	5.63	-1.82	0.07	-21.29	0.78
LexTALE	-0.81	0.36	-2.23	0.03*	-1.52	-0.10
Length (PC1)	-2.29	2.12	-1.08	0.28	-6.45	1.86
Frequency (PC4)	-8.61	2.25	-3.83	<0.001***	-13.01	-4.20
Concreteness (PC5)	-0.09	2.05	-0.04	0.96	-4.10	3.92
Language:condition1	3.15	10.51	0.30	0.76	-17.44	23.75
Language:condition2	5.09	10.30	0.49	0.62	-15.09	25.27
Language:Length (PC1)	-0.22	4.24	-0.05	0.96	-8.52	8.09
Language:Frequency (PC4)	9.34	4.48	2.08	0.04*	0.55	18.13
Language:Concreteness (PC5)	-3.29	4.09	-0.80	0.42	-11.32	4.73

There is a main effect of language; in German-English bilinguals, English (L2) words elicited larger pupillary response in general as opposed to words presented in German (L1). Again, both positive and negative words elicited a smaller pupillary response in comparison to neutral words, although the difference between negative and neutral words is marginal. There is no interaction of language and word type. There is also a main effect of frequency: lower-frequency words elicit stronger pupillary responses. There is a main effect of LexTALE but again, the effect is marginal.

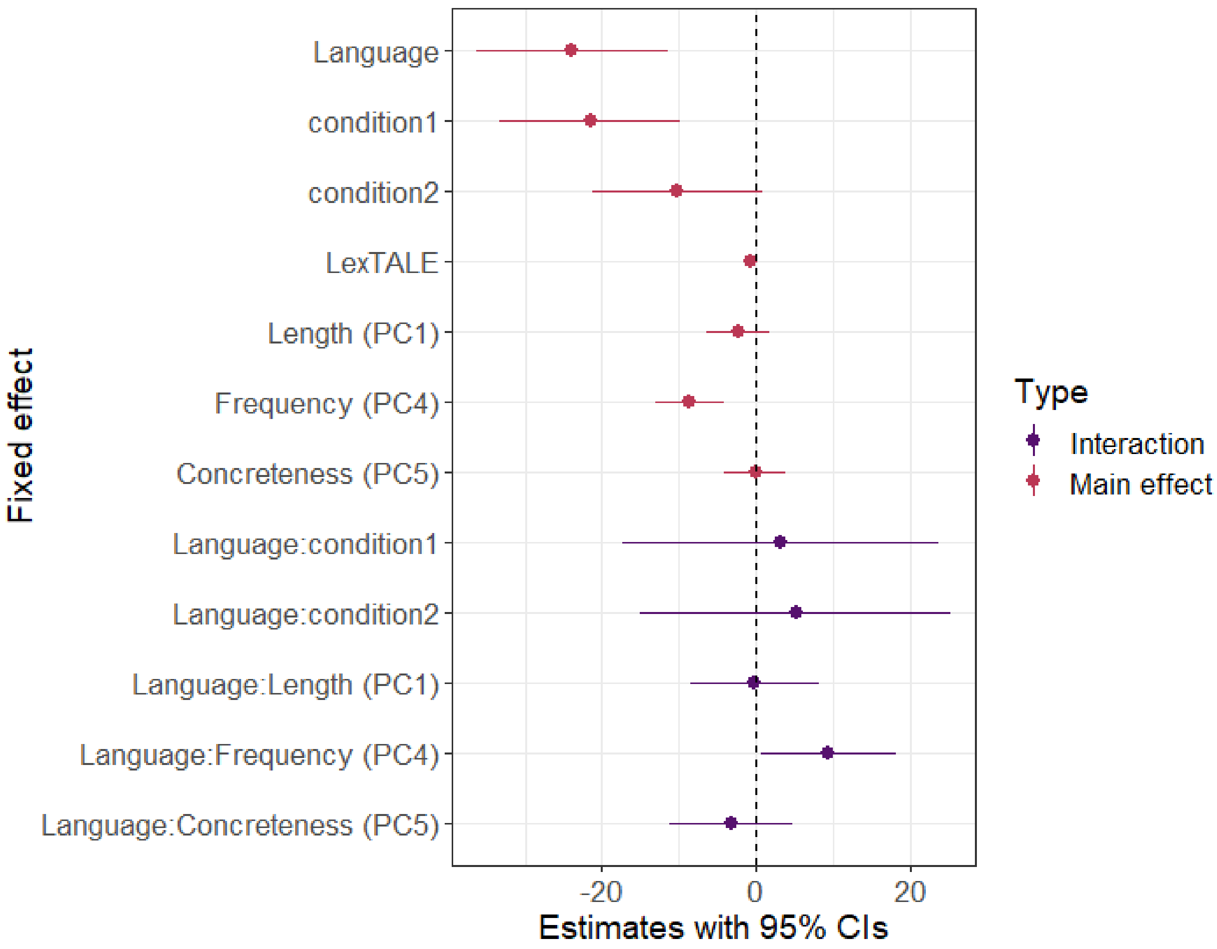

Figure 22 By-language pupillary response model estimates with 95% Confidence Intervals

To further examine the word type effect, we recoded word type. In the new model positive words were coded as baseline. Negative words were coded as condition 1, and neutral words were coded as condition 2. This model suggests that both negative (estimate=11.54, t=6.01, p=0.06) and neutral (estimate=21.72, t=5.99, p<0.001) words elicit larger pupillary responses than positive words. However, the difference between positive and negative words is marginal.

4.3.3.3.4 Model comparisons

Model comparisons were ran to examine the main effects and interactions of interest (Language, word type and Language: word type). In the first model comparison a null model with the same random and fixed structure as detailed above was run with the main effect of Language omitted. A Likelihood Ratio Test was run, comparing the maximal linear mixed-effect model for by-language with the null model (x^2 (1) = 3.50, p=0.06). The first model comparison shows there is no significant effect of language but the trend is there, suggesting words in L2 (English) elicited larger pupil response.

Next, we examined the main effect of word type. A null model without main effects for condition 1 and condition 2 was run and a Likelihood Ratio Test was conducted to compare the null model to the maximal model (x^2 (2)= 5.18, p=0.08). There is a marginal, non-significant effect of word type.

The last model comparison was run to look at the interaction of word type and group. A null model was run with the interaction of group and word type omitted, and this was compared to the full model with a Likelihood Ratio test (x^2 (2) = 0.12, p=0.94). The model comparison shows there is no significant interaction of group and word type.

4.3.3.4 Eye-tracking word recognition task

In the eye-tracking experiment, participants had a word recognition task after each trial (different to the LDT; participants were simply asked to indicate whether they recognised/know each word). Participants' word recognition accuracy in the eye-tracking task was high. On average, L1 English participants recognised 98% (ranging from 95% to 100%) of the trials, and German speakers recognised 91% of the trials in English (ranging from 65% to 99%), and 98% of the trials in German (ranging from 87% to 100%).

Binomial linear mixed effects models were run to examine which of the predictors in the pupil models predict word recognition. The analysis is split into two parts: by-group and by-language. For both parts of the analysis, the linear mixed effects models included the following structures: group (or language), condition (word type), and covariates, namely LexTALE and the three principal components (Length, Frequency and Concreteness) were entered as main effects in the fixed structure, and the covariates were also included as interactions with group (or language). For the by-group models random structure, random intercepts were included for subject and item, word type was entered as by-subject random slope, and group and LexTALE were entered as by-item random slopes. For the by-language models random structure, in addition to the random intercepts for subject and item, word type and language and their interactions were entered as by-subject random slopes, and LexTALE was entered as by-item random slope. Full model syntax can be found in Appendix D. All models were specified with binomial (logit) family argument, and optimx package L-BFGS-B optimiser.

4.3.3.4.1 By-group word recognition

The table (26) and blobbogram (Figure 23) below summarise the fixed effects specified in the by-group accuracy model. There is a main effect of group: monolingual speakers were more likely to recognise the words than German speakers. High frequency words were more likely to be recognised, and there is also an interaction of group and frequency. Higher LexTALE scores predict higher word recognition accuracy, however the effect is very small. Both positive and negative were recognised better than neutral words. The same model was run with opposite coding of condition (word type) to examine the word type differences further. The second model findings align with the equivalent LDT findings. They suggest word recognition is driven by arousal rather than valence – no difference between negative and positive words was found (estimate=-0.10, z=-0.21, p=0.83), whereas neutral words were less likely to be recognised in comparison to positive words (estimate=-1.60, z=-3.80, p<0.001).

Table 26 By-group pupil experiment word recognition model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
Intercept	4.94	0.23	21.44	< 0.001 ***	4.49	5.39
Group	1.39	0.38	3.70	< 0.001***	0.65	2.13
condition1 (positive)	1.56	0.42	3.73	< 0.001***	0.74	2.37
condition2 (negative)	1.47	0.43	3.45	< 0.001***	0.63	2.30
LexTALE	0.06	0.02	3.27	0.001**	0.02	0.09
Length (PC1)	-0.24	0.16	-1.52	0.13	-0.55	0.07
Frequency (PC4)	1.22	0.23	5.26	< 0.001 ***	0.76	1.67
Concreteness (PC5)	-0.02	0.17	-0.14	0.89	-0.36	0.31
Group:condition1	0.22	0.55	0.40	0.69	-0.86	1.30
Group:condition2	0.61	0.57	1.09	0.28	-0.50	1.72
Group:Length (PC1)	-0.32	0.18	-1.79	0.07	-0.66	0.03
Group:Frequency (PC4)	-0.81	0.30	-2.69	0.007**	-1.40	-0.22
Group:Concreteness(PC5)	0.06	0.23	0.25	0.81	-0.40	0.51
Group:LexTALE	0.03	0.03	1.09	0.27	-0.03	0.10

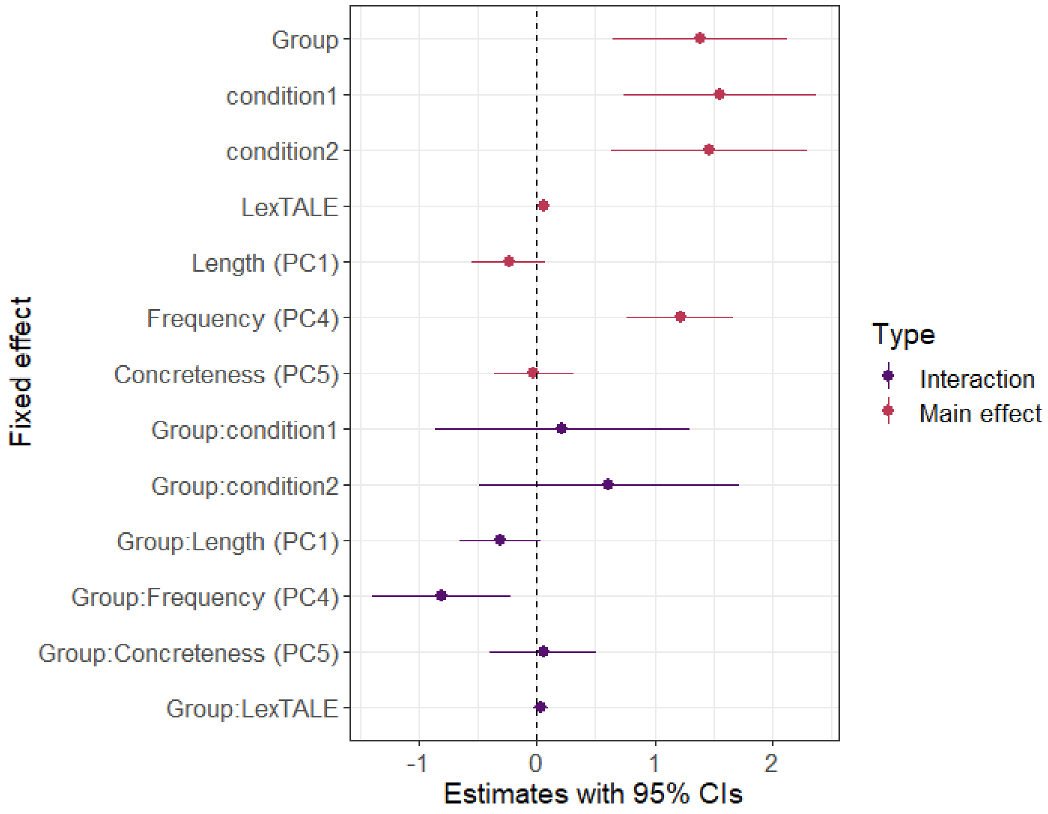

Figure 23 By-group pupil experiment word recognition model estimates with 95% Confidence Intervals

As detailed above, model comparisons were run to examine the effect of group, word type and the group: word type interaction. The first model comparison was done with the full model as specified above, and a null model from which the main effect of group was omitted. The Likelihood Ratio test (x^2 (1) =9.80, p=0.002) confirms there was a main effect of group; monolingual speakers recognised more of the words than German speakers did. Second by-group model comparison looked at the effect of condition (word type) by comparing a null model without the parameters for word type with the full model. The Likelihood Ratio test found an effect of word type (x^2 (2) =15.87, p<0.001) as suggested by the full model – both positive and negative words were recognised more often than neutral words. The third model comparison was run to look at the interaction of group and condition (word type). The interaction terms of word type: group were removed from the null model, which was compared to the full model. No interaction was found (x^2 (2) = 1.74, p=0.42).

4.3.3.4.2 By-language word recognition

The table (27) and blobbogram (Figure 24) below summarise the fixed effects specified in the by-language accuracy model. German (L1) words were more likely to be recognised in comparison to English (L2) words. Rather unsurprisingly, as found in the other analyses, high frequency words were more likely to be recognised than low frequency words, and there was an interaction of language and frequency, showing that the effect of frequency is stronger in L2. Both positive and negative words were recognised more often than neutral words. We recoded condition (word type) as detailed above, and found no difference between positive and negative words (estimate =-0.11, z= -0.31, p=0.76), while positive words were more likely to be recognised than neutral words (estimate = -0.88, z=-2.62, p=0.009).

Table 27 By-language pupil experiment word recognition model summary

Term	Estimate	Std.error	z-value	p-value	95% CI low	95% CI high
Intercept	4.63	0.21	21.54	< 0.001 ***	4.21	5.05
Language	1.62	0.38	4.22	< 0.001 ***	0.87	2.38
condition1 (positive)	0.87	0.34	2.58	0.009**	0.21	1.53
condition2 (negative)	0.75	0.34	2.19	0.03*	0.08	1.42
LexTALE	0.03	0.01	1.80	0.07	-0.002	0.05
Length (PC1)	-0.15	0.13	-1.13	0.26	-0.41	0.11
Frequency (PC4)	0.91	0.15	6.07	< 0.001 ***	0.61	1.20
Concreteness (PC5)	-0.08	0.13	-0.57	0.57	-0.34	0.19
Language:condition1	-1.03	0.66	-1.56	0.11	-2.32	0.27
Language:condition2	-0.79	0.68	-1.17	0.24	-2.12	0.54
Language:Length (PC1)	-0.15	0.26	-0.58	0.56	-0.67	0.37
Language:Frequency (PC4)	-1.37	0.30	-4.61	< 0.001 ***	-1.96	-0.79
Language:Concreteness (PC5)	-0.03	0.27	-0.11	0.91	-0.56	0.50

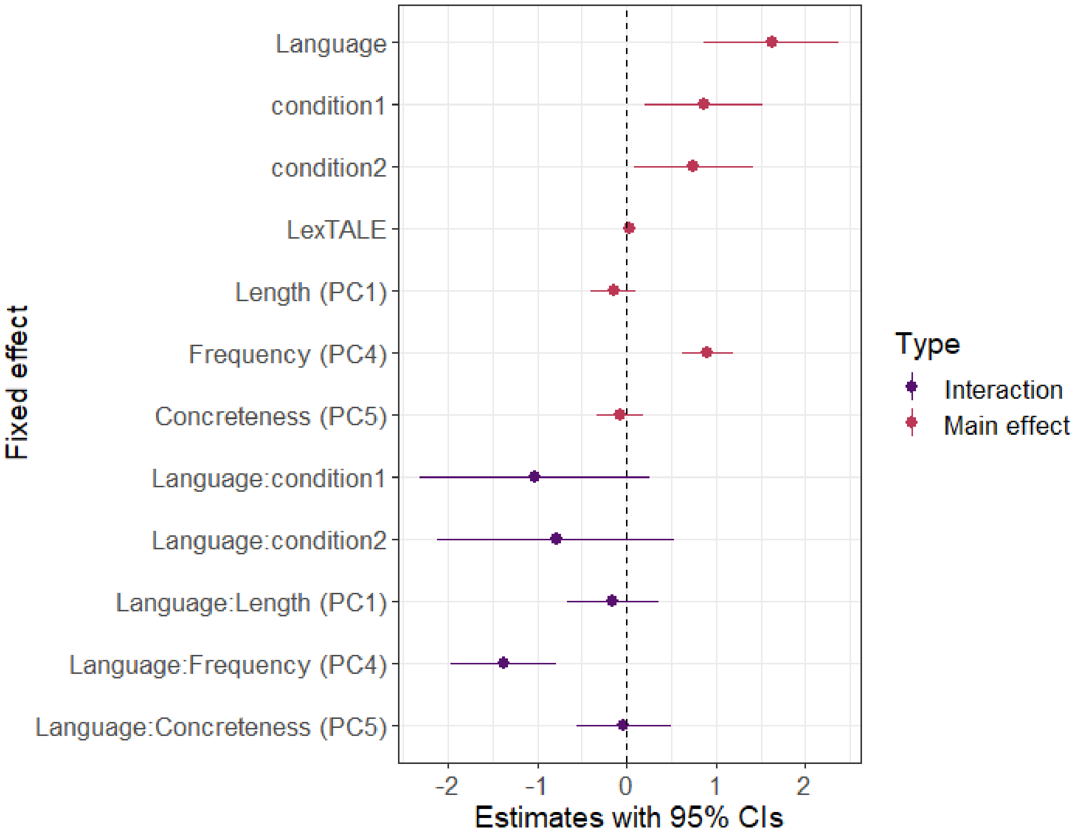

Figure 24 By-language pupil experiment word recognition model estimates with 95% Confidence Intervals

Model comparisons were run to look at the effects of language, word type and the interaction of language and word type. The first null model that was compared to the full model had the main effect of language removed. The model comparison confirms there is an effect of language (x^2 (1)=14.99, p<0.001) – words in German were more likely to be recognised than words in English (L2). The second null model had the main effect of condition (word type) removed, and the model shows there was no main effect of word type (x^2 (2)=4.64, p=0.10). The last null model that was compared to the full model had the interaction of word type and language removed (x^2 (2)=10.71, p=0.005). This null model shows an interaction of language and word type. Suggesting that high arousing words were better recognised in L1 as opposed to L2. Interestingly, this effect did not come out in the full model detailed above.

4.4 Discussion

This experiment had three aims:

1. To investigate whether a Lexical Decision Task will detect reduced emotional resonance in bilingual participants. This was done with both cross-group and cross-language comparison: we compared a monolingual group to a bilingual group, and bilingual participants' L1 to their L2.

2. To compare findings from the LDT to a pupillometry task on the same participants, using the same stimuli.

3. To investigate whether there is a positivity bias in the bilingual participants' pupillary responses. To this end, we split the stimulus word categories differently than in our previous experiments. We hypothesised there would not be a positivity bias in the pupillary response; we expected it to be driven by arousal rather than by valence.

The Lexical Decision Task did not detect reduced emotional resonance in our bilingual sample (neither in cross-group comparisons nor cross-languages comparison) – the model comparisons showed that the word type * group/language interactions were not significant. This was somewhat expected, as cognitive paradigms generally are not reliable in detecting reduced emotional resonance of L2.

In both by-language and by-group comparisons we found that high arousing words were processed the fastest, but also that positive words speed up reaction times in comparison to negative and neutral words (however, the difference between positive and negative words was marginal, and was not significant in any of the models). The main effects found in the LDT align to previous literature on affective word processing, suggesting that the effects are more strongly driven by arousal, rather than valence (Kousta et al., 2009; Kuperman et al., 2014). However, some studies have found that HA words are processed slower (Kuperman et al., 2014).

In addition to reaction times, word recognition was also facilitated by word type: high arousing words (both positive and negative) had a higher recognition likelihood in comparison to low arousing words, consistently across both the LDT word recognition analyses as well as in the word recognition task after each pupillometry trial. This is in line with our previous findings (see Experiment 1). In the pupil experiment word recognition task, it was also found that HA words were recognised more often in one's L1, which is the only (tentative) evidence of reduced emotional resonance found in this experiment.

In terms of the covariates, we found that length and word frequency had a significant effect on reaction times (such that longer and lower frequency words slow down reaction times). This is consistent with previous research suggesting that word length and frequency affect word processing (Ferrand et al., 2011). There was also an interaction of participant group and frequency, and group and concreteness in the by-group model, and an interaction of frequency and language in the by-language comparison. This suggests that even in a carefully controlled stimulus set, where the covariates are algorithmically balanced, they can affect word-processing and should be included in the analysis.

The LDT findings fall in line with previous studies which did not detect reduced emotional resonance with a Lexical Decision Paradigm (for example, Conrad et al., 2011; Ponari et al., 2015). We found no word type: group/language interactions in the RTs. However, as discussed in the introduction, a number of experiments have found reduced facilitation effects in L2 (e.g. see Chen et al, 2015), or less affective priming in L2 (Altarriba & Canary, 2004). These findings have been interpreted as reduced emotionality in L2 affecting how emotion facilitates word recognition and access. There does not seem to be a clear pattern of what predicts whether a cognitive measure, Lexical decision Task or other, will work to this end. Studies which have employed cognitive paradigms have a varying number of covariate items included, use a number of different statistical techniques, and test a varying number of participants, in different bilingual-monolingual and cross-language compositions, with a varying number of stimuli items. The tests simply seem inconsistent.

There are several possible explanations for this unpredictability. Either the tests do not capture reduced emotional resonance very well, or some of the bilingual samples would not have shown the effects regardless of the measurement technique. This is hard to assess, given that most of the experiments have been run simply to explore lexical properties of words, or to detect the effect of emotion, and not with a strict methodological focus. Variation in samples is likely to account for some of the variation in cognitive paradigm findings, as the literature is very versatile in the different types of bilinguals tested. It is possible some of the tested samples would not have shown a reduced effect of emotion regardless of the test. The literature is inconclusive as to when cognitive paradigms are an appropriate measure, and which characteristics the sample has to meet to (A) experience reduced emotional resonance in the first place (which is a much larger-scale theoretical issue, discussed more in detail in experiment 1), and (B) for the paradigms to work.

One explanation, which always must be considered, is bilingual L2 proficiency and the effect of cognitive effort. It is possible that the cognitive effort effect is a stronger confound in paradigms which are based on cognitive processing, as opposed to physiological paradigms. We attempted to control for cognitive effort by only testing highly proficient bilinguals who are immersed in their L2. We also selected the stimulus set algorithmically to consider lexical dimensions that may increase cognitive effort and included those variables as covariates in the analyses. We do not believe the present findings are due to poor participant

English proficiency, as there were no differences in word recognition by-language or by-group. It is possible that the cognitive effort effect simply cannot be disentangled from the effect of emotion in bilingual lexical processing, thus decreasing the sensitivity of cognitive paradigms as a measure of bilingual emotion. A systematic examination of this is required.

Interestingly, in some of the analyses in this chapter, we found LexTALE to be a significant predictor of word recognition, pupillary response and reaction time. This contrast with some of our previous findings (Experiment1, Experiment 4), but participant LexTALE scores were also found to predict word recognition in Experiment 2. Perhaps the score range needs to have more variance for LexTALE to have any predictive value – in Experiment 2 and the present experiment, the score range was wider than in Experiment 1. However, in Experiment 4, where the range is comparable, there was no effect of LexTALE. These findings suggest we cannot draw firm conclusions about how participant proficiency affects physiological responses. Word recognition, on the other hand, seems to be more often predicted from LexTALE proficiency scores.

If cognitive paradigms are not a sensitive measure of bilingual emotion, we need to examine why, and what exactly their difference to physiological measurement is. It seems that they tap onto different facets of word processing; simply understanding a word, as opposed to having a bodily response to the word. The underlying mechanism of this is not entirely clear. To date, the only study systematically comparing the two found reduced emotional resonance with skin-conductance measurement, but not with a Stroop task (Eilola & Havelka, 2010). The authors suggested that proficient bilinguals have quick and automatic access to the denotative meaning of the words, yet still lack the connotative and contextual meaning, and consequently also physiological arousal elicited by these words.

Pavlenko (2005, 2012) argues that bilingual emotion is “disembodied cognition”, and that we should be mindful of the distinction between a concept, and semantic meaning of a word. Concepts include the physical and mental contexts we associate with each word, and are thus embodied, whereas semantic meaning is simply our understanding of the word. It is possible cognitive paradigms do not tap on to concepts, but rather the meaning of a word and consequently are not very reliable at measuring disembodied cognition. Eilola and Havelka’s (2010) findings would support this idea – semantic access may be at least partially distinct from embodiment. However, this does not fully explain why in monolinguals, cognitive paradigms are sensitive to affective dimensions, and other word qualities that indicate embodiment (such as concreteness and frequency). It is unclear why these main effects also can be found in bilingual groups – the relationship of bilinguals “understanding” and “feeling” a word seems to be more complex and requires rigorous testing, as well as new theoretical perspectives.

This experiment highlights a problematic issue in the literature of reduced emotional resonance. A large number of experiments use cognitive paradigms to measure reduced emotional resonance. Cognitive paradigms are easy to design, quick to administer and generally widely used in psycholinguistics research. There is a lot of information on how they behave under different experimental design scenarios, and on covariates that should be included. However, as we have discussed in all the previous chapters, measuring affective responses in bilinguals is a very difficult task. Many of the studies using cognitive paradigms make strong claims about the phenomenon, yet the field is ultimately lacking a systematic approach to the different methods used, such as a series of methodology-focused experiments, or a meta-analysis.

The physiological measure used in the present experiment also failed to detect reduced emotional resonance in our bilingual sample. The interaction did not appear in either cross-group or cross-language comparisons. This was somewhat unexpected. Consequently, we cannot compare the two measurement methods, and cannot draw strong conclusions about the general suitability of these tests, as was initially intended in the aims of this experiment. Unfortunately, without a physiological comparison point the Lexical Decision Task findings simply add to the inconsistent body of literature with cognitive paradigms.

We did not replicate our previous findings from similar pupillometry tasks (see Experiment 1, Experiment 4, Toivo & Scheepers, 2019). The low arousing words (neutral valence words), unexpectedly, elicited the strongest pupillary responses, hence making us unable to detect a “baseline” comparison point for the high arousing words. This was frankly a confusing finding, which cannot be fully explained with the information we get from this experiment. There are a few possibilities which we will discuss further.

Looking at the figure summarising the pupil change across all time bins split by word type (see Figure 17 and 18), the overall pattern in by-group results seems very similar to what we have obtained from Experiment 1 and 4 (see Figures 3 and 28). It should be noted that the stimuli were split differently in this experiment – the HA words were split into positive and negative words, and the third category were low arousing, neutral valence words. The important distinction to our previous experiments is that here, we did not include words that are “neutral” in arousal (between high arousing and low arousing) in this experiment. Previously, neutral arousal words have elicited the strongest pupillary responses. Despite splitting the stimuli on a different lexical dimension, there still seems to be one word type that elicits the strongest pupillary response. This leads us to believe that pupillary response may be more sensitive to the effect surprise than it is to word arousal, and the proportions of different word types within the stimuli set of an experiment contribute to this.

In our other pupil experiments (see Experiment1, Experiment 4) neutral distractor words have consistently elicited the strongest pupillary responses. Neutral words have in all the previous experiments contributed one third of the stimuli words (1/3 high arousing, 1/3 low arousing, 1/3 neutral distractor). We have interpreted this as an effect of surprise – the participants detect a pattern of seeing “very exciting” and “very boring” words, and the neutral items do not fit this pattern.

Pupillary response is sensitive to surprise and unexpected changes in patterns (Kloosterman et al., 2015; Scheepers et al., 2013). It is possible surprise drives pupillary response up more than the arousal dimension of the stimuli items. In the present experiment we had 1/3 negative, 1/3 positive and 1/3 neutral stimuli items. Perhaps the participants started anticipating a semantic pattern of positive and negative alternating in the experiment, and deviations from that (i.e neutral valence words) contributed to a heightened pupillary response.

When compared to the word recognition findings of the present experiment, this finding provides grounds from some interesting speculations. In the word recognition tasks, the “neutral” words were consistently less often recognised than the two HA word categories (positive and negative). This is somewhat strange given that the items are all controlled for variables that affect word recognition (such as word length and frequency). It is possible that during the experiment participants implicitly categorise words into “Typical” vs. “Untypical”, and items deviating from that are harder to recognise and more surprising. This is a very interesting potential confound, which should be further examined in a series of experiments altering the proportions of word types in the experiment.

Hence, we cannot conclude from these findings that physiological measures would be better suited for detecting reduced emotional resonance, or that better control of the confounds in stimuli items would eliminate the unreliability of cognitive paradigms.

Against our expectations, we found a difference between negative and positive stimulus items in the pupillometry experiment. Previously, it has been suggested that pupil response is driven by arousal rather than valence (e.g. Bradley et al., 2008; Kuchinke et al., 2007; Partala & Surakka, 2003). There is evidence for this using words, pictures and sounds as stimuli. Here, we found that negative items overall elicited a higher pupillary response, as opposed to positive words. This findings warrants further investigation.

The unexpected pupil findings provide some evidence against the positivity bias in bilinguals; to detect a positivity bias, the difference in pupillary response between positive and negative items should have been reduced in L2. The Lexical Decision Task findings, on the other hand, provide more conflicting evidence. In the LDT, we did not find significant differences between positive and negative words in the by-group RT and word recognition models. However, in the by-language word recognition task (see section 4.3.2.3.2) we found that the difference between *positive* and neutral words was larger in the L1 German speakers, in comparison to the monolingual L1 English speakers. This can be taken as some (very) tentative evidence for positivity bias. However, we should consider these findings with caution as the LDT generally did not show an interaction of word type and participant test language/group, which means we did not detect reduced emotional resonance with the test. This, in turn, suggests that LDT (and possibly other cognitive paradigms) may not be the best measure of embodiment of the emotional language in bilinguals.

Whilst inconclusive in hypothesis-testing, this experiment has provided some interesting and potentially useful methodological considerations about cognitive paradigms and pupillometry experiments. Cognitive paradigms should not be used to draw conclusions about the nature of bilingual emotion processing until we know what exact mechanism distinguishes them from physiological measurement, and why the findings are so inconsistent. In terms of pupil measurement, the potential effect of surprise and proportions of word types in stimuli sets should be further examined in a series of controlled experiments.

Chapter 5 Metacognitive measurement on reduced emotional resonance

5.1 Introduction

In the previous experiments we have focused on physiological and cognitive measurement of reduced emotional resonance in L2. These methods rely on automatic and implicit processes of the bilingual mind. This experiment, in turn, will look into metacognitive judgment – we are using an explicit measure of emotion, namely ratings of affective words, to measure reduced emotional resonance.

The three types of measures (physiological, cognitive, and metacognitive) each are based on a different mechanism of language processing. Affective ratings require deeper and more conscious processing of the stimuli. As opposed to measuring direct effects of emotion, or how these effects affect behavioural measures, such as reaction times, affective ratings measure participant *perception* of emotional language. This is an interesting point, as the whole study of reduced emotional resonance stems from bilinguals' experience, and how they perceive feeling less in their L2 (Pavlenko, 2006).

Unexpectedly, affective rating tasks have not been used very frequently in research on reduced emotional resonance – the procedure itself is much easier to conduct than physiological measurement, and does not require specialist equipment. When used, affective rating has often been presented as a validation task, or a condition in an experiment measuring something else, and sometimes entirely omitted from the results section (e.g. Ferré, García, Fraga, Sánchez-Casas, & Molero, 2010).

In their bilingual recall task, Anooshian and Hertel (1994) had a fully crossed design with English and Spanish L1 speakers, and word stimuli presented in English and Spanish. The participants were split into different conditions according to the rating task (ratings of emotion, ease of pronunciation and word activity) they were asked to perform prior to the recall task. There was no interaction of test language and word type, or participant group and word type in the affective word ratings. This suggests the affective rating task did not detect reduced emotional resonance in their bilingual sample.

In one of the first experiments using physiological measurement of reduced emotional resonance, Harris (2004) used a rating task during the SCR measurement. It should be noted the rating in this experiment was of valence (how pleasant the participants found the words) rather than arousal. No differences between the participants' languages were found in the ratings of the stimuli, even though reduced emotional resonance was detected in the SCR measurement. In a later study, Ayçiçegi-Dinn and Caldwell-Harris (2009) used an emotional intensity rating task as a part of their word recall study. No differences between L1 and L2 words were found – however, they also did not find a decreased emotion effect on word recall in L2.

Winskel (2013) compared emotionality ratings of negative and neutral words in Thai speakers (L1 vs. L2), and between the Thai speakers and monolingual L1 English speakers. They found no interaction of the participant group or language, and word type, suggesting there was no difference in how emotional the words were rated between L1 and L2. In the emotional Stroop task, on the other hand, they found some indication of reduced emotional resonance of L2.

More recently, Iacozza and colleagues (2017) included a rating task in their pupillometry experiment (2017); after each trial participants were asked to rate how they find the emotional impact of each of the words on a 7-point Likert type scale. Again, the rating measure was not purely that of arousal, but a composite measure of valence and arousal at the same time (7 on the scale would indicate a "high, negative impact"). They found a reduced emotional resonance effect in pupil responses in the L2 group of participants, but this effect did not carry over to the ratings.

An interesting point to note is that both studies summarised above (Winskel, 2013; Iacozza et al., 2017), which used ratings as an actual measure (and not a control measure), only looked at neutral and negative words instead of using the full range of emotionally arousing words, both of high and low valence. It has been argued that only negative words are disembodied in L2 (Sheikh & Titone, 2016), but we have found no evidence for this with our balanced stimuli sets

(Toivo & Scheepers, 2019; Experiment 1, Experiment 3). Hence, it would be of interest to include both positive and negative high arousal words in a rating task to paint a more comprehensive picture of the interplay of the two in bilingual language processing.

Ong and colleagues (2017) used both valence and arousal ratings as measures of bilingual emotion processing in Chinese-English bilinguals. Contrary to their hypothesis, they found no interaction of word type and language in arousal ratings. Overall, words were rated as more arousing in participants' L2. Negative words were rated as more arousing in L1, and positive words were rated as more arousing in L2. For valence ratings, both positive and negative words had more polarised ratings in English (L2), as opposed to Chinese (L1).

Garrido and Prada's (2018) findings contrast this: testing Portuguese-English bilinguals, they found that words in L1 (vs. L2) were rated as more extreme in valence. There was no main effect of test language, but the crucial interaction of word type and test language was found, although only for taboo words. They were rated more emotionally intense in L1, as opposed to L2.

Dewaele (2016; 2018) has done some interesting work on bilingual affective ratings. In his studies the participants, however, rated the words on offensiveness rather than the standard affective ratings dimensions. In a rating study of 30 offensive English terms, the bilinguals (English L2) consistently rated the offensive words more offensive than the English as L1 speakers did, except for the word "cunt", the offensiveness of which bilinguals rated as lower than English L1 speakers did. Dewaele (2016) concluded this is possibly because the bilinguals want to mark offensive words as red flags and overcompensate for the reduced emotional resonance they are experiencing when using these words.

Dewaele's findings (2016; 2018) are consistent with those of Caldwell-Harris et al. (2011). In this study, participants were asked to think of a situation where the stimulus phrase would be used, and rate the emotional intensity of these situations, rather than the phrase itself. They found that Mandarin L1 speakers rated Mandarin reprimand situations as higher in emotional intensity than English (L2) reprimands. However, English taboo phrase situations were rated as more emotionally intense as opposed to Mandarin taboo phrase situations. This provides further support to the idea of red flags. While the ratings of these two studies are on unconventional dimensions (as opposed to simple emotional arousal of a given word or phrase), this conclusion can possibly explain why the affective rating findings are often inconsistent, with null effects, or L1 words receiving more polarised ratings.

An interesting point to note about the Caldwell-Harris et al., (2011) study is that they used a situated ratings task, as described above. It is possible that these instructions would likely direct the participants to the connotative meaning of words (Ferré et al., 2010), rather than simply denotative meaning. This may be one explanation as to why affective ratings so often do not detect the language*word type interaction.

Most of the experiments discussed above suggest that affective ratings tasks may not be a suitable measure to capture reduced emotional resonance. There is typically no interaction of word type and participant group/test language in ratings, even though the effect is detected with another task in the same participants. The purpose of this experiment is to expand this literature, and systematically investigate whether affective ratings can find reduced emotional resonance. Doing this, we would like to address a number of methodological issues prevalent in the previous studies.

All the rating studies discussed above have used standard ANOVAs or linear mixed effects models to predict ratings. Treating ordinal data as continuous is problematic (Liddell & Kruschke, 2018). Hence, analysing rating data with cumulative link mixed models would be more appropriate. Combining both negative and positive high arousal word stimuli in a carefully controlled, algorithmically selected stimuli set and rigorous statistical analysis, that is better suited for the ordinal nature of rating data, we hope to examine whether metacognitive judgments can be used as a measure to detect reduced emotional resonance in L2. Further, we will use a larger number of items (240 stimulus words) than any of the previous experiments.

This experiment will attempt to tap onto the difference between physiological response to words and an explicit evaluation of them. This hopefully will shed some light onto where the difference between semantics and embodiment of language lies. It is safe to assume the bilinguals we have tested in the other experiments, being highly proficient and immersed in their L2, know the meaning of the words. However, we have seen that their physiological responses are different, both in previous literature (Toivo & Scheepers 2019, Iacozza et al., 2017) and in our experiments (see Experiment 1). We speculate there is a difference in the perception of emotion from seeing and *understanding* the word, and *feeling* behind the language. Bilinguals can access the denotative meaning but may lack the connotative meaning, in other words, the concept of the word (Pavlenko, 2005).

The aim of this experiment is to combine an implicit, physiological measure (pupillometry) with an explicit measure of emotion (affective ratings). We expect to replicate the reduced emotional resonance effect found in experiment 1 in the pupillometry part of the experiment, but based on previous literature this effect may not be found in ratings.

5.2 Method

5.2.1 Participants

Eighty-two participants were recruited for the experiment (57 bilinguals, 25 English speakers; 56 female, 26 male). Participants were between 17-58 years of age (mean=23.68 years, SD=7.19 years). The bilingual participants were from 32 different countries (full breakdown in Appendix C), had lived in the UK an average of 6 years 8 months (SD=77 months), and mode age for English exposure was 6 years (mean=4.02 years, SD=3.57 years). Participants were paid for their participation or compensated with course credits.

5.2.2 Materials

Materials were taken from Experiment 1 (see Chapter 2, section 2.2.1). The language history questionnaire was slightly revised (see Appendix A).

5.2.3 Procedure

The procedure was nearly identical to Experiment 1 (see section 2.2.4). The only change made was the addition of a rating task after each trial – instead of pressing a button the participants were instructed to rate the word from 1 (extremely calming) to 9 (extremely exciting). After each trial participants were asked to say their rating out loud and the experimenter recorded this on the experimenter PC. Should the participant not see or know the word, they were instructed to give 0 as their rating. After completing the experiment, participants were asked to complete the LexTALE test for English proficiency (Lemhöfer & Broersma, 2012).

5.3 Results

5.3.1 LexTALE

L1 English speakers scored 94.9% on average (ranging from 77.5-100%), while the bilingual speakers scored 82.11% on average (ranging from 48.8-100%). Nine English speakers scored 100% and one bilingual speaker scored 100%. In comparison to our previous samples, it seems that the bilinguals did slightly worse on the LexTALE test. Table 28 below summarises the LexTALE scores for each participant group.

Table 28 Mean LexTALE scores and SDs by group

Group	Mean	SD	Min	Max
Bilinguals	82.1%	12.6%	48.8%	100%
English speakers	94.9%	6.1%	77.5%	100%

5.3.2 Word rating task

5.3.2.1 Descriptive statistics

Table 29 below summarises mean arousal ratings given to each word type and their standard deviations, split by group

Table 29 Mean arousal ratings and SDs split by group and Word Type.

Group	High arousing		Low arousing		Neutral distractor	
	Mean	SD	Mean	SD	Mean	SD
Bilinguals	6.45	2.06	3.22	2.11	5.11	2.12
Monolinguals	6.47	1.75	3.40	1.80	4.96	1.95

When looking at the mean ratings and the boxplot below (Figure 25) showing the distiribution of ratings, it seems that the mean ratings are very similar across both groups. However, bilinguals' ratings seem to vary more and the scores overlap more across the word types. There seems to be a clear effect of word type: low arousing words have lower ratings than high arousing words, and neutral words seem to fall somewhere in between the two.

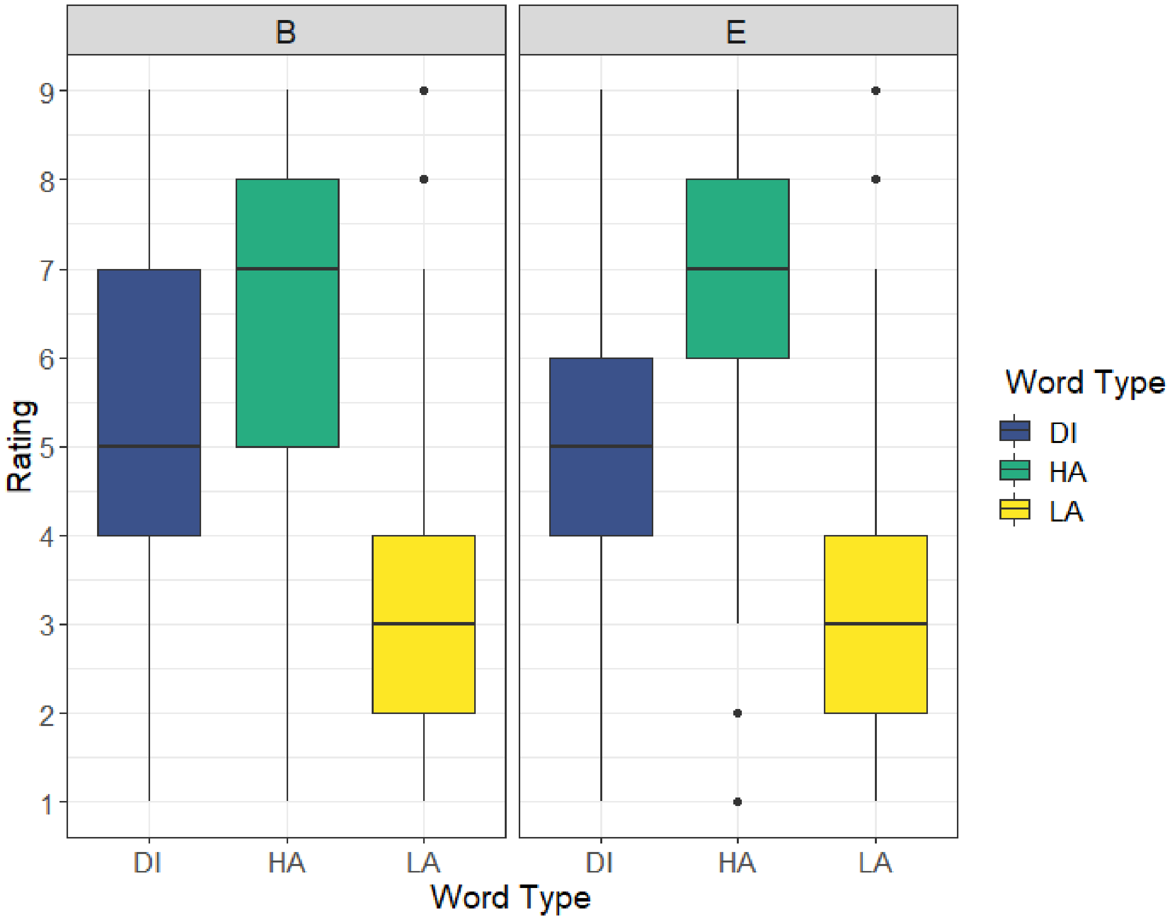

Figure 25 Rating distributions split by participant group and word type

5.3.2.2 Cumulative link mixed models

To analyse group and word type differences, cumulative link mixed models were run using the 'ordinal' package (Christensen, 2019). The models were specified with Group, Word Type, LexTALE score, and the lexical covariates (Bigram frequency, Concreteness, Lexical Frequency, Length & Orthographic neighbours) as fixed main effects. Group and Word Type were deviation coded, and the lexical covariates and LexTALE scores were mean-centered. Group interaction with the covariates and with LexTALE were entered as fixed interactions. For the random effect structure, a random intercept was included for subjects and items. Group was entered as by-item random slope and word type as well as the lexical covariates were entered as by-subject random slopes. It should be noted Group:LexTALE interaction was left out from the random structure due to computational issues with the clmm function. Full model syntax can be found in Appendix D. All models were specified with cloglog link and nlminb control, which were chosen after comparing model fits. The distractor words were not included in the models to keep them consistent with Experiment 1 and 2 analyses.

Table 30 Arousal ratings model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
1|2	-2.77	0.08	-34.32	<0.001***	-2.93	-2.61
2|3	-1.76	0.08	-22.97	<0.001***	-1.91	-1.61
3|4	-1.20	0.08	-15.85	<0.001***	-1.35	-1.05
4|5	-0.71	0.07	-9.42	<0.001***	-0.85	-0.56
5|6	-0.22	0.07	-2.93	<0.001***	-0.36	-0.07
6|7	0.34	0.07	4.52	<0.001***	0.19	0.48
7|8	0.97	0.07	13.01	<0.001***	0.82	1.11
8|9	1.68	0.08	22.22	<0.001***	1.53	1.83
Word type	1.65	0.11	14.64	<0.001***	1.43	1.87
Group	0.05	0.16	0.32	0.75	-0.27	0.37
Lexical frequency	0.01	0.03	0.47	0.64	-0.04	0.07
Bigram frequency	0.03	0.03	0.96	0.34	-0.03	0.09
Concreteness	-0.02	0.03	-0.55	0.58	-0.08	0.04
Length & orthographic neighbours	0.03	0.03	0.93	0.35	-0.03	0.09
LexTALE	0.00	0.01	-0.69	0.49	-0.02	0.01
Group: Word type	-0.32	0.21	-1.54	0.12	-0.72	0.09
Group: Lexical frequency	0.09	0.03	3.09	0.002**	0.03	0.14
Group: Bigram frequency	-0.02	0.03	-0.58	0.56	-0.07	0.04
Group: Concreteness	-0.05	0.03	-1.68	0.09	-0.11	0.01
Group: Length &orthographic neighbours	-0.04	0.03	-1.47	0.14	-0.10	0.01
Group: LexTALE	0.00	0.02	-0.05	0.96	-0.03	0.03

Table 30 and Figure 26 summarise the arousal rating model findings. There is a main effect of word type, showing that HA words were rated higher in arousal than LA words. There is no main effect of group, which suggests both groups rated the stimuli in a similar manner, despite the descriptive differences seen in Figure 25. There are no main effects for any of the lexical covariates. Interestingly, there is an interaction of lexical frequency and group. Crucially no Group: Word type interaction is found, suggesting there were no quantifiable differences in how the word types were rated between the participant groups.

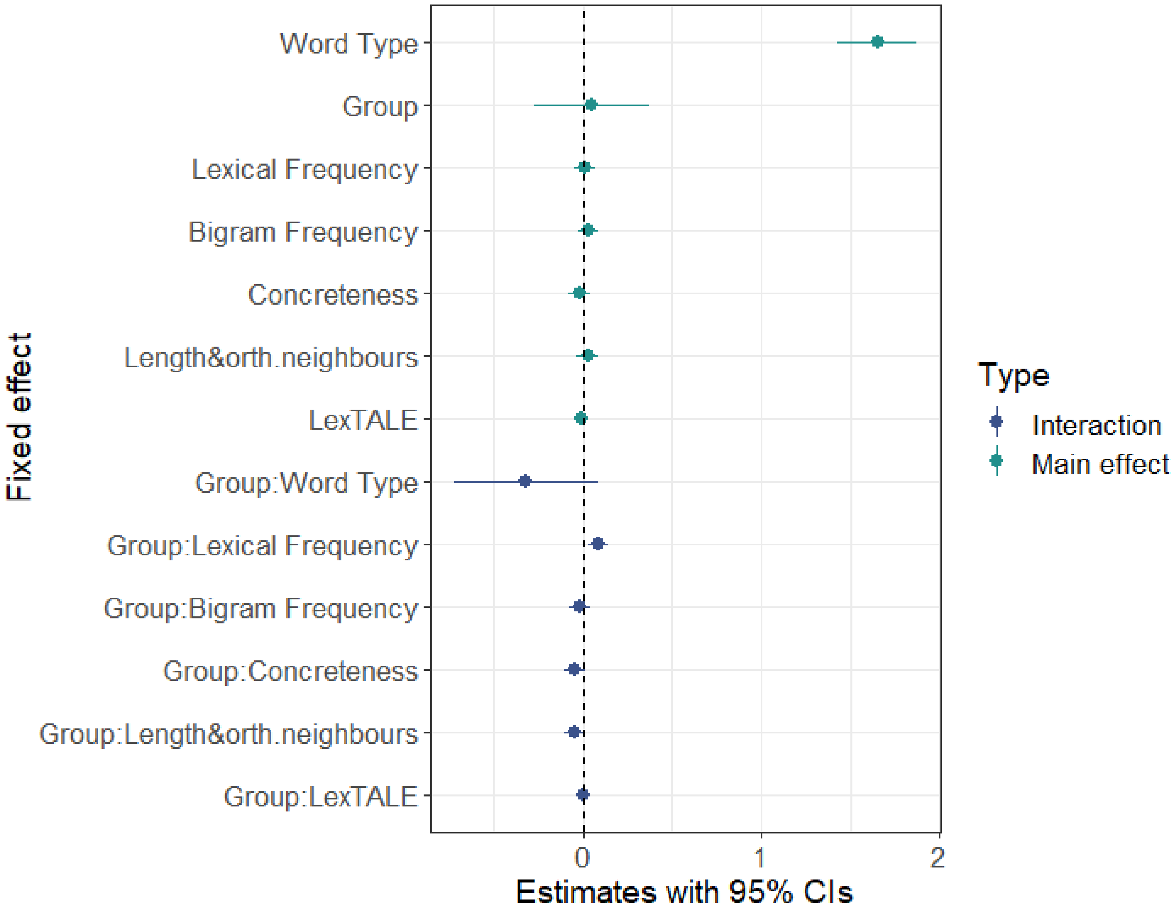

Figure 26 Ratings cumulative link mixed model estimates with 95% Confidence intervals

5.3.3 Eye-tracking task

5.3.3.1 Pupil data pre-processing

The pupil data were pre-processed using the protocol from Experiment 3 (see section 4.3.3.1). It should be noted there was an unusually large number of missing values in this dataset – 62118 observations (trial*timestamp) were removed, which constitutes 36% of the data. This is perhaps due to hardware issues that followed moving the laboratory twice during the testing period.

5.3.3.2 Pupil data descriptive statistics

Figure 27 shows the distribution of area under the curve values. Due to skewness, pupillary response mixed effects models will be specified with the Gamma family argument. These distributions only include trials where participants indicated that they recognised the word.

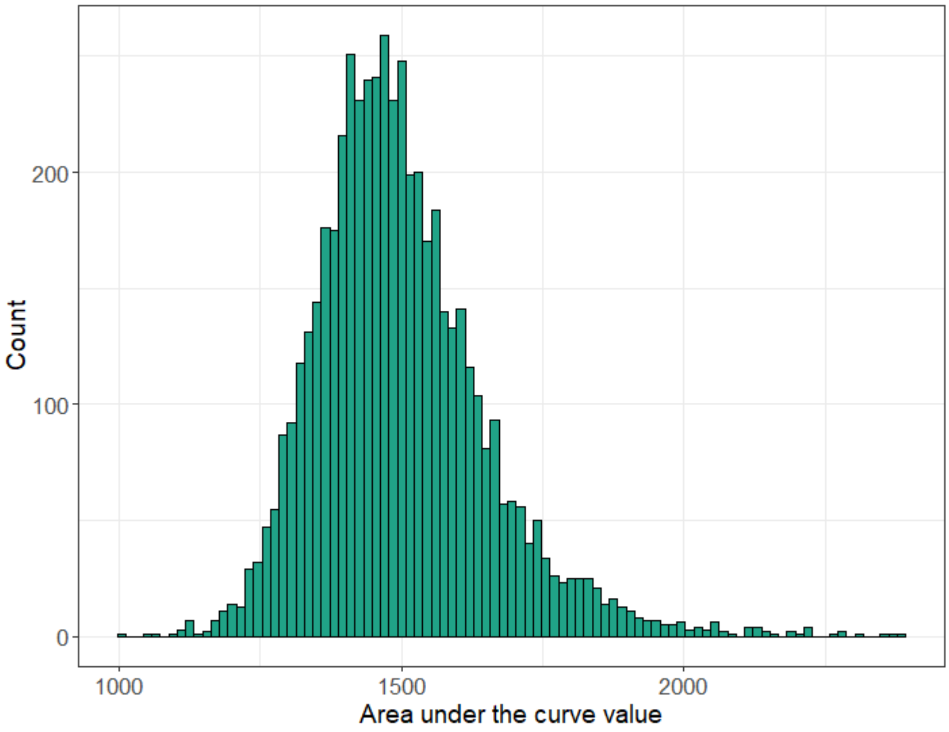

Figure 27 Pupil area under the curve distribution

Figure 28 below shows pupil change averaged across time and word type, split by speaker group (monolingual vs. bilingual speakers). From this graph it appears that neutral words elicited the strongest pupil response in the monolingual group, but not in bilinguals. The difference between high arousing and low arousing words seems to be smaller in the bilinguals as opposed to the monolingual group. The monolingual speakers overall seem to have a smaller pupillary response, which would be consistent with the cognitive effort effect on pupil response. Interestingly, in comparison to our previous pupil experiments the pupil response seems to last longer and does not fully reset within the measurement period. This is very likely due to a more cognitively demanding task and the participants actively thinking about the stimuli words – in the previous experiments participants were merely asked to indicate whether they recognised the word or not, whilst in this experiment participants were asked to rate the emotional arousal of each of the stimuli words.

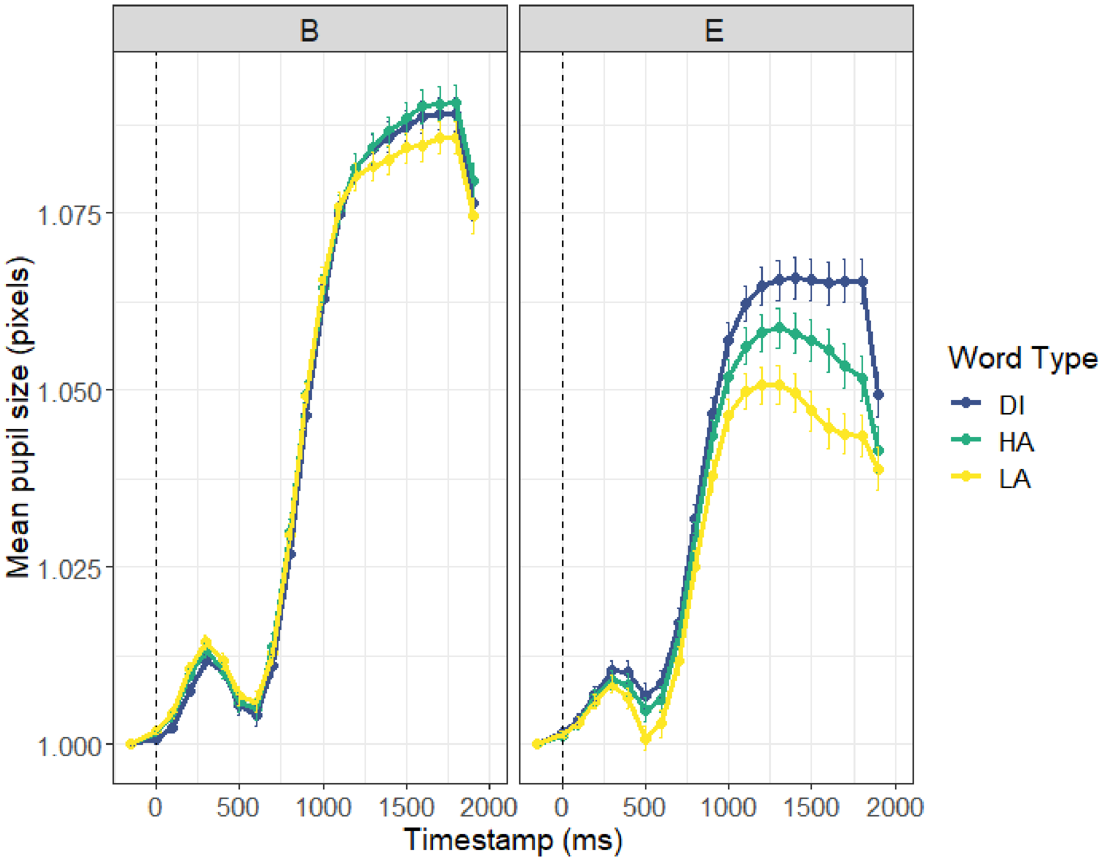

Figure 28 Mean pupil across size time and word type, split by participant group

5.3.3.3 Area under the curve mixed models

By-group pupillary data were analysed using generalised linear mixed effects models. Participant group and word type were deviation coded for analysis. LexTALE scores, as well as the principal components used as item covariates were mean-centered. The full area under the curve model (Appendix D) included main effects of group, word type, and LexTALE and Length, Frequency, and Concreteness of the items as covariates. Word type and principal components were also included as fixed interactions with group. In the random structure, random intercepts were included for item and subject, word type was included as a by-subject random slope, and group was included as by-item random slope. The models were specified with Gamma family argument, and optimx package L-

BFGS-B optimiser. It should be noted that two sets of analyses were run: one with the standard area under the curve from 500ms onwards, and one in which the modelling data were narrowed down only to include time bins from 1000ms onwards – perhaps due to the more demanding nature of the ratings task, the pupillary response does not start to separate per word type until 1000ms. Only the latter set of analyses is reported, as there were no notable differences in the parameter estimates, SEs and p-values. The last time bin (1900ms) was excluded from both sets of analyses, as there seems to be significant noise; pupil size should not drop as drastically within 100ms (Figure 28).

Table 31 summarise and Figure 29 summarise the parameter estimates of the full area under the curve model, standard errors, z-values, p-values and 95% confidence intervals for the fixed effects. There is a main effect of group, which mirrors our previous findings of bilingual participants having a stronger pupillary response overall. The effect of word type is marginal, showing that HA words predict larger pupillary responses. A main effect of Lexical frequency is observed, which aligns with our previous findings. Although from the descriptive measures (see Figure 28) it seems like there should be a Group:Word type interaction, this does not come through significant in the model, thus failing to replicate Experiment 1. We suspect this may be due to issues with the hardware (see above pupil data pre-processing and discussion, as well as limitations section in the overall discussion).

Table 31 Pupillary response model summary

Term	Estimate	Std. error	z-value	p-value	95% CI low	95% CI high
Intercept	757.62	3.02	251.19	<0.001***	751.71	763.53
Word type	4.20	2.18	1.93	0.05.	-0.06	8.47
Group	15.10	7.09	2.13	0.03*	1.20	29.00
Lexical frequency	-2.82	1.21	-2.32	0.02*	-5.20	-0.44
Bigram frequency	-1.64	1.19	-1.39	0.17	-3.97	0.68
Concreteness	0.28	1.25	0.22	0.82	-2.17	2.73
Length& orthographic neighbours	-2.00	1.21	-1.66	0.10	-4.36	0.37
LexTALE	-0.01	0.26	-0.03	0.98	-0.51	0.50
Group:Word type	-4.23	3.58	-1.18	0.24	-11.26	2.79
Group:Lexical frequency	0.85	2.15	0.40	0.69	-3.36	5.06
Group:Bigram frequency	-1.11	2.06	-0.54	0.59	-5.16	2.94
Group:Concreteness	2.74	2.21	1.24	0.21	-1.58	7.06
Group:Length & orthographic neighbours	-0.42	2.10	-0.20	0.84	-4.54	3.69
Group:LexTALE	-0.40	0.66	-0.61	0.54	-1.69	0.89

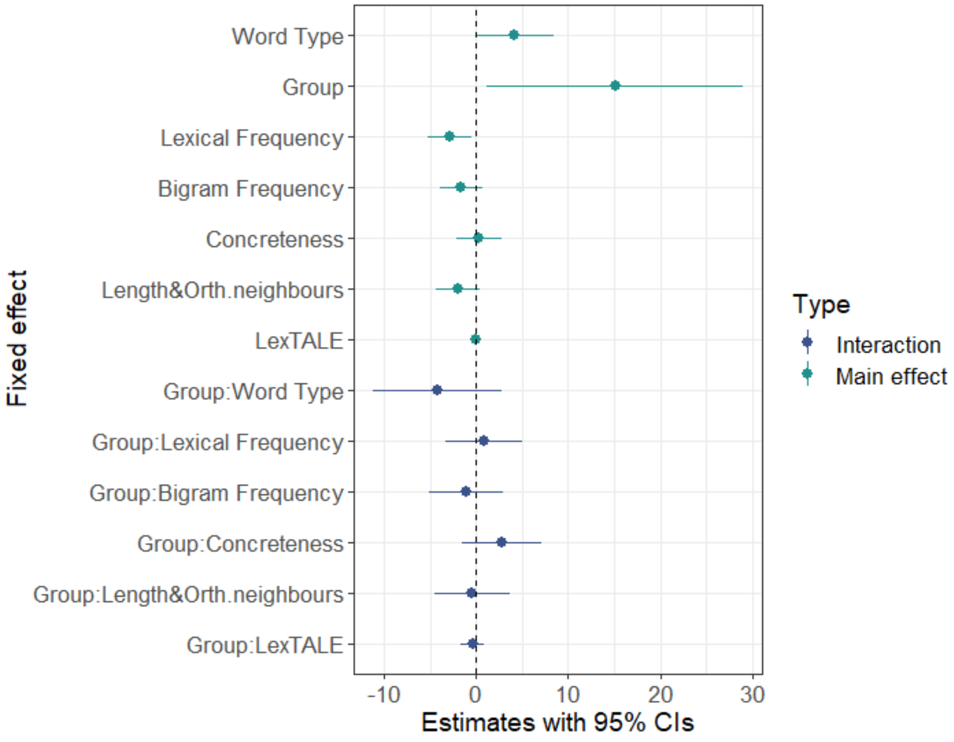

Figure 29 Pupil response model estimates with 95% Confidence Intervals

5.3.4 By-item correlations

By-item correlations were tested to see whether there is a relationship between ratings and pupillary responses on a by-item basis. To do this, a mean area under the curve value and a mean rating were calculated for each item, and these were tested with a Spearman correlation.

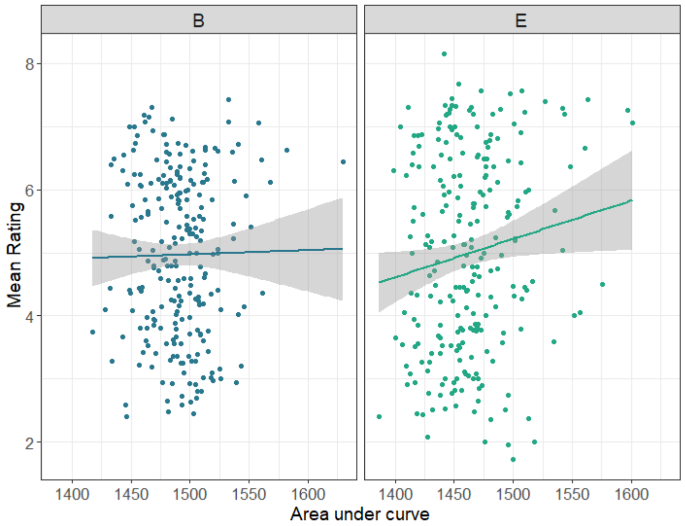

Figure 30 Scatterplot of mean rating and area under the curve values, split by participant group

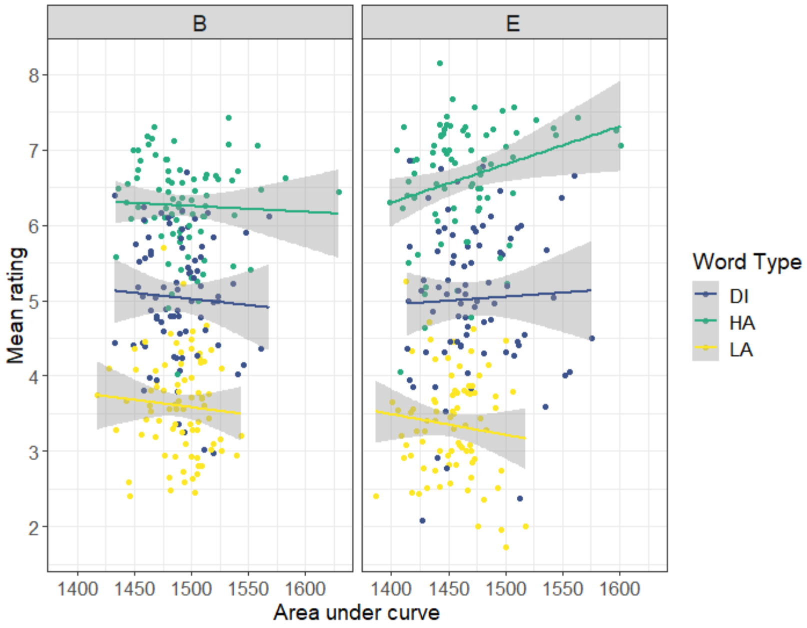

Figure 31 Scatterplot of mean rating and area under the curve values, split by word type and participant group

From the scatterplots (Figure 30 and 31) it can be seen that the associations between area under the curve values and mean ratings seem rather weak, especially for the bilingual group. When split by word type, there seems to be a positive correlation only for the high arousing words, and only for the L1 English speakers. Results of the Spearman correlation indicated that there was a weak positive association between area under the curve values and ratings in the L1 English speaker group. The association was not significant (rs(238) = 0.11, p =0.09). In bilinguals, there was a weak negative association between area under the curve values and ratings (rs(238)=-0.05, p =0.44). This association was not significant.

5.4 Discussion

The aim of this experiment was to compare a metacognitive measure (affective word ratings) to a physiological (pupillometry) measure of emotion. We expected to replicate our previous findings from pupillometry experiments. We expected there to be an interaction of word type and speaker group such that the difference between high arousing and low arousing words is smaller in bilinguals. The rating task was exploratory, but based on previous findings we did not think it was very likely for the task to detect reduced emotional resonance.

We did not find a significant interaction of group and Word Type in the pupillary task. Descriptively, it seems that the interaction is present, but the effect was not found in the GLMEM. This was unexpected, and we suspect is due to hardware problems causing the data to be very noisy – a grand total of 36% the pupil data were removed. The eye-tracker laboratory had to be moved twice during the testing period and given that Eyelink II is not a portable system, this was less than ideal. Alternatively, it is possible the effect was not going to replicate with this sample regardless of the quality of data, but we cannot draw any strong conclusions from the existing data.

Interestingly, the pupillary response to neutral items seems to be slightly different in this experiment (see Figure 28) in comparison to Experiment 1. In Experiment 1, much to our surprise, the pupillary response to neutral distractor items was stronger than to HA and LA words. Here, on the other hand, only the monolingual group has an exceptionally strong response to neutral distractor items and for bilinguals it falls between high arousing and low arousing items as was initially intended when designing these materials. We have previously interpreted the distractor effect as a possible effect of surprise – this effect, considered together with the findings from Experiment 3, will be discussed in more detail in the overall discussion. It should be noted that neutral items were not included in the Generalised Linear Mixed Effects Models, and hence any differences discussed here are merely descriptive (and in this experiment may also be due to noise).

It is possible that the rating task at the end of each trial explains this difference, although that does not explain the difference found between the two participant groups. In our previous experiments, the participants have performed a simple word recognition task after each trial. In the present experiment they had to actively think about the word they have seen and give it an affective rating. This is likely to increase the cognitive load, as can be seen in the pupil graphs (Figure 28) – typically, within the two-second recording window the pupil returns to baseline, but here we can see the pupil is still dilated at the end of the recording. Having to actively think and evaluate a word to give it an affective rating is much more demanding than a simple recognition task which does not involve deeper processing of the word.

This is an important methodological consideration for future studies measuring pupillary responses to single-word stimuli. If rating tasks are used as a validation or as an additional measure after each trial, it is important the measurement window or the time between trials is sufficiently long for the pupil to return to baseline. Otherwise the baseline measure will be conflated from previous trials. Although our measurement window in this experiment was only 2 seconds, the task took an additional 2-3 seconds as the experimenter recorded participants' rating for each word. This should give enough time for the pupil to return to baseline, which is why we are not concerned about conflated baseline measures (but would add 0.5-1 seconds to the measurement window in future studies using ratings concurrently with pupil measurement). One alternative to this is to avoid using time series data altogether and average the pupil size over a longer measurement window, as was done in Iacozza et al. (2017).

For the rating task, the cumulative link mixed models detected no interaction between word type and participant group – hence, we cannot conclude reduced emotional resonance from the affective ratings. This was somewhat expected as many of the previous experiments using this measure have not found clear evidence for reduced emotional resonance (see Iacozza et al., 2017; Ong et al., 2017; Winskel 2013). Overall, the ratings reflected the pattern found in the original ratings of the stimuli words (Warriner et al., 2013) – we found a significant difference in ratings between high arousing and low arousing words. There was no main effect of group, suggesting that both groups rated the words in a similar manner.

When looking at the descriptive differences between the speaker groups (Figure 25) we can see the bilinguals have more variance in their ratings for HA words, whereas ratings for LA and neutral distractor words are more aligned to the monolingual speaker ratings. This can be taken as tentative evidence for differential processing between the word types and how bilinguals perceive HA words as less arousing. However, no strong conclusions can be drawn from this as the cumulative link mixed model showed no effect of word type and participant group interaction.

The variance in bilinguals' ratings of the HA words, in comparison to our previous pupil findings is interesting. In some of the pupil experiments (see Toivo & Scheepers, 2019, Experiment 1) it seems that low arousing words may elicit a stronger pupillary response in bilingual participants, as opposed to high arousing words. The ratings tentatively suggest bilinguals may perceive high arousing words less arousing, or at least that the variance is larger, whereas the pupillary responses suggest the effect may in fact be due to the low arousing words being processed differently. These theoretical considerations about the word type effects and what drives them warrant systematic exploration. This will be discussed in more detail in the overall discussion.

Another interesting point is the by-item correlation analysis between pupillary data and word ratings. This was an additional exploratory analysis we conducted – we calculated mean area under the curve and rating scores for each item and there was no significant correlation between the two. There was a weak positive correlation of area under the curve and pupillary response in the English monolingual group, but this again was not statistically significant. Perhaps the most feasible explanation for this is that pupillary response typically has large individual differences and pupil data are very noisy. There was also no correlation on a by-participant basis (where words were split into word types).

The findings of this experiment provide a few interesting pointers for future research. It seems that affective ratings are not particularly sensitive to detecting reduced emotional resonance. This is consistent with previous studies using a rating measure (see Iacozza et al., 2017; Winskel, 2013).

In the language embodiment theory, Pavlenko (2005, 2012) suggests that differenced in bilingual emotion processing between the languages may be due to disembodied cognition. Understanding the meaning of a word is different to having a conceptual representation of it. The concept of a word involves a rich representation of the word in different contexts – it requires causal antecedents, appraisals and physiological responses (Pavlenko, 2005). Perhaps this distinction explains why the word ratings do not detect reduced emotional resonance while pupillometry does; when isolated from its context, affective ratings may tap onto the word meaning (which still carries affective dimensions such as arousal and valence), whereas the underlying physiological response is based on a concept.

In conclusion, this experiment sheds some (rather dim) light onto the difference between understanding the word and feeling the word. This complex relationship requires further investigation. The present experiment does not provide evidence for affective ratings as a reliable measurement tool of reduced emotional resonance. Future studies should consider this when choosing their measurement methodology.

Chapter 6 Implications of reduced emotional resonance – investigating the optimality bias in L2.

6.1 Introduction

The foreign language effect (FLE) suggests that the tendency of bilingual speakers to experience less emotional involvement in their second language (L2) can lead to a reduction in cognitive biases (e.g. Keysar et al., 2012). This means that when using their L2, bilinguals may be able to engage in more rational thinking, which in turn may lead to a reduction of typical biases in decision-making or moral judgment.

Evidence for the FLE has been provided for a number of different cognitive biases. For example, it has been found that the FLE may reduce superstitious belief (Hadjichristidis et al., 2019). Bilingual participants in this study were asked to rate how bad or good they would feel about doing an action (such as applying for a job) in different "good luck" and "bad luck" scenarios. It was found that reading the scenarios in their L2 prompted more neutral feelings towards good versus bad luck scenarios. The FLE has also been found to mitigate causality illusions in a contingency learning task, where people falsely believe that two events are related (Diaz-Lago & Matute, 2018).

Most of the research on the FLE has been conducted in the context of decision-making. For instance, Keysar et al. (2012) investigated the loss-aversion bias, i.e. whether the way a decision-making dilemma is framed affects how participants choose to respond to it (see also: Kahneman & Tversky, 1979). Their participants were presented with a hypothetical scenario in which 600,000 people were exposed to a deadly disease. The participants were presented with two choices of medicine, one of which was a "sure" option (A) and one of which was a "risky" option (B). In the *gain frame* condition, participants were told that a choosing medicine A will save 200,000 lives, whilst if they choose medicine B, there is a 33.3% chance that 600,000 people will be saved and 66.6% chance that no one will be saved. In the *loss frame* condition, they were told that choosing medicine A will cost the lives of 400,000, whilst with medicine B, there is a 33.3% chance that no one will die, and a 66.6% chance that 600,000 people will die. Hence, the outcomes were identical in both framing conditions - however, participant’s choices were not. They were more likely to choose the “risky” medicine (B) if the outcome was framed in terms of *loss* rather than *gain* – in other words, a clear framing effect was found. Crucially, being presented the dilemma in one’s L2 mitigated this bias. These findings have also been replicated by Costa, Foucart, Arnon, Aparici, and Apesteguia (2014) on a number of similar framing problems. They suggested that using L2 reduces loss aversion because it mutes the emotional involvement of participants.

In an investigation of utilitarian judgements, Costa and colleagues (2014) studied the classic ‘footbridge dilemma’, and found bilinguals participating in their L2 were more likely to opt for (hypothetically) pushing one individual off a bridge to save the lives of five others. They argued that due to reduced emotionality in L2, the emotional compromise of harming one individual does not interfere with the rational decision of saving more lives. Further research has found that the effect emerges for the ‘footbridge dilemma’, but not the ‘trolley dilemma’, which involves pushing a button to sacrifice an individual, instead of actively harming the individual (Cipolletti et al., 2015; Geipel et al., 2015a).

Emotionality of the decision-making scenario presented seems to be an important mediator. Using one’s second language only seems to mitigate the bias for more emotional and morally compromising hypothetical situations; for example, those involving actively pushing a person to their death. Corey and colleagues (2017) replicated this effect over several experiments, and found that the FLE was stronger in personal dilemmas, as opposed to impersonal ones. Importantly, it was also found that the effect decreased if emotionality was diminished by manipulating the severity of consequences, e.g. death vs. disability vs. injury. Thus, the FLE appears to be stronger in more emotional contexts, which supports a strong link to reduced emotional resonance in one’s second language.

Little research so far has focused on whether the FLE also affects judgements about other people, in particular *attributions*. Attribution is defined as the process of assigning cause and meaning to the actions of others and/or phenomena in the world around us (e.g. Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015). Previous research on attribution suggests that people often fail to provide unbiased judgements. One well-known attribution bias, for example, is the *fundamental attribution error*: people are prone to attribute their own mistakes to environmental factors, whilst attributing mistakes made by others to dispositional factors (e.g. Ross, 1977). More recently, however, some theorists argued that this divide between ‘person’ vs. ‘environment’ is too simplistic, as it fails to address the complex reasons behind responsibility, such as intervening causes, failure to act, or previous failed attempts (Alicke et al., 2015).

The aspect of emotion has also been incorporated into attribution theory. According to the ‘person-as-reconstructor’ theory (Kahneman & Miller, 1986; Kahneman & Tversky, 1982), psychological reactions to an event are reconstructed *after* the event. Tragic outcomes produce strong affective reactions, which motivate observers to reconstruct the event and look towards alternative choices. An actor may be blamed for failing to act differently, even when the outcome was not foreseeable to the actor. Similarly, the ‘person as moralist’ theory (Alicke, Rose, & Bloom, 2011; Mandel, 2010) argues for a bidirectional relationship between cause and blame. The theory suggests that assessing an actor’s causal role becomes conflated with the observer’s emotional responses. Factors like negative perceptions of an actor, or negative consequences of an action, can therefore influence blame attribution to some extent.

According to the *optimality principle*, observers assume that people are rational and strive to make the best possible decision in a complex and competitive environment (Schoemaker, 1991). This principle is often problematic when judging other people (Toda, 1991), specifically given that observers are hardly able to account for the many unknown variables that can affect the actions of others. This can lead to a discrepancy between perceived intention and behaviour, and failure to realise that ‘good intentions’ may not necessarily lead to ‘good outcomes’ (or vice versa). In other words, observers often fail to recognise the simple fact that people are fallible and make mistakes, and that optimality cannot always be achieved.

A recent study has offered a novel application of this concept, by studying *optimality bias* in moral judgements (De Freitas & Johnson, 2018). The authors argued that suboptimal choices or actions made by others are difficult to understand, because people are always expected to behave optimally, even in situations where they do not have full control. Consequently, actors making suboptimal decisions will elicit more pronounced affective reactions in observers, and thus be subject to more severe moral judgements.

In a series of experiments (De Freitas & Johnson, 2018), participants were presented with different vignettes, each describing a scenario where an actor must choose between three different alternatives, e.g., a doctor having to choose between three different treatments for a patient with hearing problems. Unbeknown to the described actor, the three options had different degrees of optimality. The vignettes always explicitly stated that the actor thought that all options were of equal efficacy, while *in fact* they had statistically different success rates. Regardless of the described actor's decision, the vignettes always described the same tragic outcome (e.g., the patient suffering from permanent hearing loss after treatment). Participants were randomly allocated to conditions in which the actor made either the *best*, *middle,* or *worst* decision from an objective, omniscient perspective. It was found that actors who made the *best* choice were assigned significantly less blame than those in either of the two suboptimal conditions. This effect emerged even though all decisions were made in the same (hypothetical) context of insufficient knowledge, and that each type of decision produced the same negative outcome. The authors replicated this effect across seven experiments with different manipulations, including varying the consequences of the action and the degree of explanation regarding the actor's intentions. De Freitas and Johnson (2018) concluded that the most important factor in this bias is the tendency to *ignore the actor's mental state*, i.e., to expect them to behave optimally even when this is not possible from the actor's point of view.

To date, there is hardly any research on linguistic background as a potential mediating factor in attribution biases, despite the wide-ranging implications such biases may have on social judgements in general, and the previously discussed Foreign Language Effect (FLE) findings in particular. The present paper is a first attempt at bridging this gap by exploring whether the FLE modulates the optimality bias in blame attribution. Specifically, we aim to replicate De Freitas and Johnson's (2018) work with slight modifications to the design. More specifically, we investigate whether the optimality bias in blame attribution is mitigated by the FLE. The original experiments had three levels of optimality (*best*, *middle*, *worst*), but found no significant difference between the two sub-optimal conditions. As we are adding a target language manipulation to our designs (L1 vs. L2), we will include only two levels of optimality.

In the following, we will report two separate experiments. The first experiment compares optimality bias across two speaker groups (L1 vs. L2 speakers of English) using vignette materials in English. The second experiment compares the effect across two target languages (Finnish [L1] vs. English [L2]) within a population of Finnish-English bilinguals.

In line with the original study, we expect that participants should ascribe *more blame* for a negative outcome to a hypothetical actor who unknowingly chooses the *worst* course of action (*suboptimal* condition) than to a hypothetical actor who unknowingly chooses the *best* course of action (*optimal* condition). We expect this to happen even though (a) the consequences of the choice are equally negative and (b) the actor is described as having insufficient information in each case. More crucially, under the assumption that this effect is mitigated by the FLE, we also expect an interaction between condition and target language. Specifically, as a result of reduced emotional involvement in L2, we predict that there should be a reliably weaker optimality bias in blame judgements when participants are tested in their second language (L2), compared to when they are tested in their first language (L1).

6.2 Method

6.2.1 Pre-registration

Hypotheses (see above), methods, and analyses (indicated in the results section) were pre-registered on the Open Science Framework (https://osf.io/w6jvs/?view_only=1e5893c1a5ab48558cc4fd8b5edbb8a9).

6.2.2 Participants

Three groups of participants were recruited across the two experiments; an L1 English-speaking monolingual group, a bilingual Finnish-English group, and a bilingual group that consisted of L1 speakers of various languages with English as their L2. All participants resided in the United Kingdom at the time of taking part in the experiment. In both experiments, bilingual participants were asked to fill out a questionnaire regarding their language background (see Appendix A). Bilingual participants were defined as speakers who are fluent in their first language and in English as their *second language*. Bilingual participants who reported having learned English before the age of six and/or having parents who speak English as their L1 were not included in the final sample. This cut-off point was chosen to exclude 'early bilinguals', i.e. participants who have learnt English from early childhood and/or in a home setting. Participant samples and further exclusion criteria per experiment are described in more detail in the following sub-sections.

In Experiment 5, an initial sample of 186 participants was recruited through convenience sampling on social media. Of these, 25 were excluded for having incomplete datasets due to technical problems in online data transfer. Another 17 were excluded for incorrect answers to comprehension questions. Finally, 25 were excluded from the bilingual subgroup for learning English before the age of 6 or having English L1 parents. The final sample consisted of 119 participants, aged from 19 to 63 years (M = 26.02, SD = 8.58). Of these, 56 were bilinguals from various L1 language backgrounds, and 63 were monlingual English speakers. Ninety-one of the 119 participants identified themselves as female, 25 as male, and 3 declined to reveal their gender. Table 32 provides a more detailed breakdown of the condition counts and gender distributions in Experiment 5.

Table 32 Participant numbers and gender distribution per condition in Experiment 5

	Bilingual		Monolingual English	
	Optimal	Suboptimal	Optimal	Suboptimal
N	30	26	36	27
% Male	30.00	23.08	16.67	14.81
% Female	70.00	73.08	77.78	85.19
% Other/Not say	0	3.85	5.56	0

In Experiment 6, a sample of Finnish-English bilinguals residing in the UK was recruited, again through social media. Half of the participants completed the study in their L1 language (Finnish), and half in their L2 (English). Of an initial set of 331 respondents, 59 gave incorrect answers to comprehension questions, and another 27 were excluded for having learnt English before 6 years of age. Finally, data sets from 34 respondents were incomplete and thus removed. The final sample therefore included 211 participants, of whom 103 had been randomly assigned to Finnish (L1) and 108 to English (L2) as the target language for testing. Participants ranged in age from 18 to 71 years (M = 36.05, SD = 11.72). Of the final sample, 187 participants reported to be female, 23 male, and one participant declined to reveal their gender. Table 33 shows a more detailed breakdown of the condition counts and gender distributions in Experiment 6.

Table 33 Participant number and gender distribution per condition in Experiment 6

	Finnish (L1)		English (L2)	
	Optimal	Suboptimal	Optimal	Suboptimal
N	48	55	51	57
% Male	10.42	10.91	13.73	8.77
% Female	89.59	87.27	86.27	91.23
% Other/Not say	0	1.82	0	0

Bilingual participants' reported age of English acquisition was comparable across the two studies (Experiment 5: M = 9.34 years; Experiment 6: M = 9.21 years). Bilinguals in Experiment 5 reported to have lived in the UK for 5.20 years on average. Bilinguals in Experiment 6 reported a longer average length of stay in the UK (9.7 years). For a full breakdown of AoA and length of stay by experiment and condition see Tables 34 and 35 below. Participants were asked to rate their English (L2) proficiency in terms of speaking, reading and writing on a scale from 1 "very poor" to 7 "excellent". After summing the scores across the three sub-scales (speaking, reading, and writing), self-assessed proficiency could range from 3 (lowest) to 21 (highest). The mean self-assessment scores were very high both in Experiment 5 (M = 18.93, SD = 2.59) and in Experiment 6 (M = 18.82, SD = 2.24). There was no reliable difference in self-assessed proficiency between the bilingual groups in the two experiments (p = 0.62 by Mann-Whitney U-test). Within Experiment 6, the bilingual subgroup who completed the task in English was slightly (but not reliably, p = .092) higher in self-assessed English proficiency

(*M* = 19.07, SD = 2.28) than the subgroup who completed the task in Finnish (*M* = 18.55, SD = 2.18).

Table 34 Bilinguals' self-reported length of stay in the UK and age of L2 acquisition (means and SDs in years), broken down by condition for Experiment 5

	Optimal		Suboptimal	
	M	*SD*	*M*	*SD*
Length of Stay	6.01	5.18	4.15	3.60
AoA	8.97	2.51	9.77	4.33

Table 35 Bilinguals' self-reported length of stay in the UK and age of L2 acquisition (means and SDs in years), broken down by condition for Experiment 6.

	Finnish (L1)				English (L2)			
	Optimal		Suboptimal		Optimal		Suboptimal	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Length of Stay	9.39	8.88	8.94	9.28	8.34	7.60	11.83	11.24
AoA	9.17	1.58	9.20	1.39	9.00	1.08	9.44	2.04

6.2.3 Materials

Both studies were carried out online using *Experimentum* (DeBruine, 2019), a platform for online surveys set up by the University of Glasgow School of Psychology and Institute of Neuroscience and Psychology. All materials used in the studies were available in both English and Finnish. Finnish materials for Experiment 6 were translated from English by an L1 Finnish (English L2) speaker, and cross-translated by two other L1 Finnish speakers (who currently reside in Finland) to ensure compatibility.

The vignette used in the study was adapted from Experiment 5 in De Freitas and Johnson (2018). The original vignette included three levels of optimality ("best", "middle", "worst"), but since the original paper did not find a difference in blame between the two suboptimal conditions, we decided to implement only two choice conditions for the sake of simplicity. The third ("middle") option was still included in the vignette in order not stray too much from original setup, but only the "best" and the "worst" option were used as choices made by the described actor (manipulated conditions). The vignette was therefore as follows:

A doctor working in a hospital has a patient who is having hearing problems. This patient has three, and only three, treatment options. The doctor believes that all treatment options have a 70% chance of giving the patient a full, successful recovery. But in fact, the doctor's belief is wrong. Actually:

If she gives the patient treatment LPN, there is a 70% chance that the patient will have a full recovery.

If she gives the patient treatment PTY, there is a 50% chance the patient will have a full recovery

If she gives the patient treatment NRW, there is a 30% chance the patient will have a full recovery.

The doctor chooses treatment (LPN or NRW) [manipulated between conditions], and the patient does not recover at all. The patient now has permanent hearing loss.

There were two versions of the vignette; in the *optimal* condition the hypothetical doctor was described to have chosen the 'optimal' treatment (LPN, 70% efficacy) and in the *suboptimal* condition the doctor had chosen the 'suboptimal' treatment (NRW, 30% efficacy). In both cases, the doctor was described as erroneously assuming equal efficacies of the treatments. The described outcome remained the same across conditions, with the hypothetical patient suffering permanent hearing loss regardless of the treatment that was administered.

A five-item "blame questionnaire" was designed to measure participants' responses to the narratives. The responses were collected on 9-point Likert scales (cf. De Freitas & Johnson, 2018) ranging from 1 (low blame) to 9 (high blame). The items addressed five different aspects of the blame judgements: (1) how much the doctor is to *blame*; (2) how much *responsibility* the doctor had; (3) how much the doctor deserved *punishment*; (4) how *seriously wrong* the doctor's decision was; and finally, (5) how *confident* the participant was in making their judgement. The last item (5) was not considered to be a direct measure of blame attribution; it rather served as an additional control metric. Full wordings of the relevant questions can be found in Appendix E. In addition, there were three comprehension questions about the content of the vignettes which were also taken from De Freitas and Johnson (2018). Comprehension questions can also be found in Appendix E. Participants were excluded if they gave wrong answers to either of the first two of the comprehension questions. The third comprehension question was not used as an exclusion criterion, due to high numbers of participants answering this question incorrectly, regardless of target language. However, this comprehension question was included in exploratory analyses (see results section).

6.2.4 Design and Procedure

In Experiment 5, all participants completed the experiment in English. We compared two groups of participants (L1 vs. L2 speakers of English) in two conditions (optimal vs. suboptimal) using a 2 × 2 between-subjects design. Assignment of participants to experimental conditions (optimal vs suboptimal) was determined at random. In Experiment 6, Finnish-English bilinguals were tested in a 2 × 2 between-subjects design crossing target language (Finnish [L1] vs. English [L2]) with condition (optimal vs. suboptimal). Participants were randomly allocated to one of the four design cells: Finnish-optimal, Finnish-suboptimal, English-optimal, or English-suboptimal. Each participant read only one vignette.

Both studies were conducted online, and each participant was sent a link to complete the experiment. Bilingual participants were first asked to fill out a short questionnaire assessing linguistic background and English (L2) proficiency. Monolingual English speakers skipped this step. Participants were then asked to read vignette allocated to them, followed by the five-item blame questionnaire (choosing appropriate scale-points via mouse click). After the blame items, participants were asked to answer the three comprehension questions about the vignette. All participants were then fully debriefed via a debriefing page. The procedure took less than 10 minutes to complete.

6.2.5 Ethics

The experiment was carried out in full compliance of the BPS Code of Ethics and Conduct (2018) and approved by the University of Glasgow College of Science and Engineering Ethics Committee.

6.3 Results

6.3.1.1 Power

Power analyses were conducted prior to recruitment of participants, using the PANGEA application (jakewestfall.org/pangea/). The analyses suggested that, assuming a conventional 'medium' effect size, 120 participants were needed to achieve 69% power, and 160 to achieve 80% power. This suggests that the final samples for Experiment 5 (N = 119) and Experiment 6 (N = 211) were reasonably sensitive to the effects of interest, although imbalances in the design (due to participant exclusion) could lower the actual power figures relative to the 'idealised' calculations reported here.

6.3.1.2 Blame scores

We combined rating responses to the first four items of the blame questionnaire (covering *blame*, *responsibility*, *punishment*, and *seriously wrong*) into a single *blame composite score* by summing them up. Since participants gave scores from 1 to 9 on the Likert scales, blame composite scores ranged from 4 (low blame) to 36 (high blame). This was treated as a continuous variable in subsequent analyses. Reliability analyses based on the R package psych (Revelle, 2018) confirmed excellent internal consistency of the 4-item composite scale, with 95% CIs for *Cronbach's alpha* of [0.923, 0.959] in Experiment 5 and [0.930, 0.957] in Experiment 6 (established via bootstrapping over 10,000 resamples per study).

6.3.1.3 Experiment 5

Table 36 shows means and SDs of the blame composite scores in each participant group and condition and the violin plot in Figure 32 provides corresponding distributional information. Participants in the optimal condition gave lower blame scores than those in the suboptimal condition. Moreover, bilinguals (performing the task in L2) tended to attribute more blame than monolingual speakers (performing the task in L1) regardless of condition.

Table 36 Means and SDs for blame attribution scores across participant group and optimality condition in Experiment 5

	Condition					
	Optimal		Suboptimal		Overall	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*
Bilingual (L2)	11.67	7.23	24.96	6.23	17.84	9.48
Monolingual (L1)	8.56	4.35	22.96	5.47	14.73	8.65
Overall	9.97	6.00	23.94	5.88		

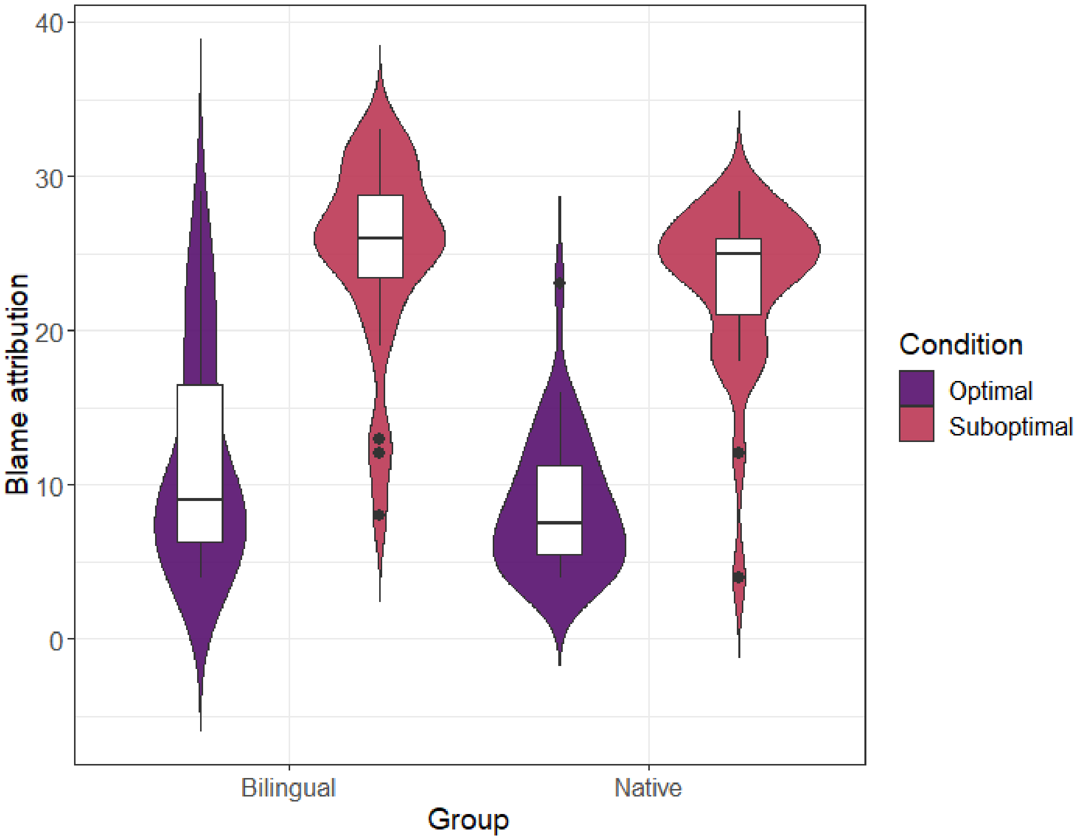

Figure 32 Blame scores by participant group and Optimality condition Experiment 6

A 2 × 2 between-subjects ANOVA was performed to test the effects of Group and Optimality on blame attribution. Overall, participants in the optimal condition attributed less blame than those in the suboptimal condition, resulting in a strong main effect of Optimality [$F(1,115) = 165.773$, $p < 0.001$, $\eta^2 = 0.577$]. A significant effect of Group was also found [$F(1,115) = 5.934$, $p = 0.016$, $\eta^2 = 0.021$], confirming that the bilingual group gave reliably higher blame scores than the monolingual group. The expected interaction between the two predictors was not confirmed [$F < 1$]. The optimality bias in Experiment 5 was therefore not mitigated by the FLE.

6.3.1.4 Experiment 6

Descriptive data for Experiment 6 are provided in Table 37 and Figure 33 below. Again, participants gave clearly higher blame scores in the suboptimal than in the optimal condition. In contrast to Experiment 5, overall blame scores were comparable across L2 vs. L1 conditions. Also note that optimality condition differences in the means were in the *opposite* direction to the expected FLE: For English (L2), the suboptimal-optimal contrast amounted to 23.46 – 8.73 = 14.73 blame-score units, and for Finnish (L1) to 21.75 – 10.65 = 11.10 blame-score units.

Table 37 Means and SDs for blame attribution scores across participant group and optimality condition in Experiment 6.

	Condition					
	Optimal		Suboptimal		Overall	
Language	*M*	*SD*	*M*	*SD*	*M*	*SD*
English (L2)	8.73	4.48	23.46	7.67	16.50	9.73
Finnish (L1)	10.65	7.90	21.75	7.58	16.57	9.49
Overall	9.66	6.41	22.62	7.64		

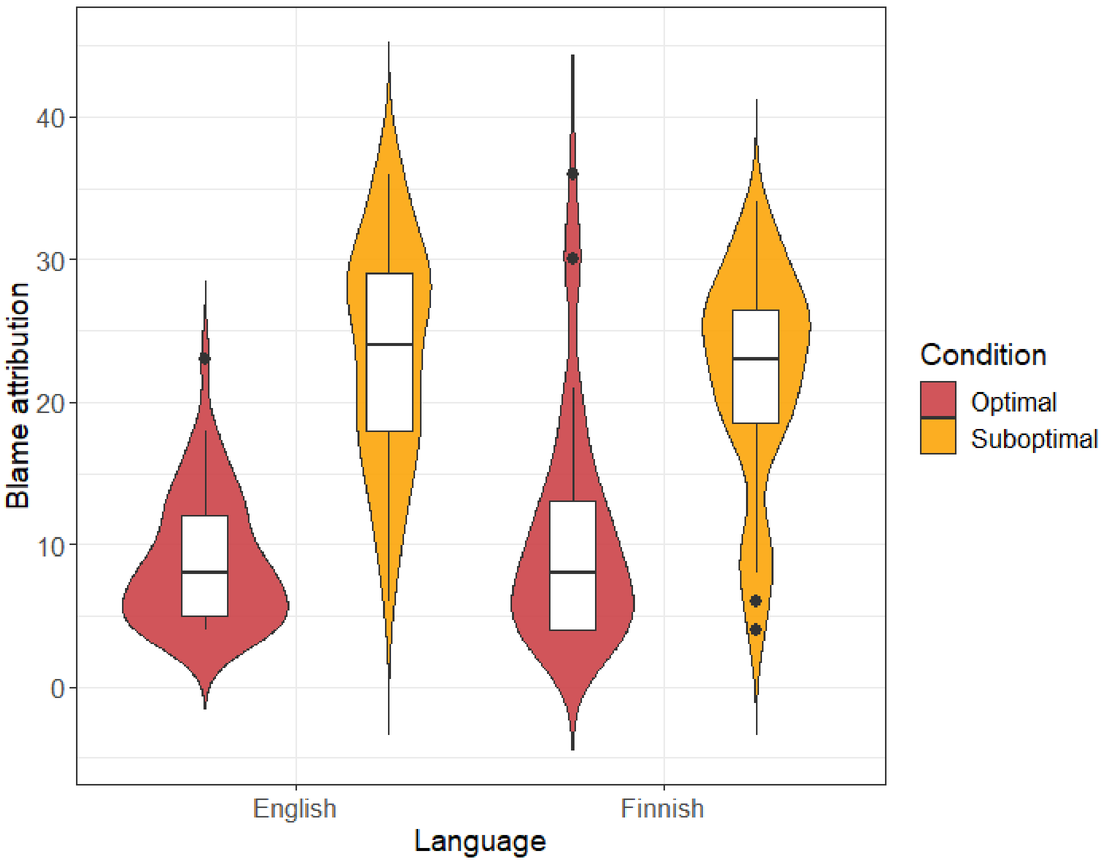

Figure 33 Blame scores by Target Language and Optimality condition Experiment 6

A 2 × 2 between-subjects ANOVA confirmed only one significant effect, namely the main effect of Optimality [$F(1,207) = 176.748$, $p < 0.001$, $\eta^2 = 0.456$]: Participants in the suboptimal condition gave higher blame scores than those in the optimal condition.

The main effect of Target Language was not significant [$F < 1$]. The interaction between Optimality and Target Language was marginal [$F(1,207) = 3.467$, $p = 0.064$, $\eta^2 = 0.009$] and in the opposite direction to the expected FLE.

6.3.1.5 Exploratory analyses

We conducted further analyses to investigate additional factors that may have affected the blame judgements. These analyses were not pre-registered, but are reported for completeness and to inspire future work.

6.3.1.5.1 Judgement confidence

Participants' confidence scores were measured by item (5) in the blame questionnaire. Since responses to this question were measured on a single, discrete but rank-ordered 9-point Likert scale, we analysed these data using ordinal logistic regression, as implemented in the R package ordinal (Christensen, 2019).

In Experiment 5, average confidence ratings did not seem to differ between the optimal (M = 6.39, SD = 2.00) and suboptimal condition (M = 6.30, SD = 1.40). Bilingual speakers (M = 6.14, SD = 1.85) tended to be slightly less confident overall than monolingual speakers (M = 6.54, SD = 1.64), but the ordinal logistic regression analysis actually revealed no reliable main effect or interaction effects (all *p*s > .2).

Ordinal logistic models of the confidence ratings in Experiment 6 showed a reliable Optimality main effect (*b* = -0.562; p = 0.023): irrespective of Target Language condition, participants in the optimal condition (M = 7.64, SD = 1.87) were more confident in their judgements than participants in the suboptimal condition (M = 6.48, SD = 1.85). By contrast, the main effect of Target Language, as well as the Optimality × Target Language interaction, did not approach significance in the confidence ratings (*p*s > .4).

6.3.1.5.2 Third comprehension question

As explained earlier, participants had to answer the first two comprehension questions correctly to be included in the main analyses. The third comprehension question (“*Did the doctor have any way of knowing that this belief about the probabilities was false or was it outside her control?*”) actually turned out to be somewhat problematic. In Experiment 5, 70 participants unexpectedly answered this question with “yes”; only 47 said “no” (as expected), and another two participants skipped this question altogether. Therefore, most participants (58%) answered this question in an unexpected manner. In Experiment 6, 82 participants unexpectedly answered “yes”, compared to 128 “no” responses and one participant skipping the question. While more in line with our expectations, the proportion of participants giving the ‘wrong’ answer was still quite large in Experiment 6 (38%).

Binary logistic regression analyses were conducted to explore whether there would be any cross-condition differences in answering the third comprehension question correctly. No clear main effects or interactions were established in either of the two experiments (all *p*s > .2). Hence, answering the third comprehension question correctly was unlikely to be predictive of the blame attribution scores of the main analyses.

6.3.1.5.3 Length of stay and Age of Acquisition as a predictors

As suggested in Tables 34 and 35 above, there were slight imbalances in length of stay in an English speaking country and in age of acquisition of English across the bilingual samples per condition. We therefore conducted additional multiple regression analyses in order to assess were these two variables were predictive of the observed blame ratings.

For Experiment 5, only bilingual participant data were considered, as we did not have information about the age of acquisition or length of stay in English speaking country for the English L1 speakers. Age of acquisition (AoA), Length of Stay (LoS), optimality condition (Condition), and all possible two-way interactions between these predictors, were included in the model as predictors of the blame composite scores.

Table 38 Regression table for Experiment 5, Bilinguals only

Predictor	*b*	*b* 95% CI	sr^2	sr^2 95% CI	Fit (R^2)
(Intercept)	31.13**	[21.92, 40.34]			
LoS	-1.13	[-2.45, 0.19]	.02	[-.03, .07]	
Condition	-15.72**	[-26.64, -4.80]	.06	[-.02, .15]	
AoA	-1.63**	[-2.63, -0.63]	.08	[-.01, .18]	.621**
LoS:Condition	-0.88	[-1.76, 0.00]	.03	[-.03, .09]	
LoS:AoA	0.16*	[0.03, 0.29]	.05	[-.02, .12]	
Condition:AoA	0.65	[-0.51, 1.81]	.01	[-.02, .04]	

Note. *** indicates p < .05. ** indicates p < .01.**

As seen in Table 38, the regression model confirmed the previously established main effect of Condition even when variation in Length of Stay and AoA was accounted for: the reliably negative estimate for Condition shows that blame judgments were harsher in the suboptimal condition. Interestingly, Age of Acquisition of English also had a significant effect; earlier acquisition of English predicted harsher blame judgments. There was also an interaction between Length of Stay and Age of Acquisition, suggesting that the effect of AoA was mitigated by LoS to some extent.

For Experiment 6, Age of Acquisition (AoA), Length of Stay (LoS), optimality condition (Condition), and test language (Language) were entered into the model as predictors of the composite blame judgments. We also included the two-way interactions between each of the predictors.

Table 39 Regression table for Experiment 6: Finnish speakers tested in Finnish or English

Predictor	*b*	*b* 95% CI	*sr*^{2}	*sr*^{2} 95% CI	Fit (*R*^{2})
(Intercept)	17.40**	[8.81, 26.00]			
LoS	-0.25	[-0.73, 0.23]	.00	[-.01, .01]	
AoA	-0.05	[-0.98, 0.87]	.00	[-.00, .00]	
Condition	-14.83*	[-28.25, -1.40]	.01	[-.01, .03]	
Language	-3.55	[-15.93, 8.82]	.00	[-.00, .01]	
LoS:AoA	0.02	[-0.02, 0.07]	.00	[-.01, .01]	.479**
LoS:Condition	-0.21	[-0.44, 0.03]	.01	[-.01, .03]	
AoA:Condition	0.42	[-1.06, 1.91]	.00	[-.00, .01]	
AoA:Language	0.39	[-0.94, 1.72]	.00	[-.00, .01]	
Condition:Language	3.73	[-0.16, 7.62]	.01	[-.01, .03]	

Note **.* indicates p < .05.**

As Table 39 shows, only the effect of optimality condition was significant (as in the pre-registered main analysis). The interaction between Condition and Language was marginal (p = 0.06) and it should be noted that its direction suggested the opposite pattern to the hypothesised FLE (same as in the pre-registered main analysis).

6.4 Discussion

In line with De Freitas and Johnson (2018), we expected blame scores to be lower in the optimal condition than in the suboptimal condition. Both studies fully supported this hypothesis, showing clear evidence for an optimality bias in blame attribution. We also hypothesised that there would be an interaction between Language/Group and Condition, such that the difference in blame judgments between the two conditions (optimal vs. subobtimal) would be smaller in L2 than in L1. This hypothesis was clearly not supported. In Experiment 5, L2 speakers were found to provide reliably higher blame attribution scores than L1 speakers, regardless of condition. In Experiment 6, no reliable difference between language conditions was found; if anything, there was a marginal interaction suggesting that the optimality bias in blame judgements was actually somewhat higher in L2 than in L1. In other words, the optimality bias in blame attribution did not appear to be modulated by a Foreign Language Effect (FLE) – or at least not in the direction we originally hypothesised.

Interestingly, in the exploratory analyses, we found lower age of L2 acquisition to be predictive of higher blame scores, and this effect to be mitigated the longer the participants have stayed in an English-speaking country. Although this pattern was found only in Experiment 5 (bilinguals from various L1 backgrounds) and not in Experiment 6 (Finnish L2 speakers of English), this may point to the importance of controlling for these variables more carefully in future research on this topic. In Experiment 6, the Finnish participants completing the study in English varied in duration of residence in the UK from a minimum of 3 months to a maximum of 50 years (average 10 years). In comparison, bilinguals in Experiment 5 only ranged in duration of residence from 2 months to 17 years (average 5 years).

6.4.1 The processes of blame attribution

In both experiments, the hypothetical actor faced significantly more blame for the same tragic outcome when they (unknowingly) made a suboptimal rather than an optimal choice. Thus, we replicated the findings from De Freitas and Johnson (2018), and found an optimality bias in blame attribution. Findings such as these are consistent with the *person-as-reconstructor* theory of blame attribution (Kahneman & Tversky, 1982; Kahneman & Miller, 1986). According to this theory, tragic outcomes motivate observers to reconstruct events *after* they happen, considering alternative choices and blaming the agent for failing to act otherwise. The doctor in our vignettes had three choices, which means that they *could* have acted differently. As a result, we observed higher blame judgements in the suboptimal condition.

This may also be explained by the *Path Model of Blame* (Guglielmo & Malle, 2017), which argues that blame is assigned systematically. Once causality is determined, observers assess whether the action was intentional. If the action was unintentional, observers then assess preventability. Our vignette was based on an unintentional scenario, so according to the theory, degree of preventability should guide blame judgements. In the optimal condition, the outcome was clearly not preventable because the patient suffers hearing loss even when the doctor picks the 'best' treatment option. In the suboptimal condition, however, it is likely that participants believed the outcome could have been prevented, had the doctor chosen the 'better' treatment. Thus, participants in the suboptimal condition seemingly based their judgments on potential alternative outcomes, while ignoring the doctor's mental state. Interestingly, exploratory analysis showed that in Experiment 6, participants in the optimal condition reported significantly more confidence in their judgement than those in the suboptimal condition, which could be seen as support for this kind of explanation.

Cushman (2008) argues that moral judgements involve two processes. The first one is triggered by negative consequences, where we search for an agent who is *causally responsible.* The second process is determined by analysing mental states, where blame is assigned only if the agent *believed* the action would cause harm. In this model, causality and foreseeability are separate processes, so causation and blame should not become conflated in moral judgements. However, our findings suggest that observers often make this mistake. Participants did not appear to engage in the second process when forming their moral judgements, i.e., they ignored the actor's viewpoint and beliefs. This contradicts the idea of two separate processes, or alternatively, suggests that the second process was given little consideration by participants: while the hypothetical doctor was causally responsible for her patient's hearing loss, analysing her mental state should have resulted in equal blame judgements across conditions, which was clearly not what the data showed.

6.4.2 The FLE in blame attribution

De Freitas & Johnson (2018) argue that factors inhibiting participants from considering the actor's mental state should enhance the optimality bias in blame attribution. Based on this assumption, and considering that emotionality might play a role in inhibiting the adoption of the actor's viewpoint, our second hypothesis was that the optimality bias in blame attribution should be stronger in L1 than in L2, particularly because previous demonstrations of the Foreign Language Effect (FLE) have pointed to reduced emotionality in L2.

In Experiment 5, we found that using L2 did not facilitate participants to think 'more rationally' about the actor's actual beliefs. Rather, L2 speakers apportioned generally more blame than L1 speakers. In Experiment 6, we found a marginal interaction in the opposite direction to our expectations, i.e., the optimality bias in blame judgements was slightly stronger in L2 than in L1. How can these unexpected results be reconciled with previous findings on the FLE?

It is possible that the FLE, by reducing emotionality, promotes consequentialist, utilitarian moral judgements. When using a foreign language, people become less sensitive to intentions and beliefs and more sensitive to outcomes (see also: Hayakawa, Costa, Foucart, & Keysar, 2016). Previous research on the FLE in moral judgement has indeed been confined to dilemmas involving utilitarian decision-making, i.e. the 'trolley' and 'footbridge' dilemma (Cipolletti et al., 2015; Corey et al., 2017; Costa, Foucart, Arnon, et al., 2014; Geipel et al., 2015a). The present study is novel in applying FLE to the attribution domain, which involves judging the intentions and actions of another person.

We conjecture that emotional involvement – in the sense of enhanced *empathy* (discussed below) – may actually be a *requirement* for considering a situation from another person's perspective. Under this view, diminishing emotion (e.g., via the FLE) might enhance the optimality bias in blame attribution, and thus partially account for the findings in both Experiment 5 (where bilinguals were found to be harsher in their blame judgments than L1 speakers) and Experiment 6 (where the optimality bias was found to be slightly stronger in L2 than in L1).

Masto (2015) argues that empathy is a crucial aspect in the forming of moral judgements. It is not enough to just observe an actor's behaviour to assess whether it is morally right, but we must also make additional evaluations regarding the motivations and thought processes of others. Previous research suggests that considering an action from *the perpetrator's* point of view can indeed reduce the severity of blame judgements. For example, in a mock-trial paradigm, Haegerich & Bottoms (2000) presented participants with a patricide scenario where a hypothetical child defendant claimed to have committed the crime in self-defence following years of abuse. Participants in the experimental condition were instructed to take the perspective of this child and imagine how they would feel and think in the same situation. This resulted in significantly lower blame judgements compared to a control group where no such instructions were provided.

Encouraging observers to think from the actor's perspective would likely also mitigate the optimality bias by directing focus away from the existence of alternative options and towards the key fact that these options are redundant (because the actor is not aware of their importance). Increased perspective-taking and empathy towards the 'doctor' in our vignettes may make participants realise that the outcome was not preventable.

Some research suggests that bilinguals may actually have advanced executive functions that are advantageous for perspective-taking (e.g. Greenberg, Bellana, & Bialystok, 2013). However, this has primarily been demonstrated for *early bilinguals,* especially those with equal proficiency in both languages (see: Rubio-Fernández, 2017). The purported bilingual advantage may not exist in *late bilinguals*. For example, Ryskin, Brown-Schmidt, Canseco-Gonzalez, Yiu & Nguyen (2014) studied visuospatial perspective-taking in a paradigm where participants completed a route-finding task by following instructions from an experimenter who had either the same or the opposite perspective. Late bilinguals struggled significantly more than monolinguals when taking opposite perspectives in their L2. Indeed, both of our experiments focused on late bilinguals, i.e. we deliberately excluded a relatively small number of bilingual participants who might have benefited from (potentially) enhanced executive functioning.

Mante-Estacio & Bernardo (2015) found a bilingual *disadvantage* in a Theory of Mind task where they asked participants to take the perspective of a character in a vignette. They studied the 'illusory transparency of intention' – originally demonstrated by Keysar (1994) – whereby readers falsely assume that characters in a story have access to the same information as the reader does. Participants were given vignettes describing a conversation and asked to judge whether the tone of a statement was sarcastic or genuine *from the perspective of the character in the vignette*. It was found that participants in L2 were more likely to focus on information that was clearly not available to the described character. Thus, these participants had more pronounced 'illusory transparency of intention' and found it more difficult to take the character's perspective in their foreign language.

Muted emotional resonance can also reduce the vividness of mental imagery. This was demonstrated by Hayakawa & Keysar (2018) on several measures. Bilingual participants reported experiencing difficulty in imagining objects in their L2. The same trend appeared also in a number of objective tasks. Participants were asked to mentally categorise objects based on visual attributes like shape. Bilinguals completing the task in their second language were less accurate than those completing the task in their L1. Importantly, Hayakawa & Keysar (2018) also found that bilingual participants completing the task in their L2 were more likely to agree to pushing a man in front of a train in the 'footbridge dilemma' and found that these participants rated the scenario as being far less visually vivid than those in L1.

As a whole, the present studies tap into a relatively new area of research. Few studies so far have investigated potential links between bilingualism and perspective-taking, and whether using a foreign language makes it difficult to imagine or consider the thoughts and feelings of others. The present research can make only tentative conclusions in this regard. In Experiment 5, L2 participants attributed significantly more blame than L1 participants, regardless of condition. In Experiment 6, the marginal interaction between language and condition suggested that L2 participants were somewhat more susceptible to the optimality bias in blame attribution than L1 participants. Together, these results could be accounted for by assuming decreased empathy (or perspective-taking ability) as a result of reduced emotional resonance in L2.

Finally, a potential issue arose from the third comprehension question in our experiments, which was also included in the original De Freitas and Johnson (2018) study: "Did the doctor have any way of knowing this belief about the probabilities was false or was it outside her control?" This question was answered incorrectly by a large proportion of participants (58% in Experiment 5 and 38% in Experiment 6) and could therefore not be used as an exclusion criterion. Participants were possibly thinking beyond what was stated in the narrative, and assumed that the doctor must have been careless in her prior research for having insufficient knowledge about the treatments' differing efficacies. That said, the exploratory analyses showed no systematic effects of language or condition in the likelihoods of answering this question incorrectly. Thus, answering this question incorrectly did not appear to be associated with participants' blame attributions.

6.5 Conclusion

The present experiments provide further evidence for the existence of an optimality bias in moral judgements. As such, they add to the existing literature on blame attribution and related theories. People find the existence of 'better' options important when morally judging the choices made by others, even when (a) all choices lead to the same (negative) outcome and (b) decision-makers are described as believing that all choices are equally optimal. More specifically, participants apportion reliably more blame (for the same negative outcome) when a described actor unknowingly made a suboptimal rather than an optimal choice. Against our expectations, we found that this optimality bias in blame attribution may be further enhanced by impaired perspective-taking, or empathy, in one's second language (L2). This contributes to the literature by suggesting that the Foreign Language Effect does not necessarily put bilinguals at an advantage in all types of moral decision-making scenarios. Indeed, there appear to be cases where reduced emotional resonance in L2 could potentially enhance irrational biases in moral judgement rather than diminish them.

Chapter 7 Overall discussion

This thesis has examined reduced emotional resonance in bilinguals’ L2 from three different angles: underlying reasons, measurement methods and potential implications for attribution and moral judgement. Chapter 2 discussed the underlying reasons and pupillometry as a measurement tool. Chapters 3, 4 and 5 focused on the measurement methods, looking at physiological measurement, cognitive paradigms and metacognitive measurement. Chapter 6 focused on the behavioural implications of reduced emotional resonance, looking at optimality bias.

7.1 Summary of the main findings

Experiment 1 replicated our previous pupillometry findings with a new set of participants and stimulus words (see Toivo & Scheepers, 2019). In terms of the underlying reasons, there was no convincing evidence for any of the standard predictors suggested in the related theoretical literature, such as age of acquisition or context of acquisition. Only bilingual participants' "preference to swear in L1" was found to have a modulating influence on pupillary responses to high versus low-arousal words in L2: Participants with a stronger preference to swear in L1 were found to have a weaker pupillary response to HA (as opposed to LA) words in L2. In other words, preference to swear in L1 was the only factor that was 'predictive' of reduced emotional resonance in L2 (as measured in pupil dilations).

While this finding does not align with the contextual learning theory, brain maturation theory, or the theory of language embodiment, there is some evidence suggesting swearing specifically is linked to reduced emotional resonance of L2 (Dewaele 2004, 2016; 2010a; 2010b, 2018). Dewaele (2016) has found that bilinguals often prefer to swear in their L1 and may overestimate the offensiveness of swearwords in L2. It has been argued this may be because the bilingual speaker attaches “a red flag” to the offensive words in L2, potentially overcompensating for the reduced emotional resonance of these words but having an awareness of the social implications of using offensive language. Hence, it is not entirely surprising that we found the preference to swear in L1 to be a predictor of reduced emotional resonance, but more empirical work is required to establish if this effect replicates, and how it relates to the theory.

Experiment 2 attempted to compare SCR measurement with pupillometry; we adapted Experiment 1 procedure to skin-conductance measurement using the same stimulus materials. We found no statistically significant interaction of word type and participant group. Intriguingly, the pattern in the SCRs suggested the exact opposite to our findings in pupillometry: Compared to L2 speakers of English, monolingual L1 speakers showed a stronger overall SCR to English stimuli, while at the same time showing smaller differences in SCRs to high vs. low-arousing words. These findings contrast findings from previous SCR studies (Caldwell-Harris et al., 2010; Harris, 2004; Harris et al., 2003; Harris et al., 2006). This may suggest that single-word SCR paradigms with many test items are more susceptible to habituation effects than designs that implement fewer items and use phrases as stimuli instead of isolated words.

In Experiment 3 we used a new, bilingual stimuli set with L1 English speakers and German-English bilingual speakers to explore why cognitive measurement is often inconclusive, and whether there is a positivity bias in how L2 words are embodied. We ran a pupillometry task and a Lexical Decision Task on the same participants.

In the LDT, we found that in by-group comparisons the German speakers had slower RTs overall, and the same was found in English (L2) when comparing by-language. This pattern was mirrored in the pupil findings, where the German group had higher pupillary responses, and also higher pupillary responses in L2 when compared by-language. This is consistent with the cognitive effort account we have discussed throughout the thesis.

In the LDT, we also found main effects of word type, suggesting that both negative and positive words were recognised better and faster than neutral words. There were no significant differences between positive and negative words, suggesting the RTs and word recognition were driven by arousal rather than valence. This is consistent with Kousta et al. (2009). The LDT detected no reduced emotional resonance in the German speakers (no interaction of word type and language), which is consistent with Kazanas and Altaribba (2016) and

Ponari and colleagues (2015). However, as we have attempted to highlight throughout the thesis, it is not meaningful to draw strong conclusions from cognitive paradigms until the field has found a methodological consensus and the systematic use of covariates is standardised.

Interestingly, in the pupil task of Experiment 3, the low arousing stimuli words, which acted as the baseline, elicited the strongest pupillary responses. Consequently, it was not possible to compare the two tasks. This unexpected finding gives ground for some methodological considerations, which will be discussed below. Unexpectedly, the negative stimulus items elicited stronger pupil response than positive words. This is inconsistent with previous pupillometry findings (for example, Kuchinke et al., 2007), and provides some evidence against the positivity bias in bilingual affective language processing.

Experiment 4 was a replication of Experiment 1 with new participants, and an affective rating task added at the end of each trial, rather than a word recognition judgement task. This was done to compare physiological measurement with metacognitive measurement. While the descriptive pattern in the pupil response data (Figure 28) was in line with the previously established pupillometry effects on reduced emotional resonance in L2, effects did not come out statistically significant, presumably because of too much noise in the data, and the fact that cross-condition differences emerged rather late in the considered time period.

The same is true for the rating data (Figure 25): while the three word categories predictably differed in terms of perceived arousal ratings, evidence for reduced emotional resonance in L2 was, at best, only descriptive. The interaction of word type and participant group was not significant in the cumulative link mixed model predicting word ratings. This falls in line with previous literature on bilingual affective ratings (e.g. Iacozza et al, 2017; Winskel, 2013), and suggests affective ratings do not necessarily capture reduced emotional resonance (or the differences are not large enough to detect with appropriate mixed effects models); embodiment and semantic access seem to be at least somewhat separate processes.

While the previous studies focused on the predictors and measurement methods of reduced emotional resonance, Experiments 5 and 6 approached the topic from a different angle. These experiments investigated optimality bias in bilingual decision-making, and whether the bias would be mitigated by the foreign language effect. We tested Finnish-English bilinguals (Experiment 6) and a general group of bilinguals from multiple countries and language backgrounds (Experiment 5) and found no foreign language effect in either sample. However, in Experiment 5 we found that bilinguals overall gave harsher blame judgments, and in Experiment 6 the Optimality condition*Language interaction was marginal, but to the opposite direction expected. We speculated these findings may be due to differences in perspective-taking in L1 and L2.

7.2 Methodological considerations

One of the aims of this thesis was to conduct systematic, methodological work to establish how reduced emotional resonance is best measured, and what factors should be considered in experiment planning and analysis stage. To this end, we compared two physiological measures (pupillometry and SCR), a cognitive paradigm (LDT) and a metacognitive measure (affective ratings).

From the methodological comparisons, the following three points can be concluded:
Firstly, affective ratings do not seem very effective at capturing reduced emotional resonance. This is consistent with previous findings from studies using affective ratings as a measurement of bilingual emotion (for example, Iacozza et al., 2017; Winskel, 2013). It is likely to be due to the different nature of words isolated from their context and concepts. Physiological measurement may tap onto the conceptual nature of words, measuring the affective response produced by the body, whereas affective ratings seem to simply measure a conscious assessment of each of the words. If this is true, it is unclear as to why affective ratings taken from databases map onto physiological responses (e.g. higher pupillary responses to words that have been rated high arousing).

Secondly, more systematic, methodology-focused work needs to be conducted into the use of cognitive paradigms. At the moment, cognitive paradigms are widely used and strong conclusions about the nature of emotional resonance of

L2 are being drawn from these studies. The underlying mechanisms are still uncertain, and the findings across the field are inconsistent (see for example: Conrad et al., 2011; Segalowitz et al., 2008; Winskel, 2013). When the use of covariates and translation equivalents differs across experiments, we cannot draw any strong conclusions from the findings of these experiments.

Thirdly, pupillometry seems to be the most consistent measurement method of reduced emotional resonance. Our findings from Experiment 1, and the descriptive findings from Experiment 4 are consistent with previous pupillometry work (Iacozza et al, 2017; Toivo & Scheepers, 2019), and replicate the effect with a new set of stimulus words.

7.3 Assessing pupillometry as a measurement tool

Overall, our pupil findings from Experiment 1 are consistent with previous physiological measurement work (Caldwell-Harris et al., 2010; Harris, 2004; Harris et al., 2003; Harris et al., 2006; Iacozza et al., 2017; Toivo & Scheepers, 2019). However, there are a few methodological issues we have observed that should be of interest to future pupillometry work.

Perhaps the most problematic issue is the difference in pupillary response to HA and LA words in L2. In Experiment 1 and 4, as well as our previous research (Toivo & Scheepers, 2019) it seems that the LA words elicit stronger pupil response in L2 than they do in L1. This, in turn, questions the origins of the interaction of word type and language/participant group, which we hold as a measure of reduced emotional resonance.

It is possible that high arousing words are less affected by cognitive effort because their processing is overall easier than the processing of low arousing words. Previous research has established that HA words typically have a processing advantage (Kousta et al., 2009). This highlights two potential issues, which warrant more methodological work. Firstly, it is extremely difficult to disentangle the effect of emotion from other potential confounding variables – these are introduced not only from using word stimuli in multiple languages, but also from the speaker language background. We believe we have made a substantial effort in controlling for the factors that contribute to the cognitive load the words have. In all the experiments in this thesis, we monitored participant proficiency and word recognition and controlled for several lexical covariates both at the stimuli selection stage as well as the analysis stage. These practices should be standardised across the field, especially when using cognitive paradigms.

Secondly, it is possible that the origins of reduced emotional resonance are in fact intertwined with, or due to cognitive effort, at least to some extent. In Experiments 1 and 4 we observed a main effect of participant group suggesting that overall the bilingual groups had a stronger pupillary response. This, in turn, suggests some degree of increased cognitive effort. The bilinguals we have tested are fully immersed in their L2 and are all very advanced L2 learners of English. Attributing the HA-LA findings solely to a bilingual cognitive effort effect and dismissing them as such would be a waste of rich data.

In terms of the heightened LA response, we cannot draw any strong conclusions from the present data. It is a possibility that the paradigm does not measure reduced emotional resonance at all, but the effect is driven by the underlying effect of cognitive effort instead. We did not explicitly ask the participants about their perception of reduced emotional resonance in L2 and whether they experience this. Thus, the heightened LA response and how that relates to the emotional and cognitive effort response, and the participant perception of L2 emotionality are topics that future methodological research should consider.

It is also possible these effects have nothing to do with cognitive effort per se. Instead of characterising the word type effect on pupil size as an "increase in response to HA words", it could be in fact more appropriately characterised as a "decrease in response to LA words". In other words, the 'calming' effect of the LA words might be reduced in bilingual speakers and consequently drive the effect. Only an appropriate 'emotionally neutral' baseline condition can resolve the issue (see discussion of Toivo & Scheepers, 2019 for more details), but as the research in this thesis shows, such a condition is difficult to establish due to the surprise/implicit word categorisation effects.

Another methodological issue arising from this thesis, across several experiments (see Experiments 1, 3 and 4), is the role of surprise as a confounding variable in pupil measurement. In Experiments 1 and 4, the neutral distractor items elicited the strongest pupillary response (in Experiment 4, only in the monolingual group). This was unexpected, as based on the arousal ratings, the magnitude of pupil response elicited by the neutral words should fall between the HA and LA words. In experiment 3, the low arousing words elicited the strongest pupil response, rendering it impossible for us to draw conclusions about reduced emotional resonance in this sample.

The common denominator of these findings is the proportion of occurrence of different stimulus categories in our experiments. In all cases the stimulus category constituting 1/3 of the stimuli elicited the strongest response, leading us to believe this is due to the participants detecting a semantic pattern and then being surprised when said pattern is violated. Perhaps the participants start implicitly categorising the stimuli as the experiment progresses, and deviations from the categorisation cause a surprise effect on the pupil response or interferes with how well the words are recognised. This finding warrants further work looking into pupillary response and stimuli proportions and highlights the importance of systematic methodological work. If we do not know what our measurement techniques are sensitive to, and fully understand the possible confounds, we cannot infer any actual effects we are interested in.

7.4 The use of lexical covariates

This thesis has aimed to address some of the methodological issues around measuring affective responses in bilinguals. One of the most striking problems in current literature is how stimuli are created and how the analyses are conducted. Throughout this thesis we have highlighted the importance of using controlled, well-balanced stimuli, and considering lexical covariates in both the stimulus selection stage as well as in the statistical analyses.

In Experiments 1, 3 and 4, we found that Lexical frequency had a significant effect of pupillary response (lower frequency effects elicit a stronger pupil response), and in some cases this effect was stronger in the bilingual groups. This effect was found on carefully balanced stimulus sets, which further supports the argument that including lexical covariates in the analyses is important to account for potential confounds. Lexical frequency and word length were also found to affect word recognition in Experiments 1-4.

Here, we will attempt to outline a guideline for creating stimuli and to the use of covariates to increase the quality of bilingual emotion research:
1. Stimuli sets should be balanced on a number of available lexical covariates that are known to affect word processing (such as length, frequency, concreteness, orthographic neighbours, valence, dominance and arousal). This should ideally be done algorithmically.
2. These lexical covariates, if possible, should also be included in the statistical analyses.
3. The use of translation equivalents should be avoided.

This approach should help create stimulus sets which take into account variation in responses that occurs due to lexical covariates (such as, words that are longer or less frequent are usually recognised more slowly). We appreciate lexical norms are not readily available for all languages, and not all research groups have the computational capacity to do algorithmic stimulus selection. Hence, well-designed stimulus sets and word databases should always be openly shared. There are also attempts to standardise psycholinguistic stimuli selection with computational approaches, such as the LexOps R package (Taylor, Beith, & Sereno, 2019). This package allows for the researcher to flexibly match their stimuli words on several lexical characteristics, and create a stimulus set controlling the Euclidean distance between these dimensions. That is, the differences across lexical covariates can be minimised, only focusing on the dimension the researcher wants to manipulate. Both the use of open stimuli sets and computational approach to stimuli matching should increase the quality of research and minimise the pervasive effect of cognitive effort and other confounds.

The lexical covariates should also be included in the analysis stage. This will allow for further control over the cognitive effort effect and will increase the accuracy of modelling. Increasing the number of predictors in a model will inevitably decrease the power of a given design, which is why we have suggested reducing the number of lexical covariates with a Principal Component Analysis. This approach helps to both account for the possible collinearity in the covariates (for example, long words are typically less frequent), as well as reduce the number of variables in the analysis models.

Trying to create a balanced stimuli set is difficult, and some variation will inevitably remain, which is why it is useful to account for this also when analysing the data. As can be seen in the experiments across this thesis, even when using balanced stimulus sets, the lexical characteristics can affect the results.

As discussed in chapter 4, the use of translation equivalents within one stimuli set is problematic, as it creates further dependencies within the data. Unfortunately, databases in languages other than English are scarce and typically less extensive, which complicates the use of stimuli in other languages. One possible approach to combat this is to use translations (see Experiment 3 method section under stimuli) but remove translation equivalents from the candidate pool of the language, which has a larger candidate pool. If translation equivalents are used, the dependencies they create should be taken into account when modelling the data (if using mixed effects models, there should be a by-subject random slope for the translation equivalents).

7.5 Limitations

It is possible that the LA-HA difference in bilinguals (LA words sometimes eliciting a stronger pupillary response) discussed above is, at least to some degree, due to cognitive effort. It is unlikely cognitive effort alone explains this, as the lexical covariates that are known to increase cognitive effort (such as length and frequency) were carefully controlled in the selection stage, as well as accounted for in the analyses. However, as speculated above, it is possible the effect of emotion cannot be fully disentangled from the effect of cognitive effort. With the present data we cannot make conclusions about this, but it is a potential confound which needs to be considered when interpreting the results.

The SCR study (Experiment 2) has perhaps the most considerable limitations of the experiments included in this thesis. Due to time constraints, it was not possible to link the SCR system with ePrime, which caused noise in the data. This noise was accounted for in the pre-processing of the data, but manual syncing has increased the potential for human error in the data. Comparing data from two separate groups of participants tested at different locations is also problematic - ideally, we would have measured pupillary response and SCR concurrently, but unfortunately this was not possible because of hardware availability.

Some of the model comparisons in Experiment 3 were computationally problematic, and we found p-values of 1, which is not feasible when comparing two different models. Given that model comparison is often held as the gold standard of obtaining p-values for mixed effect models, these findings are concerning and warrant further investigation into fitting and comparing GLMEMs.

Data from Experiments 3 and 4 had an unusual amount of noise (see pre-processing of pupil data in Experiment 3). This is possibly due to having to move the hardware to a different testing space twice during data collection. In Experiment 3 the noise did not increase the data loss, but in experiment 4 we have lost 36% of the data, which is not desirable in any way.

Experiment 4 was intended to replicate Experiment 1 fully, including the prediction of underlying reasons of reduced emotional resonance from a language background questionnaire. Due to time limits, we did not include this part in the thesis, but therefore the group sizes in Experiment 4 are unequal.

When collecting data for Experiments 1-4, effect size estimation and power calculation tools for linear mixed effects models were not as advanced as they are today. Simulation tools (e.g. DeBruine & Barr, 2019) have been developed and simulation of mixed effects models' data is now considered good practice prior to data collection (although estimating power through this would still require some prior knowledge on population parameters, which is not always possible without pre-existing data). We have simply based on our sample sizes on previous experiments where the effect was detected (Toivo & Scheepers, 2019), as was done in Experiment 3, or aimed at getting the largest and most diverse sample possible given the limited resources (Experiment 1 and 4). Given that the experiments do demonstrate the desired word type-participant interaction and are very unlikely to be overpowered due to the complex model structures involving random effects and covariates, we do not think there is a considerable issue with the power in these experiments. However, when designing future experiments, simulation prior to data collection should be included.

With the increasing use of Mixed Effects Models, other tools for calculating power and effect size have been developed in addition to simulation techniques (Brysbaert & Stevens, 2018). However, most of these techniques (as well as the simulation guides) only account for scenarios, where the model is relatively simple and does not necessarily have a “maximal” structure (Barr et al., 2013). This creates a trade-off of modelling accuracy versus estimating parameters prior to running the study. If the model structure is complex, assessing power will be more difficult and inaccurate, whereas if the model is simple and the tools for assessing power are readily available, some variation will not be captured by the model.

One approach is to simplify model structures, but as we have argued, this can be detrimental to the interpretation of the results when word stimuli are used. Power estimation for more complex random structures (particularly accounting for interactions) is a concern for psycholinguistics as a whole, and should be considered an area of interest for future research.

7.6 Future directions

The underlying reasons for reduced emotional resonance are not fully understood and the field requires more experimental evidence. Pupillary response may not vary enough to capture this. Simulation techniques for Generalised Linear Mixed Effects Models have developed since running Experiment 1, which should help with estimating whether the effect was not found due to inadequate power. We deem this unlikely, but it would be of interest to confirm this with an experimental procedure – it would be possible to simulate based on the existing data we have.

It would also be useful to look into the underlying reasons systematically using alternative techniques; developing a validated psychometric measure for this would be one option. This could also further the research on physiological and cognitive measures, and the origins of the reduced emotional resonance effect found in them.

A systematic investigation of the role of surprise in pupillary responses is warranted. Manipulating the proportions of stimuli would be of interest for this experiment to see how the surprise effect is formed. This could be done with word stimuli, but also with affective pictures as well as sounds to get a more comprehensive understanding of the possible surprise effect (or whether it is a surprise effect at all).

The work on cognitive measurement and as to why it is so inconclusive needs to be continued. Successfully comparing cognitive tests to a physiological measure could shed light into this question; repeating experiment 3 with previous materials from Toivo & Scheepers (2019) or creating a new stimulus set where the stimuli are manipulated on arousal and not on valence could yield more conclusive results in this regard. It would also be useful to conduct a meta-analysis on the studies that use cognitive paradigms. This would allow for systematic assessment of why the results are inconclusive and help identify what factors in experimental design and participant populations may affect the findings. Rather than producing new, inconsistent experimental evidence, the field should rigorously assess its methodology and identify whether the effects in fact are real.

Conducting skin-conductance measurement concurrently with pupillometry and using full sentences with target words (see Iacozza et al., 2017) would provide further information on the possible habituation effects (see chapter 3 introduction for further discussion on habituation effects) of SCR, and whether pupillometry indeed is a superior measurement technique.

Collecting more evidence about the FLE in attributions is required to understand the limits and origins of the FLE. A replication of experiments 5 and 6 would be useful. This could be done with an added level of information given to the participants as was done in one of the experiments of the original paper by DeFreitas and Johnson (2019). In other words, the vignette would be slightly altered to highlight that the doctor had no way of knowing whether their choice was optimal, and that they have done research to make the best-informed choice possible.

In chapter 6 we speculated the increased blame in the bilingual group may be due to impaired perspective-taking in L2 – this aspect should be investigated in more detail, potentially with an eye-tracking paradigm to establish whether perspective-taking affects bilingual decision-making. It would also be of interest to test other attribution biases, such as the fundamental attribution error, which is known to be sensitive to increased perspective-taking (Hooper, Erdogan, Keen, Lawton, & McHugh, 2015).

7.7 Contributions and conclusions

This thesis has examined the underlying reasons, measurement methods and implications of reduced emotional resonance. Its most substantial contribution is the systematic methodology-focused work we have conducted, pointing out inconsistencies and highlighting areas which require standardisation and more work. We have shed some light into the differences between words and concepts and how this difference may be reflected in measurement methods (physiological vs. metacognitive). We have also found some evidence suggesting swearing, and in particular the preference to swear in L1, may be a key concept when discussing predictors of reduced emotional resonance.

We have also extended the implications literature beyond moral decision-making, to attributions and whether attribution biases will be mitigated by the FLE. This has opened a new field of investigation and has provided a new potential explanation for why the FLE occurs. Alongside the dual pathway model, emotionality of a given decision-making situation, and possible focus on outcomes rather than intentions, future research should look into whether perspective-taking affects the FLE.

The study of bilingualism has been traditionally centred on the cognitive benefits and consequences of bilingualism, somewhat ignoring the breadth of topics around the bilingual speaker. It has been argued that research should strive to understand the experiences of the bilingual speaker beyond benefits and consequences of bilingualism (Grosjean, 2008). This thesis, exploring the bilingual emotional experience from multiple viewpoints, has attempted to address this and move past the simple monolingual-bilingual cognitive performance comparison. The study of reduced emotional resonance fundamentally stems from bilinguals’ self-reports of feeling “less” in their L2, and as more and more of world’s population live in bi- and multilingual settings, it is increasingly important we understand the experience of a multilingual speaker holistically.

The field of bilingualism and emotions is truly interdisciplinary and rich, but it suffers from inconsistencies and drawing strong conclusions from the findings produced by those inconsistent methodologies. This thesis has highlighted the importance of using lexical covariates and appropriate statistical methods when studying reduced emotional resonance in bilinguals, or bilingualism in general. Several fundamental methodological issues are yet to be solved – the field needs systematic method-oriented research to establish best practices.

Appendices

Appendix A: Language background questionnaires and coding keys
Language questionnaire: EXPERIMENT 1

1. Age: _______
2. Sex:___________
3. Country of origin: ______________
4. How long have you been in the country of your current residence?
_____ (years)_____ (months)
5. Native language (If you grew up with more than one language, please specify):_______________________

6. List the languages you know in order of proficiency (most proficient first):

1.	4.
2.	5.
3.	6.

7. If you have lived or travelled in other countries for more than three months, please indicate the name(s) of the country or countries, your length of stay, the language(s) you learned, and the frequency of your use of the language according to the following scale (circle the number in the table):

Never Rarely Occasionally Sometimes Frequently Very Frequently Always
1 ______ 2_______3___________4__________5__________6______________7_______

Country	Length of stay (cumulative)	Language	Frequency of use
			1 2 3 4 5 6 7
			1 2 3 4 5 6 7
			1 2 3 4 5 6 7

8. Provide the age at which you were first exposed to each language in terms of speaking, reading, and writing, and the number of years you have spent on learning each language

Language	Age first learned the language	Number of years spent learning (cumulative)

9. Write down the name of the language(s) used by your teachers for general instruction at each schooling level. If you switched language within a given school level, write a note such as “switched from X language to Y language at Grade Y”.
Primary/Elementary School: __________
Secondary/Middle School: __________
High School: __________
College/University: __________

10. Type in the box the age at which you started to learn each language in any or all of the following situations (if only one situation is relevant for one language, provide age information for only that situation).

Language	At home	At school	After immigrating to the country where spoken	At informal settings (e.g from nannies, friends)	Online games	Other (specify)

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

Activity	Language:___________	Language:_________	Language:________
Watching TV series/films/ online podcasts	________hrs	________hrs	________hrs
Listening to radio	________hrs	________hrs	________hrs
Reading for fun	________hrs	________hrs	________hrs
Online chatting/writing emails	________hrs	________hrs	________hrs
Reading for school/work	________hrs	________hrs	________hrs
Writing for school/work	________hrs	________hrs	________hrs

12. Estimate, in terms of hours per day, how often you speak your languages (both in person and online, for example over Skype/phone or an online chat) currently with the following people:

	Language:_______	Language:_________	Language:_________
Family members	________hrs	________hrs	________hrs
Friends	________hrs	________hrs	________hrs

Classmates	_______hrs	_______hrs	_______hrs
Coworkers	_______hrs	_______hrs	_______hrs
Partner/spouse	_______hrs	_______hrs	_______hrs

13. Which of your languages you became fluent with first:__________________

14. How often do you use each of the languages you have studied or learned for the following activities? Please circle the number in the table according to the scale below.

Never Rarely Occasionally Sometimes Frequently Very Frequently Always

1 ______ 2_______3___________4__________5__________6______________7______

Langu age	Thinking	Talking to yourself	Swearing	Expressing emotions **(a)**	Dreaming	Arithmetic **(b)**	Remembering numbers **(c)**
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7

(a). This includes shouting, showing affection, etc. **(b).** This includes counting, calculating tips, etc. **(c).** This includes telephone numbers, ID numbers, etc.

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

	Listening	Speaking	Reading	Writing
At home (with family)				
At home (with flatmates)				
With friends				
At school				
At work				

16. Please indicate how often you CURRENTLY use each language in the following contexts

Never Rarely Occasionally Sometimes Frequently Very Frequently Always

1 ______ 2_______3___________4__________5__________6______________7_______

Language	With family members	With friends	With a partner (e.g a girl-/boyfriend or a spouse)	At work	At school
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7

Coding of the questionnaires: EXPERIMENT 1

Questions 4 and 7 are coded together as cumulative stay in English-speaking countries. Code in months. For participants from countries such as Singapore, Hong Kong or Nigeria, where English is an official language, stay in their home country should be included in their cumulative stay.

Question 8 (Age of Acquisition) was coded as the total number of years spent learning English. Further, the age of acquisition (first exposure to English) was extracted from the question.
Question 10 (School instruction). University was excluded from the coding (all participants were students and going to university in the UK). If the participant had had general instruction at lower school levels in English, this was coded as 1, and if not, it was coded as 0.
Question 11 (English learning in different situations) is broken down by different situations of learning (hence there are 6 different sub-categories). Number of participants will have left the ‘after immigrating to a country where spoken’ question blank. Replace this with an age that is computed by subtracting participant’s answer to question 4 (‘How long have you been in the country of your current residence’) from their age. Otherwise if the participant has not said answered a sub-category, enter NA. if they have specified the situation ‘Other’, enter the age and situation in the box
Questions 11 and 12 (Number of hours per day spent using English in different activities) are coded separately as the total number of hours per day per each question.

Question 13 (order of acquisition) is coded as 0 and 1. Enter 0 if the participant became fluent in another language first, 1 if the participant indicated becoming fluent in English first.
Questions 14 and 16 are coded as a difference score between the languages participants listed. The questions are broken down into sub-categories and these sub-categories. The response scale ranges from 1 (never) to 7 (always). Positive integers indicate a ‘favour’ for the other language, negative integers indicate a ‘favour’ for English (for example if a participant has indicated 7 for their first language for a sub-category, and 2 for English, this would be coded as 5. If a participant had indicated 4 for English and 3 for their first language, this would be coded as -1). If the participant listed multiple languages, the score for each

sub-category is based on the difference score between English and the language which had a higher frequency score for that specific sub-category.
Question 15 (language dominance in different situations) is coded as the overall percentage of English across the situations. If the participant has written both, this is counted as English. If the participant has not responded to some of the situations, leave them out from your percentage calculation.

Language questionnaire: EXPERIMENT 2 and EXPERIMENT 3
1. Age: _______
2. Sex:___________
3. Country of origin: ______________
4. How long have you been in the country of your current residence?
______ (years)______ (months)
5. Native language (If you grew up with more than one language, please specify):_______________________

6. If you have lived or travelled in other countries for more than three months, please indicate the name(s) of the country, your length of stay, the language(s) you learned, and the frequency of your use of the language according to the following scale (circle the number in the table):
Never Rarely Occasionally Sometimes Frequently Very Frequently Always
1 ______ 2_______3________4__________5__________6________7_______

Country	Length of stay (cumulative)	Language	Frequency of use
			1 2 3 4 5 6 7
			1 2 3 4 5 6 7
			1 2 3 4 5 6 7

7. Provide the age at which you were first exposed to each language, and the number of years you have spent on learning each language

Language	Age first learned the language	Number of years spent learning (cumulative)

Language questionnaire: EXPERIMENT 4

1. Age: _______ 2. Sex:___________

3. Country of origin: ______________

4. How long have you lived in the UK? _____ (years)_____ (months)

5. Native language (If you grew up with more than one language, please specify):_______________________

6. List the languages you know in order of proficiency (most proficient first):

1.	3.
2.	4.

7. If you have lived or travelled in other countries for more than 3 months and used English as your main communication language, please indicate the name(s) of the country and your length of stay

Country	Length of stay (cumulative)

8. Provide the age at which you were first exposed to each language, and the number of years you have spent on learning each language

Language	Age first learned the language	Number of years spent learning (cumulative)

9. Which language was used by your teachers for general instruction at each schooling level? Also indicate how many years of each you have completed.

Primary school	Middle school	High school	Undergraduate	Postgraduate
___ years	____ years	____ years	____ years	____ years

10. Type in the box the **age** at which you started to learn each language in any or all of the following situations (if a situation is not relevant for a language, please write NA).

Langua ge	At home	At school	After immigrating to the country where spoken	At informal settings (e.g from nannies, friends)	Online games	Films, TV, music	Other (specify)

11. Estimate how often you do the following activities in each of your languages.

Never *Every year* *Monthly* *Weekly* *Every day* *Several times a day*
1 ________2__________3__________4_________5_________6__________

Langu age	Watching TV series /films	Listening to radio/pod casts/audi obooks	Online chatting/ sending emails	Reading for fun	Reading for work/univ ersity	Writing for work/universi ty
	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6
	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6
	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6

12. How often do you use each of the languages for the following activities? Please circle the number in the table according to the scale below.
Never *Rarely* *Occasionally* *Sometimes* *Frequently* *Very Frequently* *Always*
1 ______ 2_______3___________4__________5__________6___________7______

Language	Thinking	Talking to yourself	Arithmetic **(a)**	Remembering numbers **(b)**
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7

Language	Expressing anger	Showing affection	Swearing	Dreaming
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7

(b). Counting, calculating tips, etc. **(c).** Phone numbers, ID numbers, etc.

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

	Listening	Speaking	Reading	Writing
At home (with family)				
At home (with flatmates)				
With friends				
With a partner				
At school				
At work				

14. Please indicate how often you CURRENTLY use each language in the following contexts

Never Rarely Occasionally Sometimes Frequently Very Frequently Always
1 ______ 2_______3___________4__________5__________6____________7______

Language	With family	With friends	With a partner	At work	At university	With flatmates/ classmates
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7

15. Which of your languages you became fluent with first:__________________
16. Which do you consider to be your dominant language:__________________

17. How would you rate how strong swear and taboo words in each of your languages are?

Language	Not strong	Little	Fairly	Strong	Very strong

18. Do you have a preference for emotion terms and terms of endearment in one language over all other? Which language is it? _______________________________

Questionnaire coding: EXPERIMENT 4

Questions 4 and 7 (coded in column G) are coded together as cumulative stay in English-speaking countries. Code in months. For participants from countries such as Singapore, Hong Kong or Nigeria, where English is an official language, stay in their home country should be included in their cumulative stay.

Question 8 (Age of Acquisition) is coded as two columns: H and I. Column H is age of exposure, so the age first learnt the language and column I is that subtracted from their age to get the number of years spent learning (ignore what they say in the number of years spent learning, as some people mistake this).

Question 9 (School instruction). Code in column J as a sum of years in English education.

Question 10 (English learning in different situations) is coded in columns K, L, M, N, O, P, Q. It is broken down by different situations of learning (hence there are 7 different sub-categories). Code participants' response to each sub-category. Number of participants will have left the 'after immigrating to a country where spoken' question blank. Replace this with an age that is computed by subtracting participant's answer to question 4 ('How long have you been in the country of your current residence') from their age. Otherwise if the participant has not said answered a sub-category, enter NA. If they have specified the situation 'Other', enter the age and situation in the box

Questions 11 (frequency of language use) is coded in columns R, S, T, U, V, W. Code only participant's response for English as a single number that they have circled, separately for each sub-category.

Question 12 (language preference in different situations) is coded in columns X, Y, Z, AA, AB, AC, AD, AE. The question is coded as a difference score between the languages participants listed. The questions are broken down into sub-categories. The response scale ranges from 1 (never) to 7 (always). Positive integers indicate a 'favour' for the other language, negative integers indicate a 'favour' for English (for example if a participant has indicated 7 for their first language for a sub-category, and 2 for English, this would be coded as 5. If a participant had indicated 4 for English and 3 for their first language, this would be coded as -1). If the participant listed multiple languages, the score for each sub-category is based on the difference score between English and the language which had a higher score for that specific sub-category.

Question 13 (language dominance in different situations) is coded in column AF and is coded as the overall percentage of English across the situations. (number of cells with English divided by total number of cells [24]). If the participant has written "both", this is counted as English. If the participant has not responded to some of the situations, leave them out from your percentage calculation (i.e you will have a smaller total number of cells)

Question 14 (current language use in different contexts) is coded in columns AG, AH, AI, AJ, AK, AL. Code as difference scores (check question 12)

Question 15 (which language you became fluent first) is coded in column AM. The languages as coded as 0 and 1. Enter 0 if the participant became fluent in another language first, 1 if the participant indicated becoming fluent in English first.

Question 16 (which do you consider to be your dominant language) is coded in column AN. Coded as 0 if participant indicated another language, 1 if they indicated English and 2 if they say both.
Question 17 (taboo words) is coded in column AO and coded as a difference score (check question 12)
Question 18 (preference for emotion terms) is coded in column AP. Code as 0 and 1: 0 for another language and 1 for English

Language Background Questionnaire: EXPERIMENTS 5 AND 6

- Are you a native English speaker who does not speak another language on a daily basis? If YES, you can skip the other questions on this page (YES/NO)
- At what age did you start learning English?
- Are either of you parents native speakers of English?
- How long have you lived in the UK? (cumulative, please give your answer in years and months)
- Have you lived in another English-speaking country? (if yes, give your answer in years and months, if not, type 0)
- How long have you studied in the UK? (give your answer in years and months, if you haven't studied in the UK, type 0)
- How proficient are you in READING in English? (1 - 7)
- How proficient are you in WRITING in English? (1 - 7)
- How proficient are you in SPEAKING in English? (1 -7)
- Do you want to leave any comments about your language background that you think might be relevant?

Appendix B: Stimuli lists

These lists only include the selected stimuli words and their word type. Full lists with covariate values can be found on the OSF: https://osf.io/9rqbj/

Experiment 1, 2 and 4

HA= High arousing
LA= Low arousing
DI= Distractor

				pastime	
Word	Type	busy		puffing	DI
porn	HA	threaten	HA	procreation	HA
ford	LA	pronounce	LA	commemorate	LA
bolt	DI	clearance	DI	requisition	DI
leopard	HA	ejaculate	HA	competitor	HA
dryer	LA	insignia	LA	ingredient	LA
fireman	DI	dormitory	DI	manifesto	DI
score	HA	homicide	HA	arrest	HA
bunch	LA	elderly	LA	bathroom	LA
crap	DI	deposit	DI	muscle	DI
terrify	HA	excite	HA	ravishing	HA
pacify	LA	sadness	LA	resolute	LA
remarry	DI	unwise	DI	glorified	DI
tornado	HA	dominate	HA	thunderstorm	HA
broccoli	LA	clarify	LA	buttermilk	LA
soprano	DI	incentive	DI	mockingbird	DI
detonate	HA	thief	HA	celebrate	HA
parental	LA	cloud	LA	comfort	LA
maverick	DI	juice	DI	confident	DI
masturbation	HA	arouse	HA	horrid	HA
ventilation	LA	groggy	LA	comply	LA
vaccination	DI	bouncy	DI	rampant	DI
battle	HA	gunpoint	HA	jazz	HA
winter	LA	whitewash	LA	item	LA
pocket	DI	backhand	DI	pizza	DI
lottery	HA	dramatic	HA	siren	HA
housing	LA	domestic	LA	granny	LA
observer	DI	talented	DI	greasy	DI
roar	HA	nightlife	HA	avenge	HA
damp	LA	bookworm	LA	drowsy	LA
yank	DI	heirloom	DI	rascal	DI
evolution	HA	disgusting	HA	parenthood	HA
category	LA	maintenance	LA	formulate	LA
violation	DI	surrender	DI	misconduct	DI
happy	HA	destruct	HA	action	HA
quiet	LA			extra	LA
				suspect	DI
				caffeine	HA
				northwest	LA
	DI		LA	cuisine	DI

seducer	HA	independent	HA	pleasure	HA
nunnery	LA	historical	LA	patient	LA
palpable	DI	opposition	DI	survive	DI
genius	HA	gunfighter	HA	warning	HA
asleep	LA	tablecloth	LA	willing	LA
habit	DI	stethoscope	DI	stealing	DI
abrasive	HA	vaccine	HA	turbulence	HA
syllable	LA	pavement	LA	godfather	LA
breakaway	DI	receipt	DI	gymnastics	DI
agonizing	HA	attack	HA	famine	HA
uneventful	LA	accept	LA	vacant	LA
expedient	DI	buddy	DI	cipher	DI
molest	HA	outrage	HA	rush	HA
repose	LA	boredom	LA	slow	LA
shocker	DI	extinct	DI	load	DI
horrifying	HA	thrill	HA	kiss	HA
disposable	LA	pause	LA	tree	LA
prehistoric	DI	blunt	DI	paint	DI
electrocute	HA	gunfire	HA	assassination	HA
pillowcase	LA	janitor	LA	transportation	LA
antiaircraft	DI	firewood	DI	examination	DI
madman	HA	inconvenience	HA	bang	HA
pension	LA	predictable	LA	gray	LA
boogie	DI	incompetent	DI	halt	DI
handgun	HA	eruption	HA	intelligence	HA
pamphlet	LA	chromium	LA	comfortable	LA
jackal	DI	aerospace	DI	confidential	DI
gruesome	HA	perky	HA	gold	HA
ordnance	LA	misty	LA	foot	LA
fallout	DI	ailing	DI	face	DI
crazed	HA	execution	HA	hero	HA
frail	LA	sanctuary	LA	grow	LA
strive	DI	radiation	DI	grab	DI
squealer	HA	passion	HA	crocodile	HA
menthol	LA	counsel	LA	wallpaper	LA
moonbeam	DI	foolish	DI	flamingo	DI
treacherous	HA	spicy	HA	breakup	HA
undisturbed	LA	acre	LA	whatnot	LA
foolishness	DI	pushy	DI	hearsay	DI
schizophrenia	HA	crusade	HA		
equilibrium	LA	gospel	LA		
neurological	DI	ruling	DI		
panic	HA	liar	HA		
elder	LA	empty	LA		
moron	DI	virus	DI		

Experiment 3

G = German
E = English

WORD	Wordtype	Language
lebendig	Positive	G
verraten	Negative	G
endgültig	Neutral	G
schwimmen	Positive	G
schlinge	Negative	G
ohnmacht	Neutral	G
herz	Positive	G
grab	Negative	G
glas	Neutral	G
gefühl	Positive	G
krise	Negative	G
inhalt	Neutral	G
tummeln	Positive	G
gestank	Negative	G
prahlen	Neutral	G
königin	Positive	G
begraben	Negative	G
geländer	Neutral	G
geld	Positive	G
kampf	Negative	G
kreis	Neutral	G
wunder	Positive	G
ärger	Negative	G
fraglich	Neutral	G
achtsam	Positive	G
nerven	Negative	G
trödeln	Neutral	G
geil	Positive	G
fluch	Negative	G
streu	Neutral	G
diamant	Positive	G
entführer	Negative	G
plantage	Neutral	G
lekker	Positive	G
tunte	Negative	G
werfer	Neutral	G
festlich	Positive	G
ängstlich	Negative	G
bursche	Neutral	G
kichern	Positive	G
spotten	Negative	G
schurke	Neutral	G
unabhängig	Positive	G
kommunismus	Negative	G
übertragung	Neutral	G
saphir	Positive	G
henker	Negative	G
tupfer	Neutral	G
urkomisch	Positive	G
tollwütig	Negative	G
bescheuert	Neutral	G
kuss	Positive	G
flut	Negative	G
zink	Neutral	G
gelingen	Positive	G
verletzen	Negative	G
vermuten	Neutral	G
abenteuerlich	Positive	G
eifersüchtig	Negative	G
obligatorisch	Neutral	G
überglücklich	Positive	G
erniedrigung	Negative	G
demographisch	Neutral	G
hochzeit	Positive	G
einbruch	Negative	G
statisch	Neutral	G
euphorie	Positive	G
psychose	Negative	G
syndikat	Neutral	G
amüsieren	Positive	G
gewalttätig	Negative	G
artillerie	Neutral	G
brauen	Positive	G
zanken	Negative	G
leblos	Neutral	G
wach	Positive	G
gift	Negative	G
grob	Neutral	G
retter	Positive	G
giftig	Negative	G
konsul	Neutral	G
zucker	Positive	G
kotzen	Negative	G
hocker	Neutral	G
feiheit	Positive	G

angriff	Negative	G
abwarten	Neutral	G
magisch	Positive	G
kummer	Negative	G
horchen	Neutral	G
gefällig	Positive	G
entführen	Negative	G
gehorsam	Neutral	G
freudig	Positive	G
betrug	Negative	G
jammer	Neutral	G
anreiz	Positive	G
verrat	Negative	G
flaute	Neutral	G
kosmisch	Positive	G
tödlich	Negative	G
kindisch	Neutral	G
jubel	Positive	G
fieber	Negative	G
winkel	Neutral	G
leidenschaft	Positive	G
rücksichtslos	Negative	G
betrieblich	Neutral	G
erfinden	Positive	G
zerstören	Negative	G
erachten	Neutral	G
witz	Positive	G
hass	Negative	G
hohl	Neutral	G
feier	Positive	G
lepra	Negative	G
härten	Neutral	G
jugend	Positive	G
teufel	Negative	G
monat	Neutral	G
buttocks	Positive	E
headless	Negative	E
salesman	Neutral	E
feast	Positive	E
drown	Negative	E
wedge	Neutral	E
discover	Positive	E
horrible	Negative	E
amateur	Neutral	E
advancement	Positive	E
frustration	Negative	E
constitute	Neutral	E
espresso	Positive	E
dictator	Negative	E
lecturer	Neutral	E
kisser	Positive	E
kidnap	Negative	E
mumble	Neutral	E
climax	Positive	E
harass	Negative	E
absent	Neutral	E
gourmet	Positive	E
culprit	Negative	E
migrate	Neutral	E
speedboat	Positive	E
cockroach	Negative	E
boardroom	Neutral	E
prosper	Positive	E
fearful	Negative	E
polling	Neutral	E
adore	Positive	E
creepy	Negative	E
privy	Neutral	E
bargain	Positive	E
hostile	Negative	E
dismiss	Neutral	E
lasagna	Positive	E
gunfire	Negative	E
inhaler	Neutral	E
quickie	Positive	E
enslave	Negative	E
retrace	Neutral	E
incredible	Positive	E
emergency	Negative	E
federation	Neutral	E
congratulation	Positive	E
claustrophobia	Negative	E
impertinence	Neutral	E
perk	Positive	E
germs	Negative	E
hock	Neutral	E
raise	Positive	E
steal	Negative	E
ward	Neutral	E
romance	Positive	E
madness	Negative	E
mortal	Neutral	E
amazed	Positive	E
insane	Negative	E

comply	Neutral	E
elope	Positive	E
fatso	Negative	E
backer	Neutral	E
ravishing	Positive	E
penniless	Negative	E
desertion	Neutral	E
enthusiast	Positive	E
terrifying	Negative	E
disposable	Neutral	E
flirt	Positive	E
snarl	Negative	E
deuce	Neutral	E
jackpot	Positive	E
nigger	Negative	E
weekday	Neutral	E
fertile	Positive	E
wartime	Negative	E
descend	Neutral	E
adventure	Positive	E
depression	Negative	E
dependent	Neutral	E
lucky	Positive	E
nasty	Negative	E
mental	Neutral	E
exhilarating	Positive	E
excruciating	Negative	E
diversionary	Neutral	E
excite	Positive	E
morbid	Negative	E
latent	Neutral	E
pretzel	Positive	E
mugger	Negative	E
sternum	Neutral	E
caffeine	Positive	E
stalker	Negative	E
outpost	Neutral	E
pleasure	Positive	E
trouble	Negative	E
needless	Neutral	E
radiant	Positive	E
horrific	Negative	E
unlisted	Neutral	E
vibrant	Positive	E
breakup	Negative	E
abstain	Neutral	E
mastery	Positive	E
nemesis	Negative	E
interim	Neutral	E
frisky	Positive	E
quitter	Negative	E
homely	Neutral	E
cheer	Positive	E
slut	Negative	E
caste	Neutral	E
euphoric	Positive	E
deceitful	Negative	E
pacifist	Neutral	E
cash	Positive	E
shot	Negative	E
flat	Neutral	E

Appendix C: Countries of origin by experiment

Bilingual participants' countries of origin by experiment (number of participants in brackets if more than 1)

Experiment 1

Bangladesh, Brazil, Brunei, Bulgaria (9), China (2), Cyprus (3), Czech Republic, Denmark, Finland (3), France (4), Germany (6), Ghana, Greece (2), Hong Kong, India (4), Israel, Italy (7), Kazakhstan, Latvia, Lithuania (2), Malaysia (3), Mauritius, Mexico, Netherlands, Nigeria, Poland (3), Portugal, Romania (3), Russia (3), Scotland (3), Singapore, Slovakia (4), Spain (4), Sweden, UK (7), USA

Experiment 2

China (18), Colombia (2), Finland, Guatemala, Iceland, India (7), Indonesia, Italy, Lebanon, Saudi-Arabia (2), South Korea (4), Thailand, USA (16)

Experiment 4

Australia, Belarus, Britain, Bulgaria (4), Burma, China (6), Cyprus, Czech republic (2), Estonia, Finland, French, Germany (2), Greece, Hong Kong, Hungary, India (3), Italy (2), Jordan, Latvia, Lithuania (3), Malaysia (2), Netherlands, Pakistan (2), Philippines, Poland (3), Portugal (2), Romania (2), Singapore (2), Spain (2), Sri Lanka, Turkey, UK (4), USA

Experiment 5

Belarus, Belgium (2), Bulgaria (5), Canada, Cyprus, Czech Republic, Estonia, Finland, France (5), Germany (8), Greece, Hong Kong (3), Hungary (14), Ireland, Italy (3), Lithuania, Macedonia, Philippines (2), Poland (2), Slovenia, South Korea, Sweden (3), UK

Appendix D: Mixed Effects Models

Experiment 1

Button press

```
recogmodel <- glmer(button ~
          # Fixed main effects:
                condition +
                Group +
                LEXFREQcentered +
                BGFREQcentered +
                CONCRETEcentered +
                LEN_orthoNcentered +
                LexTALEcentered+
          # Fixed 2-way interactions involving Group:
                Group:condition +
                Group:LEXFREQcentered +
                Group:BGFREQcentered +
                Group:CONCRETEcentered +
                Group:LEN_orthoNcentered +
                Group:LexTALEcentered+
          # By-subject random effects:
                (1 +
                  condition +
                  LEXFREQcentered +
                  BGFREQcentered +
                  CONCRETEcentered +
                  LEN_orthoNcentered
                 |RECORDING_SESSION_LABEL) +
          # By-item random effects:
                (1 + Group
                 |item),
               data=modeldata_recog,
               family=binomial(link="logit"),
               control = glmerControl(optimizer = c("bobyqa"),
                      optCtrl=list(maxfun=50000),tol = .0001))
```

Area under the curve monolingual vs. bilingual

```
model_area_all <- glmer(area_under_curve ~
          # Fixed main effects:
             condition +
             Group +
             LEXFREQcentered +
             BGFREQcentered +
             CONCRETEcentered +
             LEN_orthoNcentered +
             LexTALEcentered+
         # Fixed 2-way interactions involving Group:
             Group:condition +
             Group:LEXFREQcentered +
             Group:BGFREQcentered +
             Group:CONCRETEcentered +
             Group:LEN_orthoNcentered +
             Group:LexTALEcentered+
        # By-subject random effects:
         (1 +
          condition +
          LEXFREQcentered +
          BGFREQcentered +
          CONCRETEcentered +
          LEN_orthoNcentered
          | RECORDING_SESSION_LABEL) +
      # By-item random effects:
         (1 + Group| item),
         data=modeldata,
         family=Gamma(link="identity"),
         control = glmerControl(optimizer = c("bobyqa"),
         optCtrl=list(maxfun=50000),tol = .0001))
```

Area under the curve PCA predictors

```
#Specified in chapter 2
```

Experiment 2

Button press

```
recog_expt2 <- glmer(Response_factor ~
        # Fixed main effects:
           Wtype_coded +
           Group_coded +
           Lexfreq_centered +
           BGfreq_centered +
           Concrete_centered +
           Length_centered +
           Lextale_centered +
        # Fixed 2-way interactions involving Group:
           Group_coded:Wtype_coded +
           Group_coded:Lexfreq_centered +
           Group_coded:BGfreq_centered +
           Group_coded:Concrete_centered +
           Group_coded:Length_centered +
           Group_coded:Lextale_centered
        # By-subject random effects:
           (1 +
             Wtype_coded +
             Lexfreq_centered +
             BGfreq_centered +
             Concrete_centered +
             Length_centered +
            |Participant) +
        # By-item random effects:
           (1 + Group_coded + Lextale_centered|Stimuli),
          data=modeldata,
          family=binomial(link="logit"),
          control = glmerControl(optimizer = "optimx",
          calc.derivs = FALSE,
          optCtrl = list(method = "L-BFGS-B",
          maxit = 10000, starttests = FALSE, kkt = FALSE)))
```

Area under the curve

```
auc_expt2 <- lmer(area_under_curve ~
         # Fixed main effects:
             Wtype_coded +
             Group_coded +
             Lexfreq_centered +
             BGfreq_centered +
             Concrete_centered +
             Length_centered +
             Lextale_centered +
         # Fixed 2-way interactions involving Group:
             Group_coded:Wtype_coded+
             Group_coded:Lexfreq_centered +
             Group_coded:BGfreq_centered +
             Group_coded:Concrete_centered +
             Group_coded:Length_centered +
             Group_coded:Lextale_centered +
         # By-subject random effects:
             (1 +
               Wtype_coded +
               Lexfreq_centered +
               BGfreq_centered +
               Concrete_centered +
               Length_centered|Participant) +
         # By-item random effects:
         (1 + Group_coded|Stimuli),
         data=modeldata,
         REML=FALSE,
         control = lmerControl(optimizer = "optimx",
         calc.derivs = FALSE,
         optCtrl = list(method = "L-BFGS-B", maxit = 10000,
         starttests = FALSE, kkt = FALSE)))
```

Experiment 3

Lexical Decision Task

Accuracy

```
#By group
by_group_accuracy <- glmer(correct ~
         # Fixed main effects:
            group_centered +
            condition1 +
            condition2 +
            LexTALE_centered+
            PC1_Length_centered+
            PC4_Freq_centered+
            PC5_Concr_centered+
         # Fixed 2-way interactions involving Group:
            group_centered: condition1 +
            group_centered: condition2 +
            group_centered: PC1_Length_centered +
            group_centered: PC4_Freq_centered +
            group_centered: PC5_Concr_centered +
         # By-subject random effects:
            (1 + condition1 + condition2|Subject) +
        # By-item random effects:
           (1 + group_centered+LexTALE_centered|Stimuli),
          data=modeldata_accuracy,
          family=binomial(link="logit"),
          control = glmerControl(optimizer = c("bobyqa"),
               optCtrl=list(maxfun=50000),tol = .0001))

#By language
by_language_accuracy <- glmer(correct ~
        # Fixed main effects:
              Language_centered+
              condition1 +
              condition2 +
              LexTALE_centered+
              PC1_Length_centered+
              PC4_Freq_centered+
              PC5_Concr_centered+
        # Fixed 2-way interactions involving Language:
               Language_centered: condition1 +
               Language_centered: condition2 +
               Language_centered: PC1_Length_centered +
               Language_centered: PC4_Freq_centered +
               Language_centered: PC5_Concr_centered +
        # By-subject random effects:
              (1 + condition1*Language_centered +
               condition2*Language_centered|Subject) +
        # By-item random effects:
              (1 + LexTALE_centered|Stimuli),
             data=modeldata_accuracy_bothlang,
             family=binomial(link="logit"),
             control = glmerControl(optimizer = c("bobyqa"),
             optCtrl=list(maxfun=50000),tol = .0001))

#Model comparisons done based on these models using the update function (null models for group compa
rison: no group, no condition and no group:condition interaction, null models for language comparison: n
o language, no condition, no language:condition interaction). Both condition 1 and 2 removed for model c
omparisons.
```

Reaction time

By group

```
#Selecting this model was particularly difficult and we had to run a number of different models to identify
the best fit. Best fit was determined based on convergence (model converged), smallest AUC and smallest s
td error. Below all the steps we took to find the best model.

#First by-group model
by_group <- glmer(StimuliDisplay.RT ~
       # Fixed main effects:
           group_centered +
           LexTALE+
           condition1 +
           condition2 +
           PC1_Length_centered+
           PC4_Freq_centered+
           PC5_Concr_centered+
       # Fixed 2-way interactions involving Group:
           group_centered:condition1 +
           group_centered:condition2+
           group_centered:LexTALE+
           group_centered:PC1_Length_centered+
             group_centered:PC4_Freq_centered+
             group_centered:PC5_Concr_centered+
        # By-subject random effects:
             (1 +condition1+condition2
              ||Subject) +
        # By-item random effects:
             (1+group_centered||Stimuli),
            data=modeldata,
            family=Gamma(link="identity"),
            control =glmerControl(optimizer =c("bobyqa"),
            optCtrl=list(maxfun=50000),tol = .0001))

#Fit the model without LexTALE
by_group2 <- glmer(StimuliDisplay.RT ~
        # Fixed main effects:
            group_centered +
            condition1 +
            condition2 +
            PC1_Length_centered+
            PC4_Freq_centered+
            PC5_Concr_centered+
        # Fixed 2-way interactions involving Group:
            group_centered:condition1 +
            group_centered:condition2+
            group_centered:PC1_Length_centered+
            group_centered:PC4_Freq_centered+
            group_centered:PC5_Concr_centered+
        # By-subject random effects:
            (1 +condition1+condition2
             ||Subject) +
       # By-item random effects:
            (1+group_centered||Stimuli),
           data=modeldata,
           family=Gamma(link="identity"),
           control =glmerControl(optimizer =c("bobyqa"),
                optCtrl=list(maxfun=50000),tol = .0001))

#Fit the model without LexTALE (just test out with glm to get an approximation of how the model estimat
es should look like)
by_group2glm <- glm(StimuliDisplay.RT ~
         # Fixed main effects:
            group_centered +
            condition1 +
            condition2 +
            PC1_Length_centered+
            PC4_Freq_centered+
            PC5_Concr_centered+
         # Fixed 2-way interactions involving Group:
            group_centered:condition1 +
            group_centered:condition2+
            group_centered:PC1_Length_centered+
            group_centered:PC4_Freq_centered+
            group_centered:PC5_Concr_centered,
            data=modeldata,
            family=Gamma(link="identity"))

# Step 1: Calibration models (stepwise approach to finding the best fit). Simple model with no covariates
by_group3x <- glmer(StimuliDisplay.RT ~
         # Fixed main effects:
            group_centered +
            condition1 +
            condition2 +
          # Fixed 2-way interactions involving Group:
             group_centered:condition1 +
             group_centered:condition2 +
          # By-subject random effects:
             (1 + condition1 + condition2 |Subject) +
          # By-item random effects:
             (1 + group_centered|Stimuli),
            data=modeldata,
            family=Gamma(link="identity"),
            control =glmerControl(optimizer =c("bobyqa"),
            optCtrl=list(maxfun=50000),tol = .0001))



# Step 2: Include LexTALE scores
by_group4x <- glmer(StimuliDisplay.RT ~
          # Fixed main effects:
              group_centered +
              condition1 +
              condition2 +
              LexTALE+
          # Fixed 2-way interactions involving Group:
              group_centered:condition1 +
              group_centered:condition2 +
              group_centered:LexTALE +
          # By-subject random effects:
              (1 + condition1 + condition2 |Subject) +
          # By-item random effects:
              (1 + group_centered+ LexTALE |Stimuli),
             data=modeldata,
             family=Gamma(link="identity"),
             control =glmerControl(optimizer =c("bobyqa"),
             optCtrl=list(maxfun=50000),tol = .0001))

#Step 3: Add covariates
by_group5x <- glmer(StimuliDisplay.RT ~
          # Fixed main effects:
              group_centered +
              condition1 +
              condition2 +
              LexTALE+
              PC1_Length_centered+
              PC4_Freq_centered+
              PC5_Concr_centered+
          # Fixed 2-way interactions involving Group:
              group_centered:condition1 +
              group_centered:condition2 +
              group_centered:LexTALE +
          # By-subject random effects:
              (1 + condition1 + condition2 +  PC1_Length_centered +
                PC4_Freq_centered+
                PC5_Concr_centered |Subject) +
          # By-item random effects:
              (1 + group_centered+ LexTALE |Stimuli),
             data=modeldata,
             family=Gamma(link="identity"),
             control =glmerControl(optimizer =c("bobyqa"),
             optCtrl=list(maxfun=50000),tol = .0001))
```

```
#Step 4: no covariates in the random structure
by_group6x <- glmer(StimuliDisplay.RT ~
        # Fixed main effects:
           group_centered +
           condition1 +
           condition2 +
           LexTALE+
           PC1_Length_centered+
           PC4_Freq_centered+
           PC5_Concr_centered+
        # Fixed 2-way interactions involving Group:
           group_centered:condition1 +
           group_centered:condition2 +
           group_centered:LexTALE +
        # By-subject random effects:
           (1 + condition1 + condition2|Subject) +
        # By-item random effects:
           (1 + group_centered+ LexTALE |Stimuli),
          data=modeldata,
          family=Gamma(link="identity"),
          control = glmerControl(optimizer = c("bobyqa"),
          optCtrl=list(maxfun=50000),tol = .0001))

#step 4: Covariate-group interactions included but covariates excluded from random structure
by_group7x <- glmer(StimuliDisplay.RT ~
           # Fixed main effects:
           group_centered +
           condition1 +
           condition2 +
           LexTALE+
           PC1_Length_centered+
           PC4_Freq_centered+
           PC5_Concr_centered+
           # Fixed 2-way interactions involving Group:
           group_centered: condition1 +
           group_centered: condition2 +
           group_centered: PC1_Length_centered +
           group_centered: PC4_Freq_centered +
           group_centered: PC5_Concr_centered +
           group_centered: LexTALE+
           # By-subject random effects:
           (1 + condition1 + condition2|Subject) +
           # By-item random effects:
           (1 + group_centered+ LExTALE|Stimuli),
          data=modeldata,
          family=Gamma(link="identity"),
          control = glmerControl(optimizer = c("bobyqa"),
          optCtrl=list(maxfun=50000),tol = .0001))

#Step 5: Center lexTALE, exclude group:LexTALE
 by_group8x <- glmer(StimuliDisplay.RT ~
         # Fixed main effects:
            group_centered +
            condition1 +
            condition2 +
            LexTALE_centered+
            PC1_Length_centered+
            PC4_Freq_centered+
            PC5_Concr_centered+
         # Fixed 2-way interactions involving Group:
            group_centered: condition1 +
            group_centered: condition2 +
            group_centered: PC1_Length_centered +
            group_centered: PC4_Freq_centered +
            group_centered: PC5_Concr_centered +
         # By-subject random effects:
            (1 + condition1 + condition2|Subject) +
         # By-item random effects:
            (1 + group_centered|Stimuli),
           data=modeldata,
           family=Gamma(link="identity"),
           control = glmerControl(optimizer = c("bobyqa"),
           optCtrl=list(maxfun=50000),tol = .0001))

#step 6: Include LexTALE in the random effect structure. THIS IS THE SELECTED MODEL
 by_group9x <- glmer(StimuliDisplay.RT ~
        # Fixed main effects:
            group_centered +
            condition1 +
            condition2 +
            LexTALE_centered+
            PC1_Length_centered+
            PC4_Freq_centered+
            PC5_Concr_centered+
        # Fixed 2-way interactions involving Group:
            group_centered: condition1 +
            group_centered: condition2 +
            group_centered: PC1_Length_centered +
            group_centered: PC4_Freq_centered +
            group_centered: PC5_Concr_centered +
            group_centered: LexTALE_centered +
         # By-subject random effects:
            (1 + condition1 + condition2|Subject) +
         # By-item random effects:
            (1 + group_centered+LexTALE_centered+
            group_centered:LexTALE_centered|Stimuli),
           data=modeldata,
           family=Gamma(link="identity"),
           control = glmerControl(optimizer = "optimx",
           calc.derivs = FALSE,
           optCtrl = list(method = "L-BFGS-B",
           maxit = 10000,
           starttests = FALSE, kkt = FALSE)))

#Model comparisons done with by_group9x (null models: group omitted, condition omitted, group:condition omitted)
```

By language

```
by_language1x <- glmer(StimuliDisplay.RT ~
        # Fixed main effects:
           language_centered +
           condition1 +
           condition2 +
           LexTALE_centered+
           PC1_Length_centered+
           PC4_Freq_centered+
           PC5_Concr_centered+
        # Fixed 2-way interactions involving language:
           language_centered: condition1 +
            language_centered: condition2 +
            language_centered: PC1_Length_centered +
            language_centered: PC4_Freq_centered +
            language_centered: PC5_Concr_centered +
         # By-subject random effects:
            (1 + condition1*language_centered +
            condition2*language_centered|Subject) +
         # By-item random effects:
            (1 + LexTALE_centered|Stimuli),
           data=modeldata_bylanguage,
           family=Gamma(link="identity"),
           control = glmerControl(optimizer = "optimx",
           calc.derivs = FALSE,
           optCtrl = list(method = "L-BFGS-B",
           maxit = 10000,
           starttests = FALSE, kkt = FALSE)))

#Model comparisons done with this model (null models: language omitted, condition omitted, language:c
ondition omitted)
```

Pupil models

Button press

```
#By group
by_group_accuracy <- glmer(button ~
           # Fixed main effects:
                group_centered +
                condition1 +
                condition2 +
                LexTALE_centered+
                PC1_Length_centered+
                PC4_Freq_centered+
                PC5_Concr_centered+
           # Fixed 2-way interactions involving Group:
                group_centered: condition1 +
                group_centered: condition2 +
                group_centered: PC1_Length_centered +
                group_centered: PC4_Freq_centered +
                group_centered: PC5_Concr_centered +
                group_centered: LexTALE_centered +
           # By-subject random effects:
                (1 + condition1 +
                 condition2|Participant) +
           # By-item random effects:
             (1 + group_centered+LexTALE_centered+
             group_centered:LexTALE|Trial_ID),
             data=modeldata_group,
             family=binomial(link="logit"),
             control=glmerControl(optimizer ="optimx",
             calc.derivs = FALSE,
             optCtrl = list(method = "L-BFGS-B",
             maxit = 10000,
             starttests = FALSE, kkt = FALSE)))

#Model comparisons done with this model (null models: groupomitted, condition omitted, group:conditio
n omitted)

#By language
by_language_accuracy <- glmer(button ~
                  language_centered +
                condition1 +
                condition2 +
                LexTALE_centered +
                PC1_Length_centered +
                PC4_Freq_centered +
                PC5_Concr_centered +
                language_centered:condition1 +
                language_centered:condition2 +
                language_centered:PC1_Length_centered +
                language_centered:PC4_Freq_centered +
                language_centered:PC5_Concr_centered +
                (1 + condition1 * language_centered +
                condition2 * language_centered | Participant) +
                (1 + LexTALE | Trial_ID),
                data= modeldata_lang,
                family=binomial(link="logit")
                control= glmerControl(optimizer = "optimx",
                calc.derivs = FALSE,
                optCtrl = list(method = "L-BFGS-B",
                maxit = 10000, starttests = FALSE, kkt = FALSE))


#Model comparisons done with this model (null models: language omitted, condition omitted, language:condition omitted)
```

Pupillary response

```
#By group
pupil_by_group4x <- glmer(area_under_curve ~
          # Fixed main effects:
                group_centered +
                condition1 +
                condition2 +
                LexTALE_centered+
                PC1_Length_centered+
                PC4_Freq_centered+
                PC5_Concr_centered+
          # Fixed 2-way interactions involving language:
                group_centered: condition1 +
                group_centered: condition2 +
                group_centered: PC1_Length_centered +
                group_centered: PC4_Freq_centered +
                group_centered: PC5_Concr_centered +
                group_centered:LexTALE_centered +
          # By-subject random effects:
                (1 + condition1 + condition2|Participant) +
          # By-item random effects:
               (1 + LexTALE_centered +group_centered +
               LexTALE_centered:group_centered |Trial_ID),
               data=modeldata_group,
               family=Gamma(link="identity"),
               control = glmerControl(optimizer = "optimx",
               calc.derivs = FALSE,
               optCtrl = list(method = "L-BFGS-B",
               maxit = 10000,
               starttests = FALSE, kkt = FALSE)))

#Model comparisons done with this model (null models: group omitted, condition omitted, group:condition omitted)

#By language
```

```
pupil_by_language2x <- glmer(area_under_curve ~
                  # Fixed main effects:
                      language_centered +
                      condition1 +
                      condition2 +
                      LexTALE_centered+
                      PC1_Length_centered+
                      PC4_Freq_centered+
                      PC5_Concr_centered+
                 # Fixed 2-way interactions involving language:
                      language_centered: condition1 +
                      language_centered: condition2 +
                      language_centered: PC1_Length_centered +
                      language_centered: PC4_Freq_centered +
                      language_centered: PC5_Concr_centered +
                 # By-subject random effects:
                   (1 + condition1*language_centered +
                   condition2*language_centered|Participant) +
                 # By-item random effects:
                      (1 + LexTALE_centered|Trial_ID),
                     data=modeldata_language,
                     family=Gamma(link="identity"),
                     control = glmerControl(optimizer = "optimx",
                     calc.derivs = FALSE,
                     optCtrl = list(method = "L-BFGS-B",
                     maxit = 10000,
                     starttests = FALSE, kkt =FALSE)))

#Model comparisons done with this model (null models: language omitted, condition omitted, language:c
ondition omitted)
```

Experiment 4

Rating models

```
#Models were selected in multiple steps of trying to establish the best fit.
#First, the different thresholds were tested (flexible thresholds were selected)

#Full model flexible thresholds
model_ratings_FULL <- clmm(rating ~
             # Fixed main effects:
                   cond1 +
                   cond2 +
                   Group_coded +
                   LEXFREQcentered +
                   BGFREQcentered +
                   CONCRETEcentered +
                   LEN_orthoNcentered +
                   LexTALEcentered +
                   VALDOMcentered +
             # Fixed 2-way interactions involving Group:
                   Group_coded:cond1 +
                   Group_coded:cond2 +
                   Group_coded:LEXFREQcentered +
                   Group_coded:BGFREQcentered +
                   Group_coded:CONCRETEcentered +
                   Group_coded:LEN_orthoNcentered +
                   Group_coded:LexTALEcentered +
                   Group_coded: VALDOMcentered +
             # By-subject random effects:
                   (1 + cond1 +cond2|Participant) +
        # By-item random effects:
            (1 + Group_coded +
             LexTALEcentered |item),
           data=modeldata_clmm,
           link="cloglog",
           control = clmm.control("nlminb"))

#Full model, symmetric thresholds
model_ratings_FULL_B <- clmm(rating ~
       # Fixed main effects:
            cond1 +
            cond2 +
            Group_coded +
            LEXFREQcentered +
            BGFREQcentered +
            CONCRETEcentered +
            LEN_orthoNcentered +
            LexTALEcentered +
            VALDOMcentered +
       # Fixed 2-way interactions involving Group:
            Group_coded:cond1 +
            Group_coded:cond2 +
            Group_coded:LEXFREQcentered +
            Group_coded:BGFREQcentered +
            Group_coded:CONCRETEcentered +
            Group_coded:LEN_orthoNcentered +
            Group_coded:LexTALEcentered +
            Group_coded: VALDOMcentered +
       # By-subject random effects:
            (1 + cond1 +cond2|Participant) +
       # By-item random effects:
           (1 + Group_coded +LexTALEcentered |item),
           data=modeldata_clmm,
           link="cloglog",
           control = clmm.control("nlminb"),
           threshold = "symmetric")

#Full model, equidistant thresholds
model_ratings_FULL_C <- clmm(rating ~
         # Fixed main effects:
            cond1 +
            cond2 +
            Group_coded +
            LEXFREQcentered +
            BGFREQcentered +
            CONCRETEcentered +
            LEN_orthoNcentered +
            LexTALEcentered +
            VALDOMcentered +
         # Fixed 2-way interactions involving Group:
            Group_coded:cond1 +
            Group_coded:cond2 +
            Group_coded:LEXFREQcentered +
            Group_coded:BGFREQcentered +
            Group_coded:CONCRETEcentered +
            Group_coded:LEN_orthoNcentered +
            Group_coded:LexTALEcentered +
            Group_coded: VALDOMcentered +
        # By-subject random effects:
           (1 + cond1 +cond2|Participant) +
        # By-item random effects:
           (1 + Group_coded +LexTALEcentered |item),
                data=modeldata_clmm,
                link="cloglog",
                control = clmm.control("nlminb"),
                threshold = "equidistant")


#Then, a stepwise model selection was employed to find out which model has the best fit
#1. No covariates
model_ratings_FULL_nocovariates <- clmm(rating ~
                  # Fixed main effects:
                      cond1 +
                      cond2 +
                      Group_coded +
                  # Fixed 2-way interactions with Group:
                      Group_coded:cond1 +
                      Group_coded:cond2 +
                  # By-subject random effects:
                      (1 + cond1 +cond2|Participant) +
                  # By-item random effects):
                      (1 + Group_coded |item),
                      data=modeldata_clmm,
                      link="cloglog",
                      control = clmm.control("nlminb"))

#2. No covariates apart from lexTALE
model_ratings_FULL_nocovariates <- clmm(rating ~
                 # Fixed main effects:
                 cond1 +
                 cond2 +
                 Group_coded +
                 LexTALEcentered +
                 # Fixed 2-way interactions involving Group:
                 Group_coded:cond1 +
                 Group_coded:cond2 +
                 Group_coded:LexTALEcentered +
                 # By-subject random effects:
                 (1 + cond1 +cond2|Participant) +
                 # By-item random effect:
                 (1 + Group_coded +LexTALEcentered |item),
                data=modeldata_clmm,
                link="cloglog",
                control = clmm.control("nlminb"))

#3. With lexical frequency, bigram frequency and lextale
model_ratings_FULL_nocovariates3 <- clmm(rating ~
                 # Fixed main effects:
                 cond1 +
                 cond2 +
                 Group_coded +
                 LEXFREQcentered +
                 BGFREQcentered +
                 LexTALEcentered +
                 # Fixed 2-way interactions involving Group:
                 Group_coded:cond1 +
                 Group_coded:cond2 +
                 Group_coded:LEXFREQcentered +
                 Group_coded:BGFREQcentered +
                 Group_coded:LexTALEcentered +
                 # By-subject random effects:
                 (1 + cond1 +cond2|Participant) +
                 # By-item random effects:
                 (1 + Group_coded +
                LexTALEcentered |item),
              data=modeldata_clmm,
              link="cloglog",
              control = clmm.control("nlminb"))

#4. Just Lexical frequency
model_ratings_FULL_nocovariates4 <- clmm(rating ~
                # Fixed main effects:
                    cond1 +
                    cond2 +
                    Group_coded +
                    LEXFREQcentered +
                # Fixed 2-way interactions involving Group:
                    Group_coded:cond1 +
                    Group_coded:cond2 +
                    Group_coded:LEXFREQcentered +
                # By-subject random effects:
                    (1 + cond1 +cond2|Participant) +
                 # By-item random effects:
                    (1 + Group_coded |item),
                    data=modeldata_clmm,
                    link="cloglog",
                    control = clmm.control("nlminb"))

#5. Lexical frequency & LexTALE
model_ratings_FULL_nocovariates5 <- clmm(rating ~
                 # Fixed main effects:
                   cond1 +
                   cond2 +
                   Group_coded +
                   LEXFREQcentered +
                   LexTALEcentered +
                 # Fixed 2-way interactions with Group:
                   Group_coded:cond1 +
                   Group_coded:cond2 +
                   Group_coded:LEXFREQcentered +
                   Group_coded:LexTALEcentered +
                 # By-subject random effects:
                   (1 + cond1 +cond2|Participant) +
                 # By-item random effects:
                   (1 + Group_coded +LexTALEcentered |item),
                   data=modeldata_clmm,
                   link="cloglog",
                   control = clmm.control("nlminb"))

#6. Length and orthogonal neighbours taken out
model_ratings_FULL_nocovariates6 <- clmm(rating ~
               # Fixed main effects:
               cond1 +
               cond2 +
               Group_coded +
               LEXFREQcentered +
               BGFREQcentered +
               CONCRETEcentered +
               LexTALEcentered +
               VALDOMcentered +
               # Fixed 2-way interactions involving Group:
               Group_coded:cond1 +
               Group_coded:cond2 +
               Group_coded:LEXFREQcentered +
               Group_coded:BGFREQcentered +
               Group_coded:CONCRETEcentered +
                 Group_coded:LexTALEcentered +
                 Group_coded: VALDOMcentered +
                 # By-subject random effects:
                 (1 + cond1 +cond2|Participant) +
                 # By-item random effects:
                 (1 + Group_coded +LexTALEcentered |item),
                data=modeldata_clmm,
                link="cloglog",
                control = clmm.control("ucminf"))

#7. Chosen model. We only decided to include HA and LA words in the final mode, and VALDOM was dropped to
#keep the model consistent with the other models in the thesis
model_ratings_FULL <- clmm(rating ~
                 # Fixed main effects:
                 condition_coded+
                 Group_coded +
                 LEXFREQcentered +
                 BGFREQcentered +
                 CONCRETEcentered +
                 LEN_orthoNcentered +
                 LexTALEcentered +
                 # Fixed 2-way interactions involving Group:
                 Group_coded:condition_coded+
                 Group_coded:LEXFREQcentered +
                 Group_coded:BGFREQcentered +
                 Group_coded:CONCRETEcentered +
                 Group_coded:LEN_orthoNcentered +
                 Group_coded:LexTALEcentered +
                 # By-subject random effects:
                 (1 + condition_coded|Participant) +
                 # By-item random effects:
                 (1 + Group_coded|item),
                data=modeldata_clmm,
                link="cloglog",
                control = clmm.control("nlminb"))
```

Pupillary response

```
model_area_expt4 <- glmer(area_under_curve ~
           # Fixed main effects:
            condition_coded +
            Group_coded +
            LEXFREQcentered +
            BGFREQcentered +
            CONCRETEcentered +
            LEN_orthoNcentered +
            LexTALEcentered +
           # Fixed 2-way interactions involving Group:
            Group_coded:condition_coded +
            Group_coded:LEXFREQcentered +
            Group_coded:BGFREQcentered +
            Group_coded:CONCRETEcentered +
            Group_coded:LEN_orthoNcentered +
            Group_coded:LexTALEcentered +
           # By-subject random effects:
            (1 + condition_coded+
              LEXFREQcentered +
              BGFREQcentered +
              CONCRETEcentered +
              LEN_orthoNcentered |Participant) +
           # By-item random effects:
            (1 + Group_coded|item),
             data=modeldataexpt4,
             family=Gamma(link="identity"),
             control  = glmerControl(optimizer  = c("bobyqa"),
             optCtrl=list(maxfun=50000),tol  = .0001))
```

Appendix E: Optimality bias experiment blame measures

The following items were used to measure blame attribution in experiments 5 and 6

Blame Questionnaire :
-To what extent does the doctor deserve blame for their patient's hearing loss?
- How responsible is the doctor for the patient's hearing loss?
-To what extent does the doctor deserve to be punished for her actions?
- How seriously wrong were the doctor's actions?
- How confident are you in your moral judgement?

Comprehension Questions:

- “TRUE or FALSE: the doctor believed that both treatments had a 70% chance of leading to recovery.
- “Given the treatment that the doctor chose, what was the actual chance of that treatment leading to recovery?”
- “Did the doctor have any way of knowing that this belief about the probabilities was false, or was it outside her control?” (answer options: “Yes, there was evidence saying that her belief was incorrect” or “No, it was outside her control.”)

References

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A Systematic Review and Meta-Analysis of the Cognitive Correlates of Bilingualism. *Review of Educational Research, 80*(2), 207-245. doi:DOI: 10.3102/0034654310368803

Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., & Lagnado, D. A. (2015). Causal Conceptions in Social Explanation and Moral Evaluation: A Historical Tour. *Perspect Psychol Sci, 10*(6), 790-812. doi:10.1177/1745691615601888

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *Journal of Philosophy, 108*, 670-696.

Altarriba, J., & Canary, T. M. (2004). The Influence of Emotional Arousal on Affective Priming in Monolingual and Bilingual Speakers. *Journal of Multilingual and Multicultural Development, 25*(2-3), 248-265. doi:10.1080/01434630408666531

Anooshian, L. J., & Hertel, P. T. (1994). Emotionality in Free Recall: Language Specificity in Bilingual Memory. *Cognition and Emotion, 8*(6), 503-514.

Ayçiçegi-Dinn, A., & Caldwell-Harris, C. L. (2009). Emotion-memory effects in bilingual speakers: A levels-of-processing approach. *Bilingualism: Language and Cognition, 12*(3), 291-303.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in Cognitive Sciences, 11*(8), 327-332. doi:https://doi.org/10.1016/j.tics.2007.06.003

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology, 59*(1), 617-645. doi:10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in cognitive science, 2*(4), 716-724.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:<doi:10.18637/jss.v067.i01>

Baumeister, J. C., Foroni, F., Conrad, M., Rumiati, R. I., & Winkielman, P. (2017). Embodiment and Emotional Memory in First vs. Second Language. *Frontiers in Psychology, 8*, 394. doi:10.3389/fpsyg.2017.00394

Bernat, E. M., Cadwallader, M., Seo, D., Vizueta, N., & Patrick, C. J. (2011). Effects of instructed emotion regulation on valence, arousal, and attentional measures of affective processing. *Developmental Neuropsychology, 36*(4), 493-518. doi:10.1080/87565641.2010.549881

Besemeres, M. (2006). Language and emotional experience: The voice of translingual memoir. In A. Pavlenko (Ed.), *Bilingual Minds* (pp. 34-58): Mutlilingual matters.

Bond, M., & Lai, T. (1986). Embarrassment and Code-Switching into a Second Language. *The Journal of Social Psychology, 126*(2), 179-186.

Boucsein, W. (2012). Methods of Electrodermal Recording. In *Electrodermal Activity* (pp. 87-258). Boston, MA: Springer US.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion, 1*(3), 276-298. doi:10.1037/1528-3542.1.3.276

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*(4), 602-607. doi:10.1111/j.1469-8986.2008.00654.x

Braun, M. (2015). Emotion and language-when and how comes emotion into words? Comment on “The Quartet theory of human emotions: An integrative and neurofunctional model” by S. Koelsch et. al. *Physics of life reviews, 13*, 36-37.

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of cognition, 1*(1), 9-9. doi:10.5334/joc.10

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904-911. doi:10.3758/s13428-013-0403-5)

Caldwell-Harris, C. L., & Aycicegi-Dinn, A. (2009). Emotion and lying in a non-native language. *International Journal of Psychophysiology, 71*(3), 193-204. doi:10.1016/j.ijpsycho.2008.09.006

Caldwell-Harris, C. L., Tong, J., Lung, W., & Poo, S. (2010). Physiological reactivity to emotional phrases in Mandarin—English bilinguals. *International Journal of Bilingualism, 15*(3), 329-352. doi:10.1177/1367006910379262

Chen, P., Lin, J., Chen, B., Lu, C., & Guo, T. (2015). Processing emotional words in two languages with one brain: ERP and fMRI evidence from Chinese-English bilinguals. *Cortex, 71*, 34-48.

Christensen, R. H. B. (2019). ordinal - Regression Models for Ordinal Data (Version 2019.4-25). Retrieved from http://www.cran.r-project.org/package=ordinal/

Cipolletti, H., McFarlane, S., & Weissglass, C. (2015). The Moral Foreign-Language Effect. *Philosophical Psychology, 29*(1), 23-40. doi:10.1080/09515089.2014.993063

Conrad, M., Recio, G., & Jacobs, A. M. (2011). The Time Course of Emotion Effects in First and Second Language Processing: A Cross Cultural ERP Study with German-Spanish Bilinguals. *Frontiers in Psychology, 2*, 1-16. doi:10.3389/fpsyg.2011.00351

Corey, J. D., Hayakawa, S., Foucart, A., Aparici, M., Botella, J., Costa, A., & Keysar, B. (2017). Our moral choices are foreign to us. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1109-1128.

Costa, A., Foucart, A., Arnon, I., Aparici, M., & Apesteguia, J. (2014). “Piensa” twice: on the foreign language effect in decision making. *Cognition, 130*(2), 236-254.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLoS One, 9*(4).

Costa, A., Vives, M. L., & Corey, J. D. (2017). On Language Processing Shaping Decision Making. *Current Directions in Psychological Science, 26*(2), 146-151. doi:10.1177/0963721416680263

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353-380. doi:10.1016/j.cognition.2008.03.006

De Freitas, J., & Johnson, S. G. B. (2018). Optimality bias in moral judgment. *Journal of Experimental Social Psychology, 79*, 149-163. doi:10.1016/j.jesp.2018.07.011

De Houwer, A. (2005). Early bilingual acquisition. In *Handbook of bilingualism: Psycholinguistic approaches* (pp. 30-48).

DeBruine, L., & Barr, D. J. (2019). Understanding mixed effects models through data simulation.

DeBruine, L. M. (2019). Experimentum.

Degner, J., Doycheva, C., & Wentura, D. (2011). It matters how much you talk: On the automaticity of affective connotations of first and second language words. *Bilingualism: Language and Cognition, 15*(01), 181-189. doi:10.1017/s1366728911000095

Dewaele, J.-M. (2004). The Emotional Force of Swearwords and Taboo Words in the Speech of Multilinguals. *Journal of Multilingual and Multicultural Development, 25*(2-3), 204-222. doi:10.1080/01434630408666529

Dewaele, J.-M. (2008). The emotional weight of I love you in multilinguals' languages. *Journal of Pragmatics, 40*(10), 1753-1780. doi:10.1016/j.pragma.2008.03.002

Dewaele, J.-M. (2016). Thirty shades of offensiveness: L1 and LX English users' understanding, perception and self-reported use of negative emotion-laden words. *Journal of Pragmatics, 94*, 112-127. doi:10.1016/j.pragma.2016.01.009

Dewaele, J. (2010a). *Emotions in multiple languages*. Basignstoke: Palgrave-MacMillan.

Dewaele, J. M. (2010b). Christ fucking shit merde! Language preferences for swearing among maximally proficient multilinguals. *Sociolinguistic Studies, 4*(3), 595-614.

Dewaele, J. M. (2011). Self-reported use and perception of the L1 and L2 among maximally proficient bi-and multilinguals: A quantitative and qualitative investigation. *International Journal of the Sociology of Language, 208*, 25–51.

Dewaele, J. M. (2017). Why the dichotomy 'L1 versus LX user'is better than 'native versus non-native speaker'. *Applied Linguistics, 39*(2), 236-240.

Dewaele, J. M. (2018). "Cunt": On the perception and handling of verbal dynamite by L1 and LX users of English. . *Multilingua, 37*(1), 53-81.

Dewaele, J. M., & Pavlenko, A. (2001-2003). Web questionnaire Bilingualism and Emotions. *University of London.*

Diaz-Lago, M., & Matute, H. (2018). Thinking in a Foreign language reduces the causality bias. *Quarterly Journal of Experimental Psychology, 72*(1), 41-51. doi:10.1177/1747021818755326

Dudschig, C., de la Vega, I., & Kaup, B. (2014). Embodiment and second-language: automatic activation of motor responses during processing spatially associated L2 words and emotion L2 words in a vertical Stroop paradigm. *Brain & Language, 132*, 14-21. doi:10.1016/j.bandl.2014.02.002

Eilola, T. M., & Havelka, J. (2010). Behavioural and physiological responses to the emotional and taboo Stroop tasks in native and non-native speakers of English. *International Journal of Bilingualism, 15*(3), 353-369. doi:10.1177/1367006910379263

Eilola, T. M., Havelka, J., & Sharma, D. (2007). Emotional activation in the first and second language. *Cognition & Emotion, 21*(5), 1064-1076. doi:10.1080/02699930601054109

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour, 8*(2), 97-120. doi:10.1016/j.trf.2005.04.012

Fan, L., Xu, Q., Wang, X., Xu, F., Yang, Y., & Lu, Z. (2018). The automatic activation of emotion words measured using the emotional face-word Stroop task in late Chinese-English bilinguals. *Cognition and Emotion, 32*(2), 315-324. doi:10.1080/02699931.2017.1303451

Fan, L., Xu, Q., Wang, X., Zhang, F., Yang, Y., & Liu, X. (2016). Neural Correlates of Task-Irrelevant First and Second Language Emotion Words-Evidence from the Emotional Face-Word Stroop Task. *Frontiers in Psychology, 7*, 1672.

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., . . . Pallier, C. (2011). Comparing Word Processing Times in Naming, Lexical Decision, and Progressive Demasking: Evidence from Chronolex. *Frontiers in Psychology, 2*(306). doi:10.3389/fpsyg.2011.00306

Ferré, P., García, T., Fraga, I., Sánchez-Casas, R., & Molero, M. (2010). Memory for emotional words in bilinguals: Do words have the same emotional intensity in the first and in the second language? *Cognition & Emotion, 24*(5), 760-785. doi:10.1080/02699930902985779

Foroni, F. (2015). Do we embody second language? Evidence for 'partial' simulation during processing of a second language. *Brain and Cognition, 99*, 8-16. doi:10.1016/j.bandc.2015.06.006

Gao, S., Zika, O., Rogers, R. D., & Thierry, G. (2015). Second language feedback abolishes the "hot hand" effect during even-probability gambling. *Journal of Neuroscience, 35*(15), 5983-5989. doi:10.1523/JNEUROSCI.3622-14.2015

Garrido, M. V., & Prada, M. (2018). Comparing the valence, emotionality and subjective familiarity of words in a first and a second language. *International Journal of Bilingual Education and Bilingualism*, 1-17. doi:10.1080/13670050.2018.1456514

Gathercole, V. C. M., & Moawad, R. A. (2010). Semantic interaction in early and late bilinguals: All words are not created equally. *Bilingualism: Language and Cognition, 13*(4), 385-408. doi:10.1017/s1366728909990460

Geipel, J., Hadjichristidis, C., & Surian, L. (2015a). The Foreign Language Effect on Moral Judgment: The Role of Emotions and Norms. *PLoS One, 10*(7), e0131529. doi:10.1371/journal.pone.0131529

Geipel, J., Hadjichristidis, C., & Surian, L. (2015b). How foreign language shapes moral judgment. *Journal of Experimental Social Psychology, 59*, 8-17. doi:https://doi.org/10.1016/j.jesp.2015.02.001

Grabovac, B., & Pléh, C. (2014). Emotional activation measured using the emotional Stroop task in early Hungarian-Serbian bilinguals from Serbia. *Acta Linguistica Hungarica, 61*(4), 423-441. doi:10.1556/ALing.61.2014.4.3

Greenberg, A., Bellana, B., & Bialystok, E. (2013). Perspective-Taking Ability in Bilingual Children: Extending Advantages in Executive Control to Spatial Reasoning. *Cognitive Development, 28*(1), 41-50. doi:10.1016/j.cogdev.2012.10.002

Grosjean, F. (2008). *Studying bilinguals*. Oxford: Oxford University Press.

Guglielmo, S., & Malle, B. F. (2017). Information-Acquisition Processes in Moral Judgments of Blame. *Personality and Social Psychology Bulletin, 43*(7), 957-971. doi:10.1177/0146167217702375

Haapalainen, E., Kim, S. J., Forlizzi, J. F., & Dey, A. K. (2010). *Psycho-physiological measures for assessing cognitive load*. Paper presented at the Proceedings of the 12th ACM international conference on Ubiquitous computing.

Hadjichristidis, C., Geipel, J., & Surian, L. (2019). Breaking magic: Foreign language suppresses superstition. *Quarterly Journal of Experimental Psychology, 72*(1), 18-28. doi:10.1080/17470218.2017.1371780

Haegerich, T. M. B., B.L. (2000). Empathy and Juror's Decisions in Patricide Trials Involving Child Sexual Assault Allegations. *Law and Human Behaviour, 24*(4), 421-448.

Harris, C. L. (2004). Bilingual Speakers in the Lab: Psychophysiological Measures of Emotional Reactivity. *Journal of Multilingual and Multicultural Development, 25*(2-3), 223-247. doi:10.1080/01434630408666530

Harris, C. L., Ayçíçeğí, A., & Gleason, J. B. (2003). Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language. *Applied Psycholinguistics, 24*(4), 561-579.

Harris, C. L., Gleason, J. B., & Ayçiçeği, A. (2006). When is a First Language more Emotional? Psychophysiological Evidence from Bilingual Speakers. In A. Pavlenko (Ed.), *Bilingual Minds: Emotional experience, expression, and representation* (pp. 257-283): Clevedon, Multilingual Matters.

Hayakawa, S., Costa, A., Foucart, A., & Keysar, B. (2016). Using a Foreign Language Changes Our Choices. *Trends in Cognitive Science, 20*(11), 791-793. doi:10.1016/j.tics.2016.08.004

Hayakawa, S., & Keysar, B. (2018). Using a foreign language reduces mental imagery. *Cognition, 173*, 8-15.

Hayakawa, S., Tannenbaum, D., Costa, A., Corey, J. D., & Keysar, B. (2017). Thinking More or Feeling Less? Explaining the Foreign-Language Effect on Moral Judgment. *Psychol Sci, 28*(10), 1387-1397. doi:10.1177/0956797617720944

Hooper, N., Erdogan, A., Keen, G., Lawton, K., & McHugh, L. (2015). Perspective taking reduces the fundamental attribution error. *Journal of Contextual Behavioral Science, 4*(2), 69-72. doi:https://doi.org/10.1016/j.jcbs.2015.02.002

Hsu, C. T., Jacobs, A. M., & Conrad, M. (2015). Can Harry Potter still put a spell on us in a second language? An fMRI study on reading emotion-laden literature in late bilinguals. *Cortex, 63*, 282-295.

Iacozza, S., Costa, A., & Dunabeitia, J. A. (2017). What do your eyes reveal about your foreign language? Reading emotional sentences in a native and foreign language. *PLoS One, 12*(10), e0186027. doi:10.1371/journal.pone.0186027

Ivaz, L., Costa, A., & Dunabeitia, J. A. (2016). The emotional impact of being myself: Emotions and foreign-language processing. *J Exp Psychol Learn Mem Cogn, 42*(3), 489-496. doi:10.1037/xlm0000179

Jonczyk, R., Boutonnet, B., Musial, K., Hoemann, K., & Thierry, G. (2016). The bilingual brain turns a blind eye to negative statements in the second language. *Cognitive, Affective & Behavioural Neuroscience, 16*(3), 527-540. doi:10.3758/s13415-016-0411-x

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*, 136-153.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-291.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201-208). New York, NY: Cambridge University Press.

Kahneman, D. M., D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*, 136-153.

Kahneman, D. T., A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York, NY: Cambridge University Press.

Kapa, L. L., & Colombo, J. (2013). Attentional Control in Early and Later Bilingual Children. *Cognitive Development, 28*(3), 233-246. doi:10.1016/j.cogdev.2013.01.011

Kazanas, S. A., & Altarriba, J. (2016). Emotion Word Processing: Effects of Word Type and Valence in Spanish-English Bilinguals. *Joural of Psycholinguistic Research, 45*(2), 395-406. doi:10.1007/s10936-015-9357-3

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*(3), 627-633.

Keysar, B. (1994). The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology, 26*, 165-208.

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect: thinking in a foreign tongue reduces decision biases. *Psychol Sci, 23*(6), 661-668. doi:10.1177/0956797611432178

Khan, A., & Rayner, G. (2003). Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem. *Journal of Applied Mathematics and Decision Sciences, 7*(4), 187-206.

Klesse, A.-K., Levav, J., & Goukens, C. (2015). The Effect of Preference Expression Modality on Self-Control. *Journal of Consumer Research*. doi:10.1093/jcr/ucv043

Kloosterman, N. A., Meindertsma, T., van Loon, A. M., Lamme, V. A. F., Bonneh, Y. S., & Donner, T. H. (2015). Pupil size tracks perceptual content and surprise. *European Journal of Neuroscience, 41*(8), 1068-1078. doi:10.1111/ejn.12859

Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition, 112*(3), 473-481. doi:10.1016/j.cognition.2009.06.007

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of memory and language, 33*(2), 149-174.

Kuchinke, L., Vo, M. L., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology, 65*(2), 132-140. doi:10.1016/j.ijpsycho.2007.04.004

Kühne, K., & Gianelli, C. (2019). Is Embodied Cognition Bilingual? Current Evidence and Perspectives of the Embodied Cognition Approach to Bilingual Language Processing. *Frontiers in Psychology, 10*(108). doi:10.3389/fpsyg.2019.00108

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. *Journal of Experimental Psychology: General, 143*(3), 1065-1081.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44*(2), 325-343. doi:10.3758/s13428-011-0146-0

Leuchs, L., Schneider, M., & Spoormaker, V. I. (2019). Measuring the conditioned response: A comparison of pupillometry, skin conductance, and startle electromyography. *Psychophysiology, 56*(1), e13283. doi:10.1111/psyp.13283

Li, P., Zhang, F. A. N., Tsai, E., & Puls, B. (2013). Language history questionnaire (LHQ 2.0): A new dynamic web-based research tool. *Bilingualism: Language and Cognition, 17*(03), 673-680. doi:10.1017/s1366728913000606

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328-348. doi:https://doi.org/10.1016/j.jesp.2018.08.009

Mandel, D. R. (2010). Predicting blame assignment in a case of harm caused by negligence. *Mind & Society, 9*, 5-17.

Mante-Estacio, J., & Bernardo, A. B. I. (2015). Illusory Transparency in Bilinguals: Does Language of Text Affect Bilingual Readers' Perspective Taking in Reading? *Current Psychology, 34*, 744–752.

Masto, M. (2015). Empathy and Its Role in Morality. *The Southern Journal of Philosophy, 53*(1), 74-96. doi:10.1111/sjp.12097

Monaco, E., Jost, L. B., Gygax, P. M., & Annoni, J.-M. (2019). Embodied Semantics in a Second Language: Critical Review and Clinical Implications. *Frontiers in Human Neuroscience, 13*(110). doi:10.3389/fnhum.2019.00110

Nash, J. C., & Varadhan, R. (2011). Unifying Optimization Algorithms to Aid Software System Users: optimx for R. *Journal of Statistical Software, 43*(9), 1-14.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13*(1), 45-52.

Nikula, R. (1991). Psychological Correlates of Nonspecific skin Conductance Responses. *Psychophysiology, 28*(1), 86-90.

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). *Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks.* Paper presented at the Proceedings of the 24th Australian Computer-Human Interaction Conference.

Ong, E. L. C., Hussain, S., Chow, Y., & Thompson, C. (2017). Variations in Bilingual Processing of Positive and Negative Information. 36-42. doi:10.5176/2251-1865_cbp17.14

Opitz, B., & Degner, J. (2012). Emotionality in a second language: It's a matter of time. *Neuropsychologia, 50*(8), 1961-1967.

Ożańska-Ponikwia, K. (2017). Expression and perception of emotions by Polish-English bilinguals I love you vs. Kocham Cię. *International Journal of Bilingual Education and Bilingualism*, 1-12. doi:10.1080/13670050.2016.1270893

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies, 59*(1-2), 185-198. doi:10.1016/s1071-5819(03)00017-x

Pavlenko, A. (2005). *Emotions and multilingualism*. Cambridge, UK: Cambridge University Press. .

Pavlenko, A. (2006). Bilingual selves. In A. Pavlenko (Ed.), *Bilingual minds: Emotional experience, expression, and representation* (pp. 1-33). Clevedon: Multilingual Matters.

Pavlenko, A. (2008). Emotion and emotion-laden words in the bilingual lexicon. *Bilingualism: Language and Cognition, 11*(02). doi:10.1017/s1366728908003283

Pavlenko, A. (2012). Affective processing in bilingual speakers: disembodied cognition? *International Journal of Psychology, 47*(6), 405-428. doi:10.1080/00207594.2012.743665

Pelham, S. D., & Abrams, L. (2014). Cognitive advantages and disadvantages in early and late bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(2), 313-325. doi:10.1037/a0035224

Ponari, M., Rodríguez-Cuadrado, S., Vinson, D., Fox, N., Costa, A., & Vigliocco, G. (2015). Processing advantage for emotional words in bilingual speakers. *Emotion, 15*(5), 644-652.

Puntoni, S., de Langhe, B., & van Osselaer, S. M. J. (2009). Bilingualism and the Emotional Intensity of Advertising Language. *Journal of Consumer Research, 35*(6), 1012-1025. doi:10.1086/595022

Räsänen, S. H. M., & Pine, J. M. (2012). Emotional force of languages in multilingual speakers in Finland. *Applied Psycholinguistics, 35*(03), 443-471. doi:10.1017/s0142716412000471

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research (Version 1.8.12). Northwestern University, Evanston, Illinois, USA. Retrieved from https://CRAN.R-project.org/package=psych

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. (pp. 173-220). New York: Academic Press.

Rubio-Fernández, P. (2017). Why are bilinguals better than monolinguals at false-belief tasks? *Psychonomic Bulletin & Review, 24*(3), 987-998.

Ryskin, R. A., Brown-Schmidt, S., Canseco-Gonzalez, E., Yiu, L. K., & Nguyen, E. T. (2014). Visuospatial perspective-taking in conversation and the role of bilingual experience. *Journal of Memory and Language, 74*, 46-76.

Scheepers, C. (2014). *Between-group matching of confounding variables: Why covariates remain important for analysis*. Paper presented at the AMLaP, Edinburgh.

Scheepers, C., Mohr, S., Fischer, M. H., & Roberts, A. M. (2013). Listening to Limericks: a pupillometry investigation of perceivers' expectancy. *PLoS One, 8*(9), e74986. doi:10.1371/journal.pone.0074986

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: evidence from pupillometry. *Front Psychol, 5*, 137. doi:10.3389/fpsyg.2014.00137

Schoemaker, P. J. H. (1991). The quest for optimality: A positive heuristic of science? *Behavioural and Brain Sciences, 14*, 205-245.

Scott, G. G., O'Donnell, P. J., & Sereno, S. C. (2014). Emotion words and categories: evidence from lexical decision. *Cognitive Processing, 15*(2), 209-215. doi:10.1007/s10339-013-0589-6

Segalowitz, N., Trofimovich, P., Gatbonton, E., & Sokolovskaya, A. (2008). Feeling affect in a second language: The role of word recognition automaticity. *The Mental Lexicon, 3*(1), 47-71. doi:10.1075/ml.3.1.05seg

Sheikh, N. A., & Titone, D. (2016). The embodiment of emotional words in a second language: An eye-movement study. *Cogn Emot, 30*(3), 488-500. doi:10.1080/02699931.2015.1018144

Surrain, S., & Luk, G. (2019). Describing bilinguals: A systematic review of labels and descriptions used in the literature between 2005–2015. *Bilingualism: Language and Cognition, 22*(2), 401-415. doi:10.1017/S1366728917000682

Sutton, T. M., Altarriba, J., Gianico, J. L., & Basnight-Brown, D. M. (2007). The automatic access of emotion: Emotional Stroop effects in Spanish-English bilingual speakers. *Cognition & Emotion, 21*(5), 1077-1090. doi:10.1080/02699930601054133

Tao , L., Marzecová, A., Taft, M., Asanowicz, D., & Wodniecka, Z. (2011). The Efficiency of Attentional Networks in Early and Late Bilinguals: The Role of Age of Acquisition. *Frontiers in Psychology,* 2(123). doi:10.3389/fpsyg.2011.00123

Taylor, J. E., Beith, A., & Sereno, S. C. (2019). LexOPS: An R Package and User Interface for the Controlled Generation of Word Stimuli. doi:https://doi.org/10.31234/osf.io/7sudw

Toda, M. (1991). The human being as a bumbling optimalist: A psychologist's viewpoint. *Behavioral and Brain Sciences, 14*(2), 235. doi:10.1017/S0140525X00066401

Toivo, W., & Scheepers, C. (2019). Pupillary responses to affective words in bilinguals' first versus second language. *PLoS One, 14*(4), e0210450. doi:10.1371/journal.pone.0210450

Vo, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods, 41*(2), 534-538. doi:10.3758/BRM.41.2.534

Warriner, A. B., Kuperman, V., & Brysbaert. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*, 1191-1207. doi:10.3758/s13428-012-0314-x)

Wierzbicka, A. (2008). A conceptual basis for research into emotions and bilingualism. *Bilingualism: Language and Cognition, 11*(02). doi:10.1017/s1366728908003362

Winskel, H. (2013). The emotional Stroop task and emotionality rating of negative and neutral words in late Thai-English bilinguals. *International Journal of Psychology, 48*(6), 1090-1098. doi:10.1080/00207594.2013.793800

Wu, Y. J., & Thierry, G. (2012). How reading in a second language protects your heart. *Journal of Neuroscience, 32*(19), 6485-6489. doi:10.1523/JNEUROSCI.6119-11.2012

Xue, J., Marmolejo-Ramos, F., & Pei, X. (2015). The linguistic context effects on the processing of body-object interaction words: An ERP study on second language learners. *Brain Research, 1613*, 37-48. doi:https://doi.org/10.1016/j.brainres.2015.03.050

Zwaan, R. A. (2014). Embodiment and language comprehension: reframing the discussion. *Trends in Cognitive Sciences, 18*(5), 229-234. doi:https://doi.org/10.1016/j.tics.2014.02.008