THE UNIVERSITY *of* EDINBURGH

Multilingualism in Later Life: Natural History & Effects of Language Learning

Brittany Blankinship

Doctor of Philosophy

The University of Edinburgh

2023

"It wasn't curiosity that killed the cat. It was trying to make sense of all the data curiosity generated"

– Halcom (pronounced "how come?" as in "why"), Michael Quinn Patton's "Internal philosophical alter ego and muse" (p524 of Patton, 2015)

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own.

Abstract

The overarching aim of this thesis is to explore the question of what role the knowledge and use of multiple languages plays in ageing. To answer this question two approaches were taken: first a natural history perspective on how languages change across the lifespan; second a training strategy investigating the effects of language learning in later life.

Chapter 1 reviews the state of the art in terms of language learning in the third age. Chapter 2 outlines various methodological considerations and contexts of the presented empirical studies. Chapter 3 presents two empirical studies of natural history: (a) a completed study where multilingual adults across the lifespan participated in an autobiographical experiment investigating the role of language in relation to the vividity and intensity of memories and (b) on-going study exploring reported changes in the language abilities in later life. These studies reveal a dynamic role of language across the lifespan of multilinguals and the inescapable importance of individual variation and contextualised differences.

The natural history investigation is followed by language learning classes across the lifespan where participants learned (a) Spanish for 2 hours 3 times a week across 4 weeks (24 hours total; Chapter 5), (b) written and spoken Mandarin Chinese 2 hours 2 times a week for 8 weeks (32 hours total; Chapter 6), or (c) Scots for 2 hours, 3 times a week for 3 weeks (18 hours total; Chapter 7). Apart from different languages being chosen, each chapter investigates a related aspect of the language. That is, the three selected languages function as an illustration of three general questions.

In Chapter 5, Spanish was selected as the target language due to being one of the most popular foreign languages in the United Kingdom, though not typically taught in schools as is the case with French or German. Therefore, participants entering the course were complete beginners, regardless of age.

This chapter concentrates on the different ages of participants and the relationships between learning success (as measured by quizzes) and the cognitive effects of language learning. Chapter 6 explores the question of modality (auditory and visual) of language learning and cognitive measures, benefitting from the unique feature of Mandarin Chinese in which the written and spoken language are dissociated to a much larger degree than in other living languages. Chapter 7 examines the issue of how the social status of a language might influence the learning process using Scots, which is debated in some circles as being a dialect of English rather than a language.

These language learning studies were designed and analysed using a mixed methods approach, integrating both quantitative and qualitative data and analysis. In adopting a mixed methods approach to answering questions around the impact of language learning across the lifespan, the participant's opinions and experiences as well as cognitive variables are taken into consideration, creating a wholistic picture representative of lived experiences.

Lastly, two primarily qualitative studies are presented in Chapter 7. One completed study which explores online language learning with older adults and provides recommendations for improving learning experiences, and another which is still in progress. Following from the practicalities of online learning with students in the third age, preliminary results are presented from a focus group study which then informed an international online survey looking at the experience of mixed age adult language learning classroom dynamics and practicalities from the point of view of teachers as well as learners.

It is my hope that this thesis will make a two-fold contribution: (1) via the empirical findings of the individual studies and (2) in terms of exploring the importance of different methodologies and methodical considerations in answering diverse research questions.

Lay Summary

The research presented in this thesis aimed to investigate language across the lifespan among adults who know and use more than one language (that is, are multilingual).

The first chapter reviews the available scientific findings on language learning in older age, which is followed in the second chapter by a discussion and examination of the various considerations around the methods of data collection and analysis in the studies presented in Chapters 3 through 7. The research presented throughout this thesis combines quantitative (meaning numbers based, countable and measurable) and qualitative (meaning descriptive, words and interpretation based) data collection; this combination of quantitative and qualitative is referred to as "a mixed methods approach". In adopting a mixed methods approach to answering questions around the impact of language learning across the lifespan, the analysis can consider both the participant's opinions and experiences as well as cognitive variables (e.g., attention).

Chapter 3 presents two studies. First, a completed study exploring how specific languages can influence the way in which participants recalled different events from their lives. This is followed by a study still in progress exploring the types and patterns of language changes reported in later life. These studies reveal that language is adaptable and changes across the lifespan of multilingual adults, rather than being determined only in childhood and locked in place throughout adulthood.

These studies addressing changes in language usage, preferences and attitudes and the influence of the choice of specific languages on memory (Chapter 3), laid the foundation for studies of language learning with mixed aged adult classes (Chapters 4, 5, and 6). Participants learned (a) Spanish for 2 hours 3 times a week across 4 weeks (24 hours total; Chapter 5), (b) written and

spoken Mandarin Chinese 2 hours 2 times a week for 8 weeks (32 hours total; Chapter 6), or (c) Scots for 2 hours, 3 times a week for 3 weeks (18 hours total; Chapter 7). Apart from different languages being chosen, each chapter investigates a related aspect of the language. That is, the three selected languages function as an illustration of three general questions.

In Chapter 5, Spanish was selected as the language to be learnt because it is one of the most popular foreign languages in the United Kingdom, though not typically taught in schools as is the case with French or German. As such, participants entering the course were complete beginners, regardless of age. This chapter concentrates on the different ages of participants and the relationships between learning success (as measured by quizzes) and the cognitive effects of language learning, focussing particularly on attention. Chapter 6 explores the question of modality (auditory and visual) of language learning and cognitive measures, benefitting from the unique feature of Mandarin Chinese in which the written and spoken language are dissociated to a much larger degree than in other living languages. Chapter 7 examines the issue of how the social status of a language might influence the learning process using Scots, which is debated in some circles as being a dialect of English rather than a language in its own right.

Lastly, two studies are presented in Chapter 7 exploring practical applications of mixed aged adult language learning classes online and from the perspective of teachers and students. One completed study which explores online language learning with older adults and provides recommendations for improving learning experiences, and another which is still in progress. Following from the practicalities of online learning with students in the third age, preliminary results are presented from a focus group study which then informed an international online survey looking at the experience of mixed age

adult language learning classroom practicalities and perceptions from the point of view of teachers as well as learners.

It is my hope that this thesis will make a two-fold contribution: (1) via the findings of the individual studies and (2) in terms of exploring the importance of different approaches to data collection and analysis in answering diverse research questions.

Acknowledgements

For and foremost I must thank my biggest supporters, my parents. I could not have done this PhD without you, both in the literal and figurative sense. Your unwavering support and encouragement are a constant blessing and the foundation and driver of my dreams and success(es) past, present, and future. Thank you for answering the non-stop onslaught of questions I asked as a child - my curiosity has not stopped, but I now have a few letters after my name. As the first Dr Blankinship in our family, my new title is as much yours as it is mine. I finished my last bit of homework, and my book is done - for now. Thank you for everything.

To my sister, thank you for being the very best friend and sibling anyone could ask for. My greatest luck in life has and always will be being your sister.

While I am surely forgetting someone, I would like to specifically thank the various teachers who have inspired and encouraged me prior to my higher education journey. I am blessed to have been taught by passionate and brilliant teachers from my very first foray into education in kindergarten with Miss Towle, through elementary school with Miss Ellis and Mrs. Boswith, and into high school where the support and encouragement of Mlle Neumann and Mrs. Karkut inspired me to dream of (and succeed in) attending a big-well established university while attending a small high school. Thank you for being role models of the type of teacher I aspire to be. An extra special not of thanks to Mlle Neumann for instilling a love of a languages and opening the world of linguistics to me. Merci beaucoup.

I could not have done this degree without the support of my PhD besties and Taco Tue-Wednesday pals including but not limited to Alicia, Carine, Eva-Maria, and Austin. Eva-Maria, I am looking forward to café dates with a book to read for fun - rather than stressing about our PhDs.

I would also like to thank Mara for being not only a great friend, but also a fantastic colleague. Collaborating with you on our paper has been one of the most pleasant academic experiences I have had. I am so happy you stopped by my poster at that online conference in 2020.

To Tanya, thank you for being the very best friend a gal could ask for and for always being there for me. You support and encouragement means the world. Your random texts of encouragement through the final weeks of my write up

provided the extra boost I needed. Looking forward to the day we can both put Dr in front of our names!

To Christie, the very best flatmate and friend a PhD student could ever dream of, I quite simply could not have survived this degree and a pandemic in the middle of it without you. It is not hyperbolic to say that I would have starved and gone insane in the last few months of this degree if it were not for you (and Rob and Sam!). Thank you for supporting me through the various stages of this summer (and the PhD generally) from amaretto biscuit to the single wafer slice of a kit kat, and everything in between.

I would also like to thank Christie and her family for helping me get a desk during lockdown when it was clear that everything would move to working from home for the foreseeable future. The very desk from which I write this! The vast majority of this PhD was written here, so thank you!.

Thank you to each and every one of my participants and collaborators on the studies presented within this thesis.

I would, of course, also like to thank my incredible supervision team.

- Dr Vega-Mendoza, thank you for being the best mentor, role model, and supporter that a young woman in science could hope for. Your support through my MSc and the beginning of my PhD means the world.
- Dr Flynn, thank you for being so generous with your time and for sharing your immense wisdom and knowledge.
- I had the privilege and pleasure of working with my primary supervisor for my master's project as well as a project as a research assistant prior to starting my PhD work. From this experience I knew there was no one else I would want to do with PhD with. Dr Bak, I cannot thank you enough for being an ever supportive and encouraging mentor, colleague, and friend. I appreciate so much the various opportunities and experiences you have provided for me during my PhD journey. It has been a dream to work with a supervisor who not only entertains my all of ideas, but also supports and encourages them.

Grandma Peggy, thank you for providing me with the greatest possible role model of what a strong, intelligent, and incredibly fashionable woman in the modern world looks like. I hope I have made you proud and will continue to do so.

Uncle Rick, I will enjoy a nice cold beverage in celebration and in honour of you.

Abbreviations

ANT = Attention Network Test

APM = (Raven's) Advanced Progressive Matrices

APQ = Autotelic Personality Questionnaire

CFL = Chinese as a Foreign Language

CLMM = Cumulative Link Mixed Model

DA = Discourse Analysis

DCI = Delayed Character Introduction

EGIDS = Expanded Graded Intergenerational Disruption Scale

EF = Executive Functions

ESM = Experience Sampling Method

ET = Elevator Counting Task

ETD = Elevator Counting Task with Distraction

ETR = Elevator Counting Task with Reversal

FDR = False Discovery Rate

FWER = Family-wise Error Rate

GAPS = General Abstract Processing System

GRS = Graphic Rating Scale

GT = Grounded Theory

ICC = Intraclass Correlation Coefficient

ICI = Immediate Character Introduction

IPA = Interpretative Phenomenological Analysis

L1 = Language one, referring to a speaker's first language

LDR = Language Dependent Recall

LX = Language X, referring to additional languages (any number, hence X) from L1

LMM = Linear Mixed Model, AKA multilevel models or hierarchical models

NSHT = Null Hypothesis Significance Testing

NRS = Numeric Rating Scale

MCMC = Markov chain Monte Carlo

RCT = Randomised Controlled Trial

RT = Reaction Time

RTA = Reflexive Thematic Analysis

SCMS = Synchronous Computer Mediated Communication

TA = Thematic Analysis

TEA = Test of Everyday Attention

VAS = Visual Analogue Scale

Contents

Chapter 1 Revisiting Old Dogs and New Tricks: Different Approaches to Language Learning in Later Life

The majority of this chapter comes from a first-author paper currently under review at the International Journal of Language and Communication Disorders as part of an invited submission to the Thematic Issue "Language as a window to cognitive changes in aging: methodological issues". Sections 1.1 and 1.7 have been written specifically for this chapter, all other sections of the chapter (1.2-1.6) are presented as in the submitted review paper.

Blankinship, B. and Bak, T. H. (under review). Revisiting old dogs and new tricks: different approaches to language learning in later life. *International Journal of Language and Communication Disorders.*

1.1 Ageing

The continuous increase in life expectancy over the last few hundred years has been hailed as one of the greatest achievements of human civilisation, from the remarkable progress of medical sciences to a general improvement in living standards, including nutrition, access to clean water, and many other related factors. However, together with the decreasing birth rate, it has also led to a radical and still ongoing change in the age structure of the populations across the world. The United Nations estimates that there are 962 million people today over the age of 60. By 2030, the number of people 60 years or older is expected to increase 34% from 2019 to 1.4 billion worldwide (UN DESA, 2020). Further, by 2050, the global population of older adults is expected to more than double from 2019 rates to 2.1 billion. In most countries, the proportion of adults aged 60 years and older will increase from 1 in 8 in 2017 to 1 in 6 by 2030 and further to 1 in 5 by 2050 (UN DESA, 2020). In fact, 2020 was the first year in history wherein people over 60 outnumbered children

under 5 in the world population (UN DESA, 2020). Within the ageing world population (i.e., 60 plus), the 80 plus population is the fastest growing. Indeed, projections suggest that globally, the 80 plus population will increase 233% between 2008 and 2040 as compared to 160% for those aged 65 and over (Kinsella & Wan, 2009).

Two topics which have been discussed in connection with ageing are (a) loneliness and (b) diseases, both of which are relevant in the context of the present work. While both loneliness and poor health can occur at any age and are by no means an automatic consequence of ageing, the likelihood of both increases with age. Loneliness in older age may be related to lifestyle changes that can accompany ageing such as retirement, children moving out of the house, or bereavement (e.g., due to widowhood) (Carr & Bodnar-Deren, 2009; National Academies of Sciences, Engineering, and Medicine, 2020). Further, these two factors also interact such that increasingly poor health due to age-related diseases (e.g., hearing problems, impaired vision, cognitive deficits, mobility issues, possible shame round physical symptoms) can also contribute to loneliness and isolation (Hackett et al., 2019; National Academies of Sciences, Engineering, and Medicine, 2020). At the same time, loneliness and isolation can contribute to poor health outcomes (National Academies of Sciences, Engineering, and Medicine, 2020).

Indeed, as a function of an ageing world population, loneliness among adults aged 50 plus is an impeding public health concern. While often considered synonymous, loneliness and isolation refer to two related, though distinct conceptual phenomena (Wigfield et al., 2022). As defined by Age UK, isolation "refers to separation from social or familial contact, community involvement, or access to services. Loneliness, by contrast, can be understood as an individual's personal, subjective sense of lacking these things to the

extent that they are wanted or needed" (Care Connect & Age UK, 2018). Thus, isolation is a more objective measure of the number of social contacts a person has, while loneliness is a subjective feeling about the gap between a person's desired levels of social contact and their actual level of social contact. Therefore, one can be isolated without being lonely, or lonely without being isolated.

Often times, however, these two notions are related to one another such that social isolation can lead to loneliness and loneliness can lead to social isolation, or they can occur concurrently. According to projections by Age UK, more than two million people aged 50 and over will be lonely by 2025-26, an expected 49% increase from the 1.36 million in 2015-16 (Age UK, 2019). Notably, the proportion of the population reporting as "often" feeling lonely is not predicted to rise, rather this projection is due to the increasing number of people over the age of 50 in the UK population. Further, those who live with a high degree of loneliness are twice as likely to develop Alzheimer's Disease than those with a low degree of loneliness (Wilson et al., 2007).

Moreover, an ageing population, although in itself a testament to the success of medicine and social care, brings with it a higher prevalence of age-related disorders. The incidence of age-related neurodegenerative diseases is projected to increase not only due to such changing worldwide population demographics, but also longer global life expectancies. Within the UK, for example, approximately 1 in 5 people will live to see their 100th birthday (Age UK, 2019; Department of Work and Pensions, 2011). However, even as the population is set to continue ageing at an unprecedented pace, there is little evidence that older adults today are healthier than previous generations (World Health Organization, 2015). As the expected proportion of an individual's life in good health has not increased with life expectancy, this

implies that for many people these additional years are spent in poor health. With an aging population, age-related neurodegenerative diseases such as Alzheimer's Disease (AD) and Parkinson's Disease (PD) have become increasingly more common (Reeve et al., 2014; Reitz et al., 2011). Indeed, the single biggest risk factor for developing Alzheimer's Disease or Parkinson's Disease is ageing (Reeve et al., 2014).

In later life, dementia is one of the leading causes of disability, ranking in the top five across 51 countries according to the Global Disease Burden 2019 data (BMJ, 2022). In the UK, dementia-related disability rates in later life are higher than some cancers, cardiovascular disease and even stroke (Age UK, 2019). Further, in the UK, dementia has been the leading cause of death for women since 2011 (Age UK, 2019; Alzheimer's Research UK, 2022). In fact, even during the Covid-19 pandemic in 2020 more women died from dementia-related health issues than from Covid-19 (Alzheimer's Research UK, 2022). On a global scale, death as a result of dementia and dementia-related complications more than doubled between 2000 and 2016 (World Health Organization, 2018). Considering all ages and sexes, dementia as of 2019 is the 7th leading cause of global death, as compared to 14th in 2000, ranking above kidney diseases, diabetes, and road injury (World Health Organization, 2020b).

As the probability of developing dementia increases with age, the global trends of both an ageing population and increased life expectancy point to the fact that age related neurodegenerative diseases will not only increase in prevalence in the coming years, but also have grievous social, economic, cultural, and political implications. The impact of dementia can be considered on three different but inter-related levels: the individuals living with dementia, their family and friends, and wider society. Outwith the obvious impact of poor health in later life on the individual and their family/support network, there are

immense social and economic consequences. In terms of annual societal and economic expense, the global cost of dementia in 2015 was US$ 818 billion, a 35% increase from 2010 (Wittenberg et al., 2019). Indeed, in 2019, the estimated global societal cost of dementia surpassed the trillion mark, at US$ 1.3 trillion (World Health Organization, 2021). Due to both the increasing cost of care and number of people living with dementia, the total global societal cost is predicted to surpass US$ 2.8 trillion by 2030. Within the UK, the cost of dementia is projected to raise to £94.1 billion in 2040 (Alzheimer's Society, 2019). In terms of social care and health care costs in the UK, dementia is higher than cancer and chronic heart disease combined (Luengo-Fernandez et al., 2015).

Therefore, there is a need for factors which could be addressed by conscious interventions or activities that may help to promote healthy ageing. In fact, the United Nations General Assembly has announced 2021-2030 to be the "Decade of Health Ageing", with the World Health Organisation taking lead on implementation (World Health Organization, 2020a). Health ageing is defined as:

> "developing and maintaining the functional ability that enables well-being in older age. Functional ability is determined by the intrinsic capacity of an individual (i.e., the combination of all the individual's physical and mental capacities), the environment in which he or she lives (understood in the broadest sense and including physical, social and policy environments) and the interactions among them" (World Health Organization, 2020a, p. 3).

Cognitive exercise/mental activity has been suggested to be a protective factor against dementia, even in later life (Valenzuela & Sachdev, 2009a, 2009b). Indeed, in a meta-analysis of 22 studies, including more than 29,000 individuals, Valenzuela and Sachdev (2006) found an overall dementia risk reduction of 46% in older adults with high level of mental activity in later life

compared to those with low activity, independent of other predictors. The magnitude of this effect was similar to that of education, occupational complexity, and cognitive lifestyle. Similarly, robust social networks have also been suggested to reduce the risk of dementia, as well as contribute to overall well-being (Fratiglioni et al., 2000). Controlling for depression and baseline cognition, Fratiglioni and colleagues (2000) found that limited social networks and dissatisfying social contacts resulted in a 60% increase in dementia risk in their community-based cohort study of 1203 older adults. Language learning is a cognitive exercise that may promote healthy cognitive ageing as well as provide a social benefit. Indeed, in the WHO *Decade of Healthy Ageing: Plan of Action (2020a)* report a key element of healthy ageing is lifelong learning.

1.2 Why language learning in later life matters

Although the interest in language learning in later life might appear as relatively recent, stimulated by demographic changes in the population, as well as research on cognitive reserve (both of which to be discussed later on), language learning across the lifespan is likely to be as old as human language itself. Most hunter-gatherer and early agricultural societies surviving today are multilingual and it is likely that language itself developed in a multilingual environment (N. Evans, 2018). This means that not only do children grow up learning different languages from early childhood, but also that language learning continues well into later life: indeed, perfecting language skills is seen as one of the central characteristics of successful aging (Sorensen, 1967).

In contrast, much of the literature on language learning in the past two centuries seemed to assume that language learning is mainly a domain of the young, happening in the context of a school or university. In situations in which older people needed to learn a new language (e.g., immigrants arriving in a

country with a different language) the target of language learning was mainly instrumental, to achieve a "communicative ability" necessary to function in the new country and society. One of the few authors who recognised the important potential of language learning in later life well ahead of his time was David Singleton (e.g., Singleton (2005)).

The recent rise in interest in lifelong learning, including language learning, has both demographic as well as neuroscientific origin. As the life expectancy grows continuously, while the retirement age in many countries remains the same, people spend higher and higher percentage of their life in retirement (UN DESA, 2019). Learning a new language or brushing up the knowledge of a language learned earlier in the course of life can appear as a worthwhile and rewarding activity. Accordingly, older people make up a growing proportion of language classes and conversational circles (as will be demonstrated in several examples in the following section).

This tendency has been additionally strengthened by recent research, suggesting that learning and using more than one language can have cognitive benefits, not only in childhood, but also in a more advanced age. There is growing evidence that bilingualism can slow down cognitive ageing (Bak, Nissan, et al., 2014), delay the onset of dementia by around 4 years (Alladi et al., 2013; Anderson et al., 2020; Bialystok et al., 2007) and counteract the cognitive symptoms after stroke (Alladi et al., 2016; Paplikar et al., 2019). This literature is further complemented by the evidence of positive effects of language learning on attention, in participants from 18 to 85 years old and after only a one-week intensive course of Scottish Gaelic (Bak et al., 2016a; Long et al., 2020).

All these findings tend to be interpreted in the light of the current notion of "cognitive reserve" (M. Antoniou et al., 2013; Bak & Mehmedbegovic,

2017; Pfenninger & Singleton, 2019; Stern, 2002), a term popularised by Stern (2002) to account for individual differences in manifested cognitive decline in the face of disease, referring to the ability to better tolerate neurodegenerative pathology in the brain before displaying behavioural symptoms. A recent review by Livingston et al (2020) estimated that 40% of dementia risk could be modified through lifestyle factors, the majority of which are thought to confer cognitive reserve. However, cognitive reserve is a multifaceted and complex construct for which there is currently not a universally agreed upon specific definition or complete set of proxies (Stern et al., 2020).

One of the distinctions made within the concept of cognitive reserve is that between a "structural/brain reserve", preventing the development of pathological changes, and "functional reserve", compensating for the emerging pathology (Stern et al., 2020; Valenzuela & Sachdev, 2009b). Both can co-exist and interact. Accordingly, some studies have reported a lesser degree of pathological changes in bilinguals (Bak & Robertson, 2017; Estanga et al., 2017), which would point to a structural reserve associated with bilingualism. However, the evidence for functional reserve is stronger, suggesting that bilinguals can achieve a better cognitive performance despite a higher level of brain atrophy (Bialystok et al., 2021; Schweizer et al., 2012).

It has been suggested that the mechanism through which bilingualism leads to a higher cognitive reserve is mediated through improved executive functioning (Grundy, 2020). In order to manage multiple language or linguistic systems, the current theory posits that bi-/multilinguals continually rely on domain general cognitive functions to analyse their linguistic environment and employ the appropriate language, simultaneously supressing the other(s) (Kroll et al., 2014). Language use, and therefore language learning, recruit a widespread neural network including the temporal, frontal, and parietal

regions, which overlaps with the network known to decline with age (M. Antoniou et al., 2013).

However, the growing interest in this field of research leads to an increasing realisation of its complexity, in particular in terms of the large number of interacting variables (Bak, 2016). In terms of linguistic features of bi-/multilingualism, while the importance of the age of acquisition has been recognised relatively early, more recent research pays more attention to the patterns of language use, in particular code-switching (D. W. Green & Abutalebi, 2013). Moreover, findings cannot be generalised across populations, as bilingualism can be associated with different variables, particularly once we move beyond the few officially monolingual countries in which most of the research has been conducted (Alladi et al., 2016).

In the case of language learning, the situation is equally complex. The frequency, intensity, length, and type of language learning can all play an important role, in addition to the complex interaction between the spoken and the written form of language. The characteristics of the language learners, from age and education, through their knowledge of other languages, to their motivation and attitudes, have all to be taken into account. The choice of language and its linguistic distance from the language(s) spoken by the participants might play a role as well. Finally, vastly different outcome measures can be used, from different measures of success in language learning, through psychometric tests of cognitive function, to subjective measures of satisfaction and wellbeing.

In this situation, a straight comparison of different studies, not taking into account their specific aims, populations and methodologies, can be misleading. The aim of this review is, therefore, not to count studies and compare the numbers of "positive" and "negative" results, but rather to

examine carefully all the relevant parameters of the available studies in order to determine possible patterns of associations between the above-mentioned variables and the cognitive outcomes.

1.3 Defining the topic: Different forms of language learning in later life

Historically, "language learning" both colloquially and as a research field has implicitly referred to language acquisition in younger age. Indeed, systematic research looking at language learning in the so-called third age has only recently started to emerge (Pot et al., 2019) though interest in the topic dates back to the 1970s (e.g., Kalfus 1977). The third age is generally thought of as beginning with retirement, meaning that the exact age boundaries are often culturally or nationally defined. The British University of the Third Age (U3A), founded in 1981, define this period as "time in your life (not necessarily chronological) where you have the opportunity to undertake learning for its own sake. There is no minimum age, but a focus on people who are no longer in full-time employment or raising a family" (*U3A – Our Story*, n.d.). Likewise, the lower age boundary and definition of "older person" or "elderly" varies according to different sources.

Also, outside of academia, it is more and more common for seniors to seek out new learning opportunities, in particular languages. In the UK, for example, the U3A currently has over 400,000 members across 1,000 different groups (*U3A – Our Story*, n.d.). In Edinburgh, Scotland, for example, there are a range of venues for language learning for adults of all ages such as the Languages for All classes at the Centre for Open Learning who provide a wide variety of classes at different levels for "all adult learners, from 16 to 99+ year old" (Center for Open Learning, n.d.). Yakety Yak Language Café, in both

Edinburgh and Glasgow, is another popular language learning avenue for adults of all ages, particularly seniors, where students are able to "yak" (i.e., speak) in small groups in a relaxed, friendly café environment with an experienced tutor (Yakety Yak, n.d.). In practice, language learning nowadays is no longer by default for teenagers or young children.

There is a multiplicity of reasons why a senior learner might wish to learn a new language including for fun, as mental exercise, for social reasons, to communicate with family, to travel, etc. (Garcia, 2017). The often-implicit assumption of a younger learner brings with it inferred motivations (e.g., primary communicative or employment purposes) and markers of success (e.g., exams) which match this age group, but may not be appropriate for third age learners.

1.4 Language learning in older age: A review of the literature

Similar to the question of the cognitive impact of lifelong bi- or multilingualism, findings in the field of third age language learning, burgeoning though still relatively small, are also mixed. In a recent systemic review of empirical studies published prior to November 2020, Ware and colleagues (2021) found that of the nine studies included, four of the five highest quality rated studies, as measured by Risk of Bias per the US Department of Health and Human Services, showed "significant increases in attentional switching (Bak et al., 2016a), cognitive inhibition (Pfenninger & Polz, 2018), working memory (Wong et al., 2019), or functional connectivity (Bubbico et al., 2019), therefore providing some moderate evidence for increases in cognitive and cerebral functioning after short-term second language training in seniors" (p. 7). Language learning has also been found to provide a myriad

of positive socio-affective effects including increased social contact as well as feelings of social inclusion and improved self-esteem (Klimova et al., 2020; Pfenninger & Polz, 2018). While there are a handful of high-quality reviews in this area (e.g., Pot et al., (2019); Ware et al., (2021)), quite a few studies have been published around third age language learning in the past two years (three in 2020 and five in 2021, respectively), meriting an updated review.

1.4.1 Conditions and findings: Summary of the available studies

To our knowledge, there are 14 published studies, across 18 papers, which have experimentally investigated language learning in later life among healthy older adults (i.e., no cognitive decline beyond what is typical with age; see Table 1.1). Indeed, many of the experimental studies published in this emerging field are pilot studies or studies of feasibility. Table 1.1 outlines the research questions and design elements, Table 1.2 describes the participants and interventions/training, and Table 1.3 details the outcomes measures and results of experimental health-ageing, third-age language learning studies to date. Of note are the differences in methodologies within this limited sample (e.g., inclusion of control groups and follow-up testing, different language levels, different delivery or teaching methods, or the duration and intensive of the language learning experiences; see Table 1.2 and Table 1.3). Ten of the thirteen studies published after 2013 cite Antoniou and colleagues (2013) influential paper as a foundation for the principles behind why language learning might be an effective means of combating cognitive decline in older age, though as seen in Table 1.1, Table 1.2, and Table 1.3 there is a wide variety of operationalization of these suggestions. While not all studies set out to investigate the cognitive impact of language learning, among the majority that

did (79%), there is a wide variety of tests used with very little overlap between research groups.

Table 1.1 Design variables of experimental healthy-ageing, third-age language learning studies to date

Authors	Primary Research question/hypothesis	Country	Design	Group assignment	Groups	Periods of measurement
Mackey & Sachs, 2012	"Do older learners of English as an L2 show improvement in L2 question formation following task-based interaction with feedback? Is there evidence for a relationship between working memory capacity and interaction-driven learning in these older learners?" (p. 713)	USA	No control group	NA	(1) language learning group	Four • pre-test • post-test • 1-week follow up • 5-week follow up
Bak et al., 2016	"To determine whether learning a new language would lead to an improvement in cognitive performance as early as one week after an intensive course" (p.2)	Scotland	Controlled trial	Non-random, age, gender & education matched	(1) language learning group (2) active control group (other courses) (3) passive control group	Three • pre-test • post-test • 9 months follow-up
Cox, 2017	"Learning in older age through interactions of learner-internal and -external variables; specifically, late-learned L2 (bilingualism) and provision of grammar explanation (explicit instruction, EI)" (p. 29)	USA	Quasi-experimental	Semi-random (order of recruitment)	mono-/late-bilingual with/without explicit instruction (4 groups)	Three • pre-test • post-test • 2-week follow-up
Ramos et al., 2017	"Exploring the relationship between language learning and switching ability in elderly monolingual	Spain	Controlled trial	Non-random, age matched	(1) language learning group	Two • pre-test • post-test

	participants who learned a second language during a whole academic year." (p. 39)				(2) passive control group	
Ware et al., 2017	"To examine the participant's subjective experience of learning a foreign language in order to explore the feasibility of the program" (p. 2)	France	Primary qualitative MM; no control group	NA	language learning group	Two • pre-test • post-test
Kliesch et al., 2018; Kliesch, Girous & Meyer, 2021	"To investigate successful language learning in old age…exploring the factors that account for the individual differences between older learners themselves" (p. 54); "to assess whether neurological substrates of L2 learning identified in younger adults can be found in older learners" (p. 145)	Switzerland	No control group, semi-MM but no description of qualitative analysis	NA	language learning group	Two • pre-test • post-test
Pfenninger & Polz, 2018	"To what extent does an intensive, four-week EFL training in a school context impact on (a) FL gains, (b) general performance of cognitive skills known to deteriorate as a function of age, and (c) (linguistic) self-confidence, FL learning motivation, general communicative skills and overall well-being?" (p. 3)	Austria	Equal-status concurrent mixed methods design	Non-random	(1) bilingual language learning group (2) active control group (monolingual language learning group)	Three • pre-test • after 2 weeks • post-test
Bubbico et al., 2019	"Investigating neuroplastic-related effects of second language learning in terms of cognitive and brain	Italy	RCT	Random	(1) language learning group	Two • pre-test • post-test

	networks functional connectivity changes" (p. 8)				(2) passive control group	
Wong et al., 2019	"To prospectively test the potential cognitive enhancing effect of foreign language learning in older adults with no clear signs of cognitive decline beyond what is age typical" (p. 2441)	Hong Kong/China	RCT	Stratified randomization (strata: age, years of education)	(1) language learning group (2) active control group (playing games like sodoku) (3) passive control group (music appreciation)	Three • pre-test • post-test • 3-month follow-up
Valis et al., 2019; Klimova et al., 2020	"To what extent foreign language learning may enhance cognitive functions among healthy older population" (p. 1311); "the effect of learning a non-native language on the enhancement of cognitive performance in healthy native Czech elderly with a focus on qualitative analysis" (p. 2)	Czech Republic	MM RCT	Random	(1) beginner language learning group (2) low intermediate language learning group (3) passive control group	Two • pre-test • post-test
Berggren et al., 2020; Nilsson et al., 2021	Testing the hypothesis that "foreign language learning in older age results in larger improvements of cognitive ability than participation in an active control condition" (p. 213); "to investigate the effect of language training on brain structure in older adults in several language-and memory-related gray matter regions and	Sweden	RCT	Stratified randomization (strata: age, word-word associative memory performance at pretest)	(1) language learning group (2) active control group (relaxation training)	Two • pre-test • post-test

	white matter tracts, all specified *a priori"* (p. 2)					
Long et al., 2020	"Factors responsible for individual differences in domain general cognitive functions following language learning" (p. 807)	Scotland	No control group	Non-random	(1) LX beginner (2) LX elementary (3) LX intermediate	Two • pre-test • post-test
Kleisch & Pfenninger, 2021; Kleisch et al., 2021	"Nonlinear cognitive trajectories across a 30-week training period" (p. 1)	Switzerland	Controlled trial	Non-random (voluntary group assignment), background variables and socio-affect matched	(1) language learning group (2) active control group (strategy game training) (3) passive control group (movie screenings)	Thirty (weekly basis)
Meltzer et al., 2021	"To examine the cognitive benefits of language learning in a very specific app-based format, comparing it to a well-matched condition involving the same amount of time spent engaged in a learning app on a tablet or smartphone." (p. 13)	Canada	RCT	Stratified randomization (strata: age, education level)	(1) language learning group (2) brain training group (BrainHQ game) (3) passive control group	Two • pre-test • post-test

Note: MM, mixed method; RCT, randomised controlled trial

Table 1.2 Description of participants and interventions in experimental healthy-ageing, third-age language learning studies to date

Authors	**Participants: *n* analysed of *n* recruited**	**Participants: gender, age range**	**L1 and LX learned**	**Intensity and frequency of language learning**	**Duration of language learning**	**Total hours of language learning**	**Delivery method**
Mackey & Sachs, 2012	9 of 9	5 female, 4 male 65-89	L1: Spanish LX: English	5 sessions	5 weeks	Sufficient information not provided	Communication sessions with trained native speakers
Bak et al., 2016	67 of 76	46 female, 31 male 18-78	L1: English LX: Gaelic	1 week	1 week	~14 hrs	In person, classes taught by trained teachers
Cox, 2017	45 of 62	27 female, 18 male 60–82	L1: English or L1/2: English-Spanish LX: Latin	2 sessions within a week	Sufficient information not provided	Sufficient information not provided	Online software
Ramos et al., 2017	43, number originally recruited not reported	21 female, 22 male 60–80	L1: Spanish LX: Basque	2 hrs sessions 2x and 1.5 hrs 1x a week	8 months (32 weeks)	176 hrs	In person, classes taught by trained teachers
Ware et al., 2017	14 of 14	9 female, 5 male 63–90	L1: French LX: English	2 hrs 1x a week	4 months (16 weeks)	32 hrs	In person, translating sentences from L2 to L1 in class with a trained teacher and using online support (dictionaries, videos)

Kliesch et al., 2018; Kliesch, Girous & Meyer, 2021	10 of 10	4 female, 6 male 65–73	L1: German or Swiss-German LX: English	1 hr 4x a day	3 weeks	60 hrs	In person, classes taught by trained teachers
Pfenninger & Polz, 2018	12 of 19	8 female, 4 male 63–89	L1: German L1/2: German-Slovenian LX: English	2 hrs 3x a week	4 weeks	24 hrs	In person, classes taught by trained teachers
Bubbico et al., 2019	26 of 30	19 female, 7 male 59–79	L1: Italian LX: English	2 hrs 1x a week	4 months (16 weeks)	32 hrs	In person, classes taught by native teachers
Wong et al., 2019	153 of 235	130 female, 23 male 60–85	L1: Cantonese LX: English	5hrs a week plus occasional social activities	6 months (24 weeks)	78-130 hrs	Online software (Rosetta Stone)
Valis et al., 2019; Klimova et al., 2020	42 of 60 screened	Not reported Age range not reported (mean age 70.9)	L1: Czech LX: English	45 min 3x a week	3 months (12 weeks)	27 hrs	In person, classes taught by trained teachers
Berggren et al., 2020; Nilsson et al., 2021	160 of 170, subset of 76 for MRI	100 female, 60 male 65–75	L1: Swedish LX: Italian	2.5 hrs classes 2x a week	11 weeks	55 hrs	In person, classes taught by trained teachers
Long et al., 2020	105 of 132 (27 removed from analysis)	67 female, 38 male 21–85	L1: English LX: Gaelic	1 week	1 week	~ 14 hrs	In person, classes taught by trained teachers
Kleisch & Pfenninger, 2021;	61 of 65	26 female, 35 male 64-75	L1: German or Swiss German LX: Spanish	5 hrs a week	30 weeks	150 hrs	App software (Duolingo) with weekly classroom

Kleisch et al., 2021							sessions taught by trained teachers
Meltzer et al., 2021	76 of 95	51 females, 25 males 65-75	L1: English LX: Spanish	30 minutes a day 5x a week	16 weeks	40 hrs	App software (Duolingo)

Table 1.3 Outcome measures and results of experimental healthy-ageing, third-age language learning studies to date

Authors	**Outcome measures (domain + tests)**	**Analysis**	**Result**	**Broad Classification of Findings**
Mackey & Sachs, 2012	Working memory • listening-span task, nonword recall task Communicative tasks • spot-the- difference, picture-drawing, and picture-sequencing Question-form production in recorded sessions	correlation (Pearson's r), question production based on a scale from previous literature	Participants who showed LX progression in advanced question formation were those who scored higher in L1 working memory tasks	NA – cognitive effect not the primary outcome
Bak et al., 2016	Attention • TEA	2-way mixed ANOVA, linear trend analysis, t-test	Significant improvement in cognitive abilities tested, maintained in 9 month follow up for those who practice the language 5+ hours a week	"positive"
Cox, 2017	Processing speed • Digital-Symbol Coding Task Education level • NAART 35 4 LX assessments • interpretation both written and aural, grammaticality judgment, and written production	repeated-measures ANOVA, planned one-way ANOVA and Tukey posthoc tests	No significant overall influence of explicit instruction, bilinguals outperformed monolinguals in learning of Latin morphosyntax	NA – cognitive effect not the primary outcome

Ramos et al., 2017	Switching ability • Colour-shape switching task Language knowledge • LexTALE	ANOVA, Bayesian Null Hypothesis Testing, Bayesian *t*-test	No significant improvement in cognitive abilities tested	"negative"
Ware et al., 2017	Global cognition • MoCA Loneliness and social isolation • University of California Loneliness Assessment Scale Semi-structured interviews	paired sample *t*-test, content/theme analysis	No significant improvement in cognitive abilities tested or loneliness scores, program found to be feasible	"negative"
Kliesch et al., 2018; Kliesch, Girous & Meyer, 2021	EEG • language-switching ERP experiment Language test • C-Test, Hueber assessment test Next A1, oral translation test Inhibition • Stroop Task, Eriksen Flanker Task Switching • Categorical Shifting Task, Figural Switching Task Working memory • RST, Digit Span, Verbaler Lern- und Merkfähigkeits- test Delayed recall • Verbaler Lern- und Merkfähigkeits-test	EFA, *t*-test, correlation (Pearson's *r* and Spearman's ρ*), qualitative analysis not described;* pairwise Spearman correlation, LMM	All participants improved in LX but those with higher verbal fluency and working memory, higher power in the beta1 band during resting state EEG scores improved the most and those with highest LX proficiency showed smaller N400 effects during a language switching task	NA – cognitive effect not the primary outcome

	Phonetic verbal fluency • Regensburger Wortflüssigkeits-Test IQ • KAI Qualitative investigation of motivation			
Pfenninger & Polz, 2018	LX proficiency tests • C-Test, production and reception Receptive vocabulary, odd-one-out task, association task Inhibition • Stroop task Attention • Alters-Konzentrations-Test Well-being • Multidimensional Mood State Questionnaire Socio-affective questionnaire	Wilcoxon tests, Mann-Whitney U-tests; content/theme analysis	Significant improvement in both groups in (linguistic) self-confidence, communicative skills subjective well-being, and some cognitive abilities tested (inhibition); language learning perceived by participants as social activity, cognitive stimulation and means of connecting with society	"positive"
Bubbico et al., 2019	rs-fMRI Global cognition • MMSE Short- and long-term episodic memory • Babcock Memory test Attention • TMT A, TMT B Cognitive flexibility	correlation (Pearson's *r*), one-way ANOVA with Duncan's post hoc MRT, ANCOVA	Significant improvement in cognitive abilities tested and increased functional connectivity in the rIFG, rSFG, and lSPL	"positive"

	• TMT AB Phonological lexicon • Verbal Fluency (FAS) Executive functioning • Verbal Fluency Test, FAB			
Wong et al., 2019	Working memory • Auditory Reading Span, Wechsler Digit Span Attention • Attention Network Test Inhibition • Simon Task Global cognition, dementia screening • ADAS-Cog • CDR	3x3 repeated measured ANOVAs, *t*-test	Significant improvement in cognitive abilities tested compared to passive control. Language learning and games improved overall cognitive abilities (global cognition, working memory), which was maintained 3 months after the intervention. Language training group greater improvement in working memory and games group in attention.	"positive"
Valis et al., 2019; Klimova et al., 2020	Global cognition • MoCA Language test • Englishtag Reflective writing	two-sample *t*-test, exact Fisher test, paired Wilcoxon test, two-way repeated measures ANOVA, Qualitative/Content Analysis	No significant improvement in cognitive abilities tested over and above control groups; most of the participants found the course to be useful in particular as in terms of their social and mental well-being and cognitive activity	"negative"
Berggren et al., 2020; Nilsson et al., 2021	Verbal intelligence • Analogies, Syllogisms, and Verbal Inference Spatial intelligence • Raven's matrices, WASI-II Matrix Task	SEM, Bayesian linear mixed models; repeated measures ANOVA, SEM	No significant improvement in cognitive abilities tested over and above control groups. No evidence of structural changes in memory-related grey and white matter structures following language	"negative"

	Working memory • Numerical updating, n-back Item memory • word-word, face-name, picture-picture Structural MRI		learning activity, regardless of vocabulary proficiency. Baseline hippocampal volume and associated memory ability were robust predictors of vocabulary proficiency.	
Long et al., 2020	Attention • TEA	LMM	LX level predicted attention switching performance: those in higher levels initially outperformed lower levels, however lower levels improved the most	"positive"
Kleisch & Pfenninger, 2021; Kleisch et al., 2021	Working memory • two-Back, Operation Span task Alertness • simple alertness task responding to the presence of a cross Divided attention • following dot with a mouse while doing the Stroop task Verbal fluency • Regensburg Word Fluency Test Socio-affect • overall well-being question • training motivation	GAMM, PCA	No significant improvement in cognitive abilities tested over and above control groups; individuals with lower cognitive baseline improved most in language learning group	"negative"
Meltzer et al., 2021	Executive function • N-back, Simon Task, DKEFS battery conditions 1 and 3	LMM, repeated measures ANOVA	Language learning more enjoyable than brain training games. App based language learning provided	"positive"

	Global cognition • MoCA Satisfaction questionnaire		similar cognitive effects as brain training games (incongruent Stroop colour naming, 2-back task), though less impact than on processing speed (reaction times in N-back and Simon task). Both experimental conditions outperformed the control group in cognitive functions test.	

Note: ADAS-Cog, Alzheimer's Disease Assessment Scale – Cognitive subscale; ANOVA, Analysis of Variance; CDR, Clinical Dementia Rating; DKEFS, Delis-Kaplan executive function system; EFA, Exploratory Factor Analysis; FAB, Frontal Assessment Battery; GAMM, Generalized Additive Mixed Models, ISPL, left superior parietal lobule; KAI, Kurztest für allgemeine Informationspsychologische Basisgrössen; LMM, linear mixed model (AKA mixed-effects models or multilevel models); MMSE, Mini Mental State Examination; MoCA, Montreal Cognitive Assessment; MRT, multiple range test; NAART, North American Adult Reading Test; PCA, Principal Component Analysis; rIFG, right Inferior Frontal Gyrus; rSFG, right Superior Frontal Gyrus; RST, Reading Span Task; SEM, structural equation modelling; TEA, Test of Everyday Attention; TMT, Trail Making Test.

1.4.2 Factors which may influence the results

All included studies are interventional, that is some sort of language learning takes place; therefore, they are all longitudinal (i.e., at least two time points of measurement – before and after the intervention). Only four studies utilized longitudinal follow-up periods, ranging from two-weeks to nine months. All studies included cognitive outcome measures of some variety, though these measures were not always the primary outcome. Three studies (Bubbico et al., 2019; Kliesch, Giroud, et al., 2021; Nilsson et al., 2021) used both neuroimaging and cognitive outcome measures.

Concerning study designs, nine studies included a control group, six of which included some level of randomization. In terms of control groups, seven studies employed passive control groups, six active control groups, and four both active and passive control groups. Only three studies did not include control groups in their design. Most studies employed a solely quantitative methodology; four studies included some level qualitative methods as well (Kliesch et al., 2018; Klimova et al., 2020; Pfenninger & Polz, 2018; C. Ware et al., 2017). Ware et al. (2017), in fact, employed a primary qualitative mixed methods design, meaning the primary research question was a qualitative one. Only Pfenninger and Polz (2018) employed an explicitly described mixed methods design, while the other two studies utilized qualitative methods though the study design or qualitative approaches were not sufficiently described. Only two studies (Bak et al., 2016a; Long et al., 2020) included participants across the lifespan or mixed-aged adult groups, while the remaining 12 recruited solely older adults ranging in age from 59 to 90.

The cognitive outcome measures employed across the studies vary quite drastically, with forty-one different tests and 13 described domains,

though "executive functioning" broadly defined covers the vast majority of measures. The depth of heterogeneity can be seen in the divergent operationalization of measures within a particular domain. For example, eight working memory, seven attention, and five inhibition/switching tasks were used. In fact, even the description of cognitive domains under investigation across studies vary, with "executive functioning" broadly defined or more specific aspects of executive functioning used. It is not a surprise, in light of this heterogeneity, that there are not consistent findings across the field. Furthermore, measures of global cognition such as the MMSE or MoCA can be useful for screening purposes but are not appropriate as primary outcome measures in a non-patient population.

With regard to the intervention itself, the studies employed a multitude of frequencies, durations, and delivery methods. Language learning interventions varied from 30 minutes to 2.5-hour sessions and from one week to eight months in duration. A myriad of session frequencies across the period of the interventions were also used, ranging from four times a day to one time a week. The intensity of training programs also fluctuated from around 14 hours to approximately 176 hours (M = 64 hours, SD = 54 hours, Mdn = 40 hours). Similarly, the delivery method of interventions varied from entirely independent online study to in-person, group-based classes. Typologically, a variety of languages were included as the target language: Scottish Gaelic, Italian, Spanish, Latin, Basque, and English. English was the most common target language, seen in seven studies, followed by Spanish and Scottish Gaelic employed in two studies each. Although the second language learning field has developed from work in the area of bi-/multilingualism, the vast majority of studies included functionally monolingual participants, with only Cox (2017) and Pfenninger and Polz

(2018) setting out to compare language learning in bilingual versus monolingual groups. Cox (2017) found that bilinguals performed better in interpretation of Latin morphosyntax regardless of the instruction style condition. Pfenninger and Polz (2018) found both groups significantly improved in the outcomes measured before and after language learning, though monolinguals performed better on a test of inhibitory attention (Stroop task) and significantly improved on a concentration test compared to bilinguals.

Taken together, it is not surprising that such a heterogenous approaches have led to mixed results, especially as it relates to cognition. In terms of overall findings in regarding to the question of a cognitive effects of language learning, six study results could be considered "positive" (i.e., findings supporting a cognitive effect of language learning) and five "negative" (i.e., findings failing to support a cognitive effect of language learning or language learning groups not performing significantly different from control groups; see Table 1.3).

Importantly, more rigorous study designs and larger samples are not more likely to produce negative results, with 50% of the "positive" and 40% of the "negative" studies employing a randomised controlled trial design. Similarly, the mean sample size across both categorizations is 64 ("positive": *SD* = 49, *Mdn* = 67; "negative": *SD* = 50, *Mdn* = 43). The idea that positive results are due to chance findings in small, sloppily designed studies, which have not been replicated in larger and more rigorous ones lacks any empirical basis.

No clear associations could be found between the cognitive measures and different measures of the amount of language learning. On one hand, "negative" studies tend to have more total hours (*M* = 88, *SD* = 62.5, *Mdn* =

55; "positive" *M* = 44.8, *SD* = 37.8, *Mdn* = 32) and to last longer (*M* = 20.2 weeks, *SD* = 8.9, *Mdn* = 16; "positive" *M* = 9.2, *SD* = 8.5, *Mdn* = 4). On the other, "positive" studies seem to be more intense and frequent within a week (*M* = 6.9, *SD* = 5.7, *Mdn* = 5; "negative" *M* = 1.8, *SD* = 0.7, *Mdn* = 2). Any comparison in this field needs, therefore, to distinguish between the total amount and duration on one hand and frequency and intensity on the other, ideally by changing systematically one of the variables, while keeping the others constant.

Finally, we have also observed a small effect of gender, in that "positive" studies tended to have a higher percentage of females (ranging from 40% to 85% of the sample) than the "negative" ones (ranging from 43% to 65% of the sample). However, like in the case of duration and intensity, the small number of studies conducted, and the heterogeneity of their designs precludes any clear conclusions.

1.4.3 Positive and negative results: A useful distinction?

Although in the preceding section we set out to compare studies showing "positive results" (confirming positive effects of language learning on cognitive functions) and "negative results" (not replicating such effects), a closer look at the reported evidence shows that such a dichotomy is by no means clear-cut and self-evident (see Table 1.3). Several studies described dissociations between different aspects of cognitive functions. In line with the expectation that language learning will have a strongest effect on switching behaviour, Bak et al. (2016a) and Long et al. (2020) observed an improvement in attentional switching subtest of the Test of Everyday Attention TEA (Elevator Task with Reversal), but not in other aspects of attention. Wong et al. (2019) reported significant improvement in language learning group compared to

active and passive controls in most cognitive abilities tested (ADAS-Cog, Auditory Reading Span, Boston Naming Test, Attention Network Test), only the Boston Naming Test improved at both post-test and follow up with the language learning group showing greater improvement in working memory and the active control in attention. Meltzer et al. (2021) found that language learning was more enjoyable than brain training games. App based language learning provided similar cognitive effects as brain training games (incongruent Stroop color naming, 2-back task), though less impact than on processing speed (reaction times in N-back and Simon task). Both experimental conditions outperformed the control group. Accordingly, even studies classified as "positive" did not find effects on all measured variables. Indeed, current models of bilingualism, such as the adaptive control hypothesis (D. W. Green & Abutalebi, 2013) predict highly-specific cognitive effects, so a global effect could signal an artefact rather than a genuine finding.

On the other hand, studies which did not find measurable cognitive effects, often found significant results in other domains. Valis et al. (2019) and Klimova et al. (2020) detected no significant improvement in cognitive abilities tested over and above control groups, but most of their participants found the course to be useful in particular in terms of their social and mental well-being and cognitive activity. In the studies by Kleisch and Pfenninger (2021) and Kleisch et al. (2021) there was no significant improvement in cognitive abilities tested over and above control groups; however, individuals with lower cognitive baseline improved most in the language learning group.

Other studies analysed subgroups of participants, reporting significant sub-group differences. In Kleisch et al. (2018) and Kliesch, Girous, & Meyer (2021), all participants improved in the target language but those with higher verbal fluency and working memory, higher power in the beta1 band during

resting state EEG scores improved the most and those with highest target language proficiency showed smaller N400 effects during a language switching task.

In summary, a dichotomization into "positive" vs. "negative" does not adequately capture the full picture. In forcing such a dichotomization, potentially valuable information such as which cognitive tests did or did not show improvement after language learning or which training regimens were viewed favourably by participants risks being lost. Like in other fields of bilingualism research, asking a straightforward yes or no question is likely to produce an answer which is short, simple and clear, but meaningless (Bak & Alladi, 2016).

1.5 Suggestions for future research

1.5.1 Randomised Control Trials: A gold standard for language learning studies?

Randomised controlled trials (RCT) are considered to be the "gold standard" in study designs when investigating casual relationships, particularly in medicine when testing the effectiveness or safety of new drugs and interventions (Hariton & Locascio, 2018). The overarching goals of RCTs is to reduce bias and confounding, that is systematic differences between groups influencing outcomes rather than exposure to an intervention. This study design is prospective, meaning cohorts are followed across time, and comparative, that is different cohorts are contrasted in the variable(s) of interest. Furthermore, RCTs are characterised by (a) the inclusions of at least two groups/cohorts or so-called study "arms" (e.g., an experimental group of interest and one or more control groups), (b) random allocation, (c) allocation

concealment, and (d) single or double blinding (Bhide et al., 2018; Grossman & Mackenzie, 2005; Hariton & Locascio, 2018). In the most straightforward application, there are two groups in a parallel group design: a treatment group/experimental group of interest and a control group. In a parallel group design only one group undergoes the experimental intervention. A more complicated approach are cross-over trials whereby all participants sequentially experience all conditions, which can be particularly relevant if there are numerous interventions of interest. Moreover, control groups can take various forms, receiving no intervention, a standard intervention, or a placebo. In passive control groups, participants often do not experience an active research intervention, that is, they go about living their normal daily life. Active controls, on the other hand, undergo some sort of research intervention, though not the primary experimental condition of interest. Typically, active control groups will partake in an intervention that has been previously shown to be effective or one that controls for additional variables beyond living life as normal.

Random allocation and allocation concealment are related concepts which typically occur at the same time. In an effort of minimizing selection bias, participants have equal opportunity to be allocated to the intervention or control group(s) and are randomly allocated to one of them (i.e., random allocation) and the person conducting the randomization does not know the next treatment allocation (i.e., allocation concealment). In other words, allocation concealment can be thought of as concealment of the randomization sequence. Randomization can take various forms as well – for example, simple, block, stratified or cluster randomization. Simple randomization is based on a single sequence of random assignments, such as a coin flip. With small sample sizes in particular, however, simple randomization

can result in unequal group sizes. Block randomization, on the other hand, ensures participants are randomised into groups of similar sizes. Even so, it could be the case, for example, that at random one group contains the majority of female participants or a much larger age range than another. Stratified block randomization considers potentially influential covariates at the randomization stage by using separate block randomization schemes for each covariate or stratum. Finally, cluster randomization is useful when individuals cannot be randomised themselves, and so hospitals, geographic locations, etc are used as units of randomization allocation instead.

Blinding is a related, though distinct concept which takes place after randomization wherein the treatment condition is hidden during the course of the experiment or trial. Sometimes called masking, blinding can take place on multiple levels: the study participants themselves (i.e., single blind), the participants and the investigators or clinicians managing participants (i.e., double blind), or the participants, investigators or clinicians managing participants as well as data analysts collecting or analysing the data (i.e., triple blinding, assuming clinicians and analysts are different people). As with other conditions in RCTs, blinding is an attempt to reduce bias, principally observer and performance bias, or the famous placebo-effect (Finniss et al., 2010). It is therefore particularly important when the outcome variables of interest are subjective, such as intensity of pain or discomfort. Typically, double blinding is preferred, wherein the study participants and people delivering the treatment or measuring the outcome variables are unaware of which arm of the study/treatment group a participant falls in. Double-blinding, however, is only feasible in studies where a placebo can be given, such as the classical drug trial. The name of this study design, therefore, is descriptive of some of its core conditions: random referring to random allocation of participants to groups

and control referring to the inclusion of a control group(s). Trials wherein no level of blinding is used are referred to as open trials or open-label trials.

RCTs, however, are not without their pitfalls, such as high expense financially and temporally, issues with randomization disincentivizing participants or when using small sample sizes generally (Grossman & Mackenzie, 2005). Furthermore, while evidence from RCTs is considered to be "at the top of the evidence pyramid" (Bhide et al., 2018, p. 380), they are not suitable, or even necessary, for all research questions or fields of enquiry. For example, it is unlikely for there to ever be an RCT conducted on the efficacy of parachutes in preventing deaths while skydiving, though few skydivers will likely argue against parachute use due to the lack of evidence (G. C. S. Smith & Pell, 2003). It is unreasonable to expect a single study design to be applicable to all situations, research questions, or lines of enquiry. For instance, it has been argued that it is "often bad science, and sometimes simply impossible" to expect public health interventions to be evaluated in the same way as pharmacological interventions (Grossman & Mackenzie, 2005, p. 517). Nevertheless, researchers in various fields often "consider that RCTs are *always* superior to *all* other types of evidence" as it is "the most rigorous and scientific study design available" (Grossman & Mackenzie, 2005, pp. 516–517).

As Meltzer and colleagues (2021) highlight, an RCT designed to answer the overarching question in the field – if specific intervention programs like language learning can provide a similar protective effect again dementia and neurodegenerative disease as lifelong bilingualism – would need to span several decades, the cost, time, and complexity of which are not viable in the current grant/funding system and larger academic structure. Regardless, a key issue in RCTs studying language learning in the third age is blinding – unlike a drug trial where a placebo pill can be easily used, it is impossible to conceal if

participants have been allocated to a language learning condition or an active or passive control. Participants will of course quickly realise if they are learning a language. The aims of the study (e.g., language learning as the experimental condition of interest) may be concealed from participants, but blinding participants, researchers, or teachers proves more difficult. At best, as in Meltzer et al. (2021) and Berggren et al. (2020), participants and researchers can be blind to group assignment during the pre-testing phase, but once randomization occurs and/or the intervention begins this is not possible. If funding allowed and the research teams were large enough, theoretically a different researcher could conduct post-testing phases and therefore be blind to participant assignment. However, double blinding remains unattainable.

Furthermore, even RCTs evaluating the short-term impact of language learning inventions are not necessarily ideal for an older population who may experience age-related decline or health issues during course of the trial. For example, in Wong et al.'s (2019) RCT, twenty-five percent of dropouts were due to major health-related issues preventing participation, such as a heart attack or stroke. Similarly, in Kleisch and colleagues (2021) recent controlled trial spanning eight months, four participants unfortunately suffered from a stroke during the training period. While in this case four participants of 65 is not a substantial portion, in other situations even four participants could have negative implications for any matching or stratification across groups – such as if those four participants happened to be randomised into the same group. Likewise, randomization itself can prove to be disincentivising for participants, particularly those allocated to passive control groups. For example, the passive control group in Melzter and colleague's (2021) RCT experienced the highest rate of withdrawal. In particular for longer studies using passive control groups who are living life as normal, it can be an insurmountable challenge without an

excess of funds to incentivize participants to return to a second or third testing session. Power and sample size are crucial in any study, though can provide additional difficulties in the face of dropouts and attrition, in particular in RCTs which are resource-intensive by design.

Aside from these fundamental hurdles inherent in the applicability of the design, we argue that RCTs in the present third age language learning field are akin to testing the efficacy of a drug when it is not clear what is the best dosage, frequency of administration, or route of administration. In this simile, dosage corresponds to intensity of language learning classes, frequency of administration to the frequency of classes, and route of administration can be thought of as delivery method (e.g., online, in person, or blended), teaching style (e.g., explicit or implicit instruction), and learning dynamics (e.g., individual learning or group learning). As seen in Table 1.2, these variables are highly heterogenous in the current literature. In accordance with previous researchers (Pot et al., 2019; van der Ploeg et al., 2020; C. Ware et al., 2021), we also suggest that one of the driving factors influencing the inconclusive findings in the field are related to methodological issues and study design. However, we argue one step further that, less time should be dedicated towards testing hypotheses around the influences of third age language learning and more time laying the foundation to develop the framework of the intervention itself, leading to hypotheses which can later be tested. Indeed, methodological homogeneity cannot be expected when crucially relevant variables have not yet been defined.

Scheel et al. (2021) elegantly outline that much of the replication crisis in psychology is driven by an often blind and unconscious adoption of the hypothetico-deductive (HD) stance on science which was further adapted by Popper's (1959) critical rationalism. HD is "the philosophy of science that

focuses on designing tests aimed at falsifying the deductive implications of a hypothesis" (Fidler et al., 2018, p. 238). Popper's (1959) critical rationalist modification to HD further heightened the role of falsification by outlining that empirical data cannot provide truth statements, but theories which endure repeated testing in environments with the potential for high falsification allow for more strongly "corroborated" predictions (Fidler et al., 2018). Indeed, Popperian hypothetico-deductivism is so engrained in research that many field equate it with the scientific method itself (Scheel et al., 2021). Further, this is confounded by the "null ritual" within statistical analyses (Gigerenzer, 2004). Gigerenzer (2004) outlines the "null ritual" as a ritualistic, and therefore often mindless, adoption of Null Hypothesis Significance Testing (NHST) as a three-step process "(1) set up a statistical null hypothesis, but do not specify your own hypothesis nor any alternative hypothesis, (2) use the 5% significance level for rejecting the null and accepting your hypothesis, and (3) always perform this procedure" (p. 587). Under an NHST framework, the null hypothesis (H_0), often that there is no difference between means, can only be rejected. Without sufficient detail to make the alternative hypothesis (H_1) statistically falsifiable, researchers typically default to defining H_1 as the complement of H_0 (i.e., there is a difference) (Morey & Lakens, 2016; Scheel et al., 2021). However, compelling researchers to specify what would sufficiently falsify their hypothesis when lacking the theoretical basis for doing so "can lead to testing against arbitrary values and runs the risk of replacing one mindless ritual with another" (Scheel et al., 2021, p. 3). The focus on RCTs and confirmatory research without solid groundwork ensuring a strong link between the test and tested theory (i.e., establishing a strong "derivation chain") (Scheel et al., 2021) is equivalent to jumping out of a plane without a parachute or checking the plane has jet fuel in the first place.

Meehl (1990) coined the term "derivation chain" to refer to this linking concept between the test and tested theory, the union of theoretical and auxiliary principles that are necessary for observable outcomes to be predicted. There are five links for a strong derivation chain, all of which are necessary before informed hypotheses can be tested: "(a) concept formation, (b) developing measures, (c) establishing relationships between concepts, (d) specifying boundary conditions and auxiliary assumptions, and (e) deriving statistical predictions" (Dubin, 1969; Scheel et al., 2021, p. 3). Within the third age language learning field, clear concepts have been formed, though as seen by the present review, there is a scarcity of consistently developed and operationalized measures both in terms of cognitive outcomes and the intervention itself. Without a consensus around how measures should be defined and operationalized, it is unlikely for a clear picture to emerge in regards to the cognitive effects of third age language learning. Similarly, further down the derivation chain, boundary conditions (i.e., within what operationalized space the theory applies) also remain unclear. It is necessary to formalize the crucially relevant variables of language learning interventions (i.e., boundary conditions) such as frequency, intensity, duration, and delivery method before any meaningful evaluation of evidence for or against the theory can be made.

The variety and heterogeneity at every level of the intervention design itself (see Table 1.2, Table 1.3)highlights the fact that the field is still working through the derivation chain, that is in a hypothesis development stage. It is not known what the appropriate dosage is, therefore, all studies consistently using the same frequency and duration would be futile. This heterogeneity is in fact a healthy sign of the field as different groups are testing different ideas and lines of enquiry. While heterogeneity in a field, often in systematic reviews

in particular, can be viewed in a negative light and even argued to be due to lack of researcher discipline, we suggest instead that it reflects the current research phase – hypothesis generation. This exploration is positive and something to be continued in the future as methodological standards are developed. However, when evaluating the field, the current heterogeneity needs to be taken into account.

1.5.2 Determining the parameters instead of asking yes/no questions

A common dichotomy in science is confirmatory versus non-confirmatory (AKA exploratory) research. Although both are important, and complementary, strands of scientific research, the nonconfirmatory research (e.g., research that does not fall within the Popperian hypothetic-deductivist framework) tends to be regarded as a "second-class citizen" (Klahr & Simon, 1999, p. 526). However, non-confirmatory research plays the imperative role of strengthening elements of the derivation chain (Scheel et al., 2021). It forms the necessary basis and a pre-requisite to establish the crucial parameters, on which confirmatory research can be developed.

The current interest in the cognitive effects of language learning and their potential clinical application in an aging population can easily lead to the temptation of designing confirmatory studies (indeed, even RCT's) which should determine, once and for all, whether cognitive benefits of language learning are real or not. However, as has been argued above, the research on the cognitive effects of language learning is not yet at a stage in which the use of RCT's could be meaningful. Without an adequate knowledge of the decisive parameters, methods etc, the results of such research would be hostage to fortune. If a study evaluating a language course of let's say three weeks, with a

once weekly group session of one hour does not produce any measurable effects, how can we decide whether it was due to the intensity, frequency, length, or indeed the teaching method.

Rather than trying to solve an equation with too many unknown variables, a better way to advance our knowledge might be through a systematic, controlled manipulation of one variable at a time. Does it make a difference whether language lessons take place once, twice or three times a week? How long does a course need to last until we can detect measurable differences in our variables? Does the class size matter? Or the teaching method? Is the choice of the language to be studied relevant? Studies addressing these questions will not be able to give a definitive answer about cognitive (and psychological) effects of language learning. But they are essential to prepare the ground for later confirmatory studies, which, without such preparation, would be premature.

However, a major obstacle for this type of research is that studies searching for a simple yes/no answer are easier to design, conduct and above all, interpret. Importantly, studies conveying a straightforward yes/no message are more likely to be cited, and hence to influence subsequent research. This can be illustrated by a comparison of two pairs of studies in bilingualism research and their citation numbers, coming from the same research group, using a comparable methodology and addressing the same research questions. Alladi et al. (2013) has a relatively straightforward message, expressed in its title: "Bilingualism delays age of onset of dementia, independently of education and immigration status". A follow-up paper, Alladi et al. (2017), looking at the almost 200 patients with Fronto-temporal Dementia (a type of dementia which showed the strongest bilingualism effect in Alladi et al., 2013 study) comes to a more nuanced conclusion: "Bilingualism delays the

onset of behavioural but not aphasic form of frontotemporal dementia". The findings were supported independently by a disease- as well as symptom-driven analysis. According to Google Scholar on 29 April 2022, the first paper was cited 499 times, the second, 39 times.

A similar phenomenon can be observed with two papers examining cognitive functions in older bilinguals, using the Lothian Birth Cohort 1936. The first study (Bak, Nissan, et al., 2014) showed a positive impact of bilingualism on cognitive ageing independently of childhood intelligence; the second (S. R. Cox et al., 2016) stressed the interaction between the childhood intelligence and cognitive functions. Again, the first paper achieved, on 29 April 2022, according to google scholar 369 citations, the second 73. This means that apart from the comparative ease of the study itself, aiming at simple yes/no messages is likely to lead to a higher impact: another example of incentivising what is arguably the most popular, but not necessarily the most meaningful type of research and penalising more nuanced approaches.

1.5.3 Incorporating qualitative methods

The hierarchical dichotomy of confirmatory versus non-confirmatory research extends into the similarly tiered distinction of qualitative versus quantitative methods, where qualitative methods are synonymous with confirmatory research and qualitative with exploratory. In fact, the preference/bias for quantitative approaches is so prevalent that Guest, MacQueen and Namey (2012) suggest "to gauge what bias a particular journal might have, and if you're still in doubt, lead with the quantitative findings" (p 269).

Within the current academic incentivization system, a mixed methods approach can function to satisfy the bias towards quantitative methods while

still exploring qualitative lines of enquiry. Mixed methods is an approach to research whereby the "the investigator collects and analyses data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or program inquiry" (Tashakkori & Creswell, 2007, p. 4). When used in conjunction, qualitative and quantitative approaches can not only supplement one another, but also produce more robust analysis which takes advantage of the strengths of each approach (Tashakkori et al., 2021). As various designs exist within the overall mixed methods framework (see Tashakkori, Johnson, & Teddlie, (2021) or Pfenninger & Singleton (2019)), it is both flexible and appropriate for the field of third age language learning. Indeed, mixed methods designs allow for predetermined (hypothesis driven) and emerging (nonconfirmatory) research questions to be investigated within a single study. By adopting mixed methods, the derivation chain can be strengthened via nonconfirmatory elements in studies, descriptive and naturalistic observations, and defining boundary conditions which allow for a refined understanding of theory leading to informative hypothesis generation.

Employing mixed methods approaches at both the study design and analysis stages will encourage and allow for researchers to study a more comprehensive understanding of the impact of third age language learning, a future direction also outlined in van der Ploeg et al. (2020). Comparisons between groups (e.g., quantitative) can only provide so much information, individual differences and experiences (e.g., qualitative) in the process of language learning must also be investigated and given precedence. The question of if there is any cognitive impact of language learning is moot if the interventions are not view positively by participants, which can only be investigated in depth by qualitative means. Qualitative methods encourage the

inclusions of more socio-effective measures, which are a promising and important avenue to continue investigating. An increasing number of studies in the field have started investigating subjective variables such as wellbeing, motivation, loneliness, and satisfaction with the experience (Kliesch et al., 2018; Kliesch, Pfenninger, et al., 2021; Klimova et al., 2020; Pfenninger & Polz, 2018; C. Ware et al., 2017). Indeed, one of the modifiable risk factors identified by Livingston et al. (2020) is "infrequent social contact" or loneliness and isolation. The evidence thus far points to language learning as a uniquely meaningful activity which can also provide social stimulation. If the field continues to work through the derivation chain, further research will be able to appropriate test and clarify the hypothesis that language learning in older age acts as a cognitive reserve building activity.

1.6 Conclusions

Research on the cognitive effects of language learning stands currently at crossroads. Converging evidence suggesting positive effects of bi/multilingualism and language learning on cognitive functions in healthy as well as in pathological ageing (including dementia and stroke) makes the topic clinically relevant. Given the growing interest in language learning among the older part of the population (particularly retirees), language learning could be employed as a simple, easily applicable and relatively cheap activity contributing to "healthy ageing" and "cognitive reserve".

On the other hand, however, the empirical evidence is complex and at times contradictory. Not all studies have replicated cognitive benefits of language learning and those who have used a bewildering variety of populations, methods, and outcome measures. The intensity, frequency, length, and method of language learning varies so much between the studies,

that no clear conclusions can be drawn as to what the optimal parameters could be. Simply dichotomising studies into results which are "positive" or "negative" fails to capture the nuanced findings in the field and encourages further investigation of research questions with "yes" or "no" answers. However, there is a trend worth investigating further, where "positive" studies tend to have a higher intensity of language learning experiences, and "negative" studies a longer duration.

However, this situation is not necessarily a product of lacking rigour, insufficient co-ordination or inappropriate design. It reflects the fact that we are still in the early stages of exploration, identifying the relevant variables and determining the optimal parameters. Until this is done, attempting to conduct a study in a Randomised Control Trial (RCT) design would be premature, producing unreliable and potentially misleading results.

Instead, we are proposing a more explorative approach, examining systematically specific variables rather than trying to answer all question at the same time. We stress the importance of linguistic and population diversity, as findings confined to a specific language and society might not be generalizable beyond its confines. Finally, we advocate an integration of qualitative and quantitative methods, ideally in a mixed methods design, in which they can complement each other.

1.7 Where this PhD fits in

This PhD aims to provide new empirical data to contribute to the wider field of language and aging as well as more specifically third age language learning, ideally helping to further the discussion and calibration of key variables outlined above. By adding new data, the aim is to contribute relevant observations, hypotheses, and findings to some of the empirical questions

which remain open, rather than seek to find a consensus to these questions. As outlined above, not only is more data needed, but also a wide range of methodologies and variables to appropriately address these open questions. Therefore, before empirical chapters are presented in Chapters 3 to 7, Chapter 2 will discuss in more detail the methodology adopted in the investigations of this thesis.

Chapter 2 Qualitative and Quantitative Methods in the Study of Multilingualism and Language Learning

As will become clear in the following empirical chapters, the present PhD work includes a wide range of methods and methodological considerations, which are justified, introduced, and explained in this chapter. This methods chapter is thorough as doing justice to the topics requires a combination and synthesis (i.e., mixing) of traditional qualitative and quantitative research approaches. The chapter begins with a discussion of the philosophy of science behind these methods, then the rationale for the methodological decisions within this thesis is provided, moving to more technical issues relevant throughout the presented work at the end.

The overarching aim of the present PhD work, as reflected in the title, was two-fold. To explore the influence of multilingualism across the lifespan via (a) natural histories, meaning studies investigating the lived experience of multilingual people as they age and (b) examine the impact of language learning experiences. The data collected in both parts of the project is quantitative and qualitative in nature. A mixed-methods approach to data collection and analysis was adopted throughout this thesis as it not only provides analytical rigour but also is better suited to answer the nuanced questions under investigation.

2.1 Philosophical introduction to questions of methodology

This chapter does not seek to provide a definitive description of philosophical or theoretical perspectives, but rather to outline the broad categorisations relevant to understanding the core methods and

methodologies considered for the present work. **Ontology** is a branch of philosophy concerned with the nature and structure of reality by studying the concepts of existence, reality, being, and becoming (Coyle, 2016b). **Epistemology**, often discussed in conjunction with ontology, is a branch of philosophy concerned more with *how* knowledge is acquired and *what* knowledge can be acquired. In other words, ontology examines the nature of reality that exists independent of human knowledge (i.e., "truth"), and epistemology examines how that knowledge is gathered (i.e., "knowing"). As such, "all research approaches and methods are based on a set of epistemological assumptions that specify what kinds of things can be discovered by research which uses those approaches and methods" (Coyle, 2016b, p. 11). Therefore, epistemological stances emerge from ontological stances. From this, arises the decision of which method or methodology is best suited for the data and research question(s). In fact, Guba and Lincoln (1994) defined paradigms as representing a worldview meaning "basic belief systems based on ontological, epistemological, and methodological assumptions" and proposed an "order of influence" where ontology informs epistemology which then informs methodology (p. 107). As they are irrevocably intertwined, ontological and epistemological stances constrain the research process within specific logistical elements, particularly in qualitative approaches where researchers are more deliberate in acknowledging their stances (Braun & Clarke, 2021a).

2.1.1 Ontology

As elegantly outlined by Braun and Clarke (2021a), in a simplified understanding, there are three possible answers to the ontological question "is there a reality that exists separately from our research practice?": (a) "well of

course there is", which represents realism, (b) "definitely not, how naïve are you?", which represents relativism, and (c) "um yes but also no", which represents critical realism (p. 167).

Realism posits that reality exists independently, regardless of the unique observer, and therefore can be accessed through research (Braun & Clarke, 2021a; Coyle, 2016b). That is, a realist ontological position assumes that truth/the world can not only be known but is also waiting to be discovered. Further, in a pure or "naïve" stance on realism, the information or truths discovered in research are independent of the tools or methods used, including the researcher themselves (Braun & Clarke, 2021a). Consequently, the language that participant's use provides a window into that reality. Pure realism has been criticised for falling victim what Bhaskar (1997) calls the "epistemic fallacy" where ontology is "reduced" to epistemology (discussed below), meaning representations of reality are confused with reality itself (Pilgrim, 2014).

Critical realism, as per the name, takes a more critical or contextualised view about the world/reality and our relationship to it by advancing that while reality exists independently, there is no way to know this reality with certainty as reality is not the same as representations of reality (Bhaskar, 1997; Braun & Clarke, 2021a; Coyle, 2016b). Therefore, researchers can gather knowledge of the observed world, but it is not possible to know the so-called real world. In some considerations of critical realism, the line of ontology and epistemology is blurred by combining ontological realism with epistemological relativism (Braun & Clarke, 2006). As such, the notion of truth and reality are maintained but human practices and experiences influence and obscure this (Maxwell, 2012). Ontological realism advances the belief that reality, and components of reality, exist independent of perception (Patton, 2015). Epistemological realism,

similarly, holds that knowledge can be gathered by assuming proposals of reality are either true or false.

Relativism is both an ontological and epistemological stance which rejects the idea that there is a single independent reality (Braun & Clarke, 2021a). Realism and critical realism assume a single reality which provides a foundation or bedrock for knowledge that is produced through research. Relativism, on the other hand, is an anti-foundationalism ontology which rejects the very notion of a singular foundation of reality or truth (Braun & Clarke, 2021a; Guba & Lincoln, 1994). Reality, instead, is constructed and dependent upon the unique ways in which individuals have come to know it. Therefore, "truth" is relative: "constructions are not more or less 'true', in any absolute sense, but simply more or less informed and/or sophisticated" (Guba & Lincoln, 1994, p. 111). As such, under relativist ontological accounts, there is no final arbiter of 'truth' but rather 'truth' is the outcome of the analysis (Braun & Clarke, 2021a). The ontologies discussed here can be thought of as a continuum with realism and relativism on the ending anchors and critical realism somewhere in between the two.

2.1.2 Epistemology

Epistemological stances outline what the researchers believe to be true regarding what knowledge is feasible given the state of reality/truth (ontology), and how that knowledge can be gathered (Braun & Clarke, 2021a). The cornerstone of **constructionism** is the view that research practices do not reveal evidence but rather produce it (Willig, 1999). Under a constructionist framework, language does not reflect a separate reality as in (post)positivist (i.e., an objective reality) or contextualist (i.e., perspectival realities) (both discussed further below), but rather creates or "brings realities into being"

(Braun & Clarke, 2021a; Willig, 1999). Therefore, within epistemologically constructionist research, regardless of the particular approach taken, language plays a central analytic role. Ontologically, constructionism fits within relativism as constructionists reject the concept of a foundation for knowledge by which ultimate truth is determined (i.e., is an anti-foundationalist approach). **Social constructionism** is an extension of relativism which holds that the means by which we as human beings understand ourselves and the world are constructed by social processes and human practice, implying that nothing is fixed or certain as each individual has experienced unique social and economic factors (Coyle, 2016b; Potter & Wetherell, 1987). Consequently, each person and their unique understanding of the world is a product of their particular cultural and historical context.

Epistemological **contextualism** stresses the ambiguous and context-dependent nature of meaning, the inescapably political and ideological nature of research, as well as the reliance on theory and interpretation for meaning to be derived from data (Jaeger & Rosnow, 1988; Madill et al., 2000). Knowledge, therefore, is "local, provisional, and situation dependent" (Jaeger & Rosnow, 1988; Madill et al., 2000, p. 9) meaning that results or findings will fluctuate according to the context (including the individuals involved) in which the data was gathered, produced, analysed, and interpreted. For example, four dimensions which effect the both the data itself and the results of an analysis include: "(1) participants' own understandings, (2) researchers' interpretations, (3) cultural meaning systems which inform both participants' and researchers' interpretations, and (4) acts of judging particular interpretations as valid by scientific communities" (Pidgeon & Henwood, 1997, p. 250). Thus, the researcher and participant are not two entities, but rather 'in relationship' to co-produce meaning (Madill et al., 2000). As such, contextualists are

comfortable with the notion of multiple versions of reality where knowledge itself is meaningfully evaluated in the context of utility rather than accuracy (which loses meaning without a single, independent reality) (Madill et al., 2000; Pidgeon & Henwood, 1997). Knowledge is inextricably linked to the knower/participant as well as the researcher themselves, referring to both the researcher's values and practices (Braun & Clarke, 2021a). Ontologically, contextualism fits broadly within critical realism. Similar to the position of critical realism on the ontological continuum as a sort of middle ground, contextualism can be thought of as the falling somewhere between the (post)positivism and constructionism on an epistemological continuum (Henwood & Pidgeon, 1994).

Though ontological and epistemological stances are not often explicitly addressed or discussed by purely quantitative researchers, quantitative methods (e.g., Frequentist statistical inference methods, discussed in Section 2.4.2) adopt a realist ontology and (post)positivist-empiricist and hypothetico-deductive epistemological stances. **Positivism** adopts a straight-forward assumption about the relationship between the world/reality and our perception such that there is a direct correspondence between events, objects, and other phenomena and our perception of them, assuming skewing factors which may damage the correspondence (e.g., a vested interest in what is being perceived) are controlled (Coyle, 2016b). Consequently, a positivist viewpoint hinges on a realist ontology and holds that it is possible to gather accurate knowledge of the world as long as the observers take an impartial, unbiased viewpoint. That is, research has the ability to "converge on the 'true' state of affairs" (Guba & Lincoln, 1994, p. 109). Positivism further adopts with it a dualist and objectivist epistemological stance by assuming the researcher is able to study the topic of interest without influencing or being influenced by it (Guba

& Lincoln, 1994). As such, research and inquiry can be thought of as taking place in a one-way mirror where replicability is synonymous with 'truth'. As Guba and Lincoln (1994) outline, positivism is the "received view that has dominated the formal discourse in the physical and social sciences" for several centuries (p.108). **Post-positivism** developed in response to critiques of positivism and has replaced positivism as the dominant paradigm within scientific research (Braun & Clarke, 2021a; Patton, 2015). As with positivism, the ultimate goal of any research practice is to obtain objective knowledge, though it is recognised that observations of the world cannot be perfect or unbiased (Braun & Clarke, 2021a; Guba & Lincoln, 1994). Post-positivists refine pure positivism by accepting that observations of the world (i.e., data) are inherently selective and we cannot perceive the world or truth in totality. In this way, post-positivism relies on a critical realist ontology and objectivity persists as a "regulatory ideal" (Guba & Lincoln, 1994). Replication of findings are likely to be true, though are subject to falsification. A related epistemological stance is **empiricism**, which posits that knowledge arises from the collection and categorization of our observations and perceptions of the world (Coyle, 2016b). Therefore, the development of increasingly complex knowledge of the world leads to theories that explain it. The fundamental empiricist claim central to research is that knowledge requires collection and analysis of data. This claim is shared by qualitative researchers (e.g., constructionists), however, the definition of data, how it should be generated, and how it should be analysed varies.

One of the most influential thinkers in the theory of science was Karl Popper. Popper's influence on scientific thinking cannot be overstated. Indeed, many of the principles he introduced are so ingratiated into scientific practice that they are assumed to be self-evident. **Hypothetico-Deductivism,** most

famously brought into prominence by Popper (1959, 1969), developed in reaction to the positivist and empiricist suggestion that observations of the world provide a direct representation of the truth or factual state of affairs. In particular, Popper took issue with the empiricist reliance on veritificationism, which links meaningfulness (and therefore the demarcation of science versus non-science) with confirmation (Creath, 2021). Popper's (1959) critical realist approach (i.e., "any claims to knowledge should be open to rational criticism", Afisi, 2013, p. 507) replaced positivist induction with deduction (discussed in more detail in Section 2.2) in the form of the falsification principle, which argues that a scientific theory must be able to be conceivably tested and proven false. Therefore, critical rationalism demarcates science from non-science using the falsification criterion. Popperian hypothetico-deductivism adopts a non-inductive view of "corroboration" as opposed to "confirmation". Godfrey-Smith (2009) uses the example of an academic transcript compared to a letter of recommendation to distinguish Popper's view of "corroboration" and logical empiricist's view of "confirmation":

> "An academic transcript says what you have *done*. It measures your past performance, but it does not contain explicit predictions about what you will do in the future. A letter of recommendation usually says something about what you have done, and it also makes claims about how you are likely to do in the future. Confirmation, as understood by logical empiricists, is something like a letter of recommendation for a scientific theory. Corroboration, for Popper, is only like an academic transcript. And Popper thought that no good reasons could be given for believing that past performance is a reliable guide to the future. So, corroboration is entirely 'backward looking'" (p. 68).

Under the hypothetic-deductive model, the scientific method develops by cultivating falsifiable hypotheses which are then tested on observable data (Godfrey-Smith, 2009). Tests are subsequently run using this observable data and outcomes which are contrary to the predictions of the hypotheses are

considered a falsification of that hypothesis. Therefore, the goal of science under this framework, and consequently the furthering of knowledge, is to attempt to disprove a theory (e.g., error elimination) rather than strive to support theoretical hypotheses (Afisi, 2013).

2.2 Three influential dichotomies

Different methodologies presented above can be conceptualised against three influence dichotomies. The following three dichotomies tend to be used in different contexts, though do seem to reflect or apply to a more general dichotomy. A discussion of the conceptual boundaries of this more general distinction is outwith the scope of this thesis, but merits thorough study and consideration.

The purpose of research in a **qualitative** paradigm is to produce contextualised and situated knowledge, focussing on meaning and meaning making practices (Braun & Clarke, 2021a). In contrast, in a **quantitative** paradigm research functions to test hypotheses, often in the larger search for explanatory models or theory refinement. Not only are the data themselves different (e.g., simply put numbers versus words), but also the approach to analysis. Data analysis in a qualitative paradigm concentrates on text and meaning whereas numbers and relationships between measured variables are the focus of quantitative data analysis. It follows that the purposes of data and subsequent sampling methods also diverge. Data functions as a means to gain rich, nuanced, and in-depth understanding in a qualitative paradigm, which may then be transferred (or analytically generalised) (Polit & Beck, 2010). In contrast, data collected under a quantitative paradigm aims at producing findings which are statistically generalisable, allowing inferences to be made about the population from which the sample was gathered. Small samples are

valued in qualitative research while larger, representative samples are the ideal to allow for statistical generalisability in quantitative research.

A **nomothetic** approach to research, where the aim is to for generalisable findings which uncover laws explaining objective phenomena, has been historically dominant in psychology (Coyle, 2007; J. A. Smith et al., 1995). Historically, psychology has roots in both natural sciences, particularly medicine and within medicine specifically physiology, but was also closely related to the humanities of the 19th century. Indeed, Wilhelm Wundt, the first self-proclaimed psychologist and the father of experimental psychology was himself first a professor of physiology (Holt et al., 2019, p. 58; Martin et al., 2010, p. 24). In the 20th century, however, the natural sciences aspect of the became predominant with the advent of behaviourism, at least in Anglo-Saxon countries. Coming from "nomos" meaning "law" and "thetikos" meaning "thesis" or "proposition" in Greek, nomothetic approaches align with quantitative methods whereby individual participants, often taken as a sample from a larger representative population, are compared to models and theories. Alternatively, **idiographic** approaches seek to examine individual cases in detail in order to understand an outcome, thus, aligning with qualitative methods where the focus is on the unique individual rather than generalizability. In practice, these two approaches to knowledge have often been dichotomized and treated as mutually exclusive in psychology, and social sciences generally, though such a withstanding dichotomy was not intended when first used by German philosopher Wilhelm Windelband in his address as Rector of the University of Strasburg in 1894 (Windelband, 1998). Rather, Windelband posed the two methods as offering integrating points of view, which any branch of science could capitalize on: "The law [nomonthetic] and the event [idiographic] remain to exist alongside one another as the final,

incommensurable forms of our notions about the world" (Windelband, 1998, p. 21). Indeed, Windelband (1998) provides an example of the varied, yet composite, solutions the two approaches might offer to explaining the cause of an explosion:

> "But just as, logically, the conclusion requires those two premises, so also does the event require two kinds of causes: on the one side the timeless necessity, in which the eternal essence of things is expressed; on the other side the particular conditions which surface at a particular moment in time. The cause of an explosion is, according to the one meaning—the nomothetic—in the nature of the explosive material, which we express as chemico-physical laws; according to the other meaning—the idiographic—it is a spark, a disturbance, or something of the sort. Only the two together cause and explain the event, but neither of the two is the consequence of the other; they are not in and of themselves bound to one another" (p. 20).

A fundamental distinction in approaches to reasoning which maps onto nomothetic and idiographic approaches is deductive and inductive reasoning. **Deductive** reasoning, also referred to as 'top down' research starts with theory, moving from the general (theory) to the specific (the data). Deductive research starts with theory, which is then refined and operationalised into hypotheses or predictions. These hypotheses are subsequently tested via experimentation which produces data or observations that are analysed in light of confirming or rejecting the hypotheses. This broadly aligns with nomothetic approaches which aim for generalizable findings as deductive approaches start on a more general theoretical level. Alternatively, **inductive** reasoning, also referred to as 'bottom up' research, starts with the data and moves from the specific (the data) to the general (theory). Beginning with the data, inductive approaches discern and label patterns within it. Some inductive methods then generalise these patterns and link them to existing theory or use them to develop new theories. Inductive approaches generally map onto idiographic approaches

which seek to examine individual cases as they start with the data (i.e., individual cases).

2.3 Overview of qualitative methods

In the following 2 sections, first a discussion of qualitative approaches will be presented, followed by quantitative. This order was chosen as it is in line with the development of many of the studies presented within this thesis, in which findings from certain qualitative aspects of studies have informed and influenced subsequent quantitative investigations.

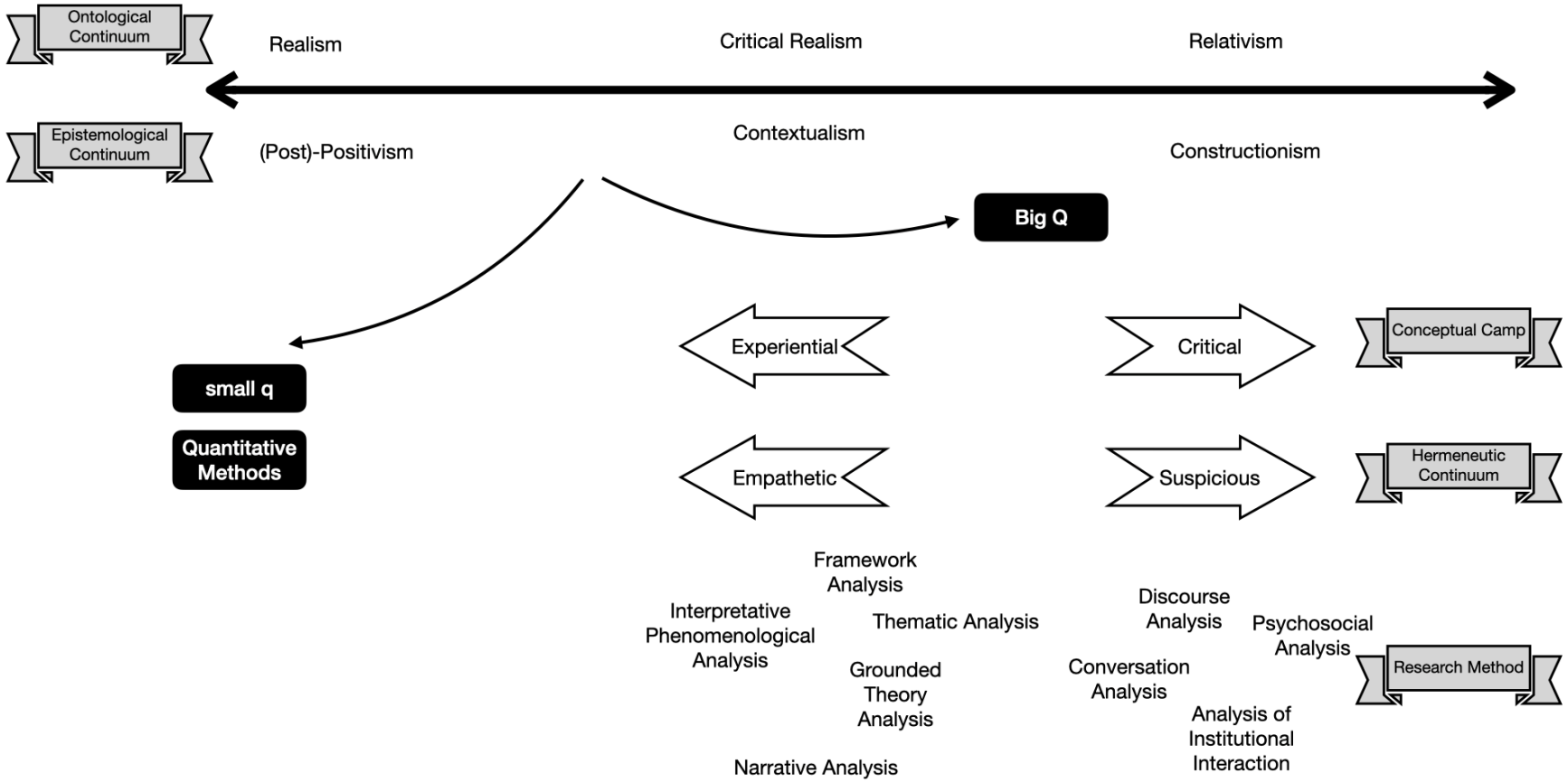
Qualitative research methods and approaches are not homogenous; different approaches are associated with different epistemological and ontological assumptions. As Patton (2015) states "there is no definitive way to categorize the various philosophical and theoretical perspectives that have influenced and that distinguish the types of quantitative inquiry" (p. 85).

Within the qualitative framework, there are two primary categorisations of qualitative paradigms: "small q" and "Big Q". "**small q**" qualitative research employs qualitative tools and techniques (methods) within a hypothetico-deductive framework, adopting realist ontology and (post)positivist epistemology (Braun & Clarke, 2021a; Kidder & Fine, 1987). Given the hypothetic-deductive framework, these approaches concern themselves with traditionally quantitative elements such as quantification of qualitative data (e.g., frequency a word is used), reliability, and generalisability. Researcher subjectivity is viewed as a source of bias, which must be managed (Braun & Clarke, 2021a). On the other hand, "**Big Q**" qualitative research employs qualitative tools and techniques (methods) within qualitative paradigms (i.e., relativist ontology and non-positivist epistemologies) which emphasize contextualised understandings and nuance over and above universal truth or

objective reality (Braun & Clarke, 2013; Coyle, 2016b; Kidder & Fine, 1987). In other words, "small q" paradigms view qualitative research as tools and techniques, while "Big Q" paradigms take a more wholistic stance where qualitative research is not only a source for a variety of research tools and techniques but also philosophies (Braun & Clarke, 2021a). Within this there are two major conceptual approaches or camps in qualitative research into which the five main "Big Q" methods discussed here fall: experiential and critical or discursive (Reicher, 2000). These conceptual orientations influence the type of research questions posed as well as modes of data collection and analysis (Braun & Clarke, 2021a). **Experiential** qualitative research focuses on inductively revealing the participant's experience and meanings or interpretations of their experience (Braun & Clarke, 2013; Coyle, 2016b). That is, conceptually researchers are principally concerned with inferring meaning from the data. The primary analytic aim of the experiential orientation, therefore, is to understand how the phenomenon of interest is understood by the participant(s) (Byrne, 2022). On the other hand, **critical** qualitative research, also referred to by some researchers as discursive, adopts a more critical orientation towards participant discourse, analysing it as integral as opposed to reflective of participant's experiences (Braun & Clarke, 2013; Byrne, 2022). Conceptually, researchers are more concerned with how conversation or discourse is used in social realities as opposed to drawing inferences about how participant feel or think. Ontological relativism and epistemological constructionism (including social constructionism in particular) align with the critical or discursive qualitative approach. This distinction in orientation mirrors the researcher's stance or intentions to reflect participant experience of their social reality (experiential) or to critically examine what constitutes their social reality (critical) (Byrne, 2022).

Similarly distinguishing is the interpretative stance taken in different qualitative analyses, which can be thought of as falling on a hermeneutic continuum from empathetic to suspicious (Steffen, 2016; Willig, 2012). The **empathetic** interpretative stance is grounded in the data and seeks to amplify and elaborate on the meanings contained within the data. The researcher stays within the confounds of the meanings presented in the data, focusing on what is manifest rather than hidden (Willig, 2012). Alternatively, **suspicious** interpretative stances, often theory-driven, aim to "get to the truth of the matter" by revealing latent or hidden meanings in the data which are not immediately obvious (Willig, 2012, p. 11). As opposed to empathetic interpretations, suspicious interpretations often aim to identify a causal mechanism of some form underpinning the phenomenon of interest. Empathetic interpretations amplify meaning rather than de-code or translate it from one concept to another (Willig, 2012). The aim of an empathetic interpretative stance is to "complete the picture" whereas suspicious interpretative stances intend to "boil things down to their underlying meanings" (Willig, 2012, p. 13). As Ricoeur (1996) outlines, the ends of this hermeneutic continuum produce different types of knowledge with empathetic stances generating understanding and critical stances yielding explanation. These two approaches are not mutually exclusive, but rather through some level of combination of two can produce greater understanding (Ricoeur, 1996; Willig, 2012). Empathetic interpretative stances tend to map onto experiential approaches to qualitative research and suspicious onto critical (see Figure 2.1).

Figure 2.1 Visualisation of relationships between ontological and epistemological stances, qualitative paradigms, qualitative conceptual camps, hermeneutics, and research methods stances discussed in this chapter



2.3.1 The main approaches to qualitative methodology

Decisions made in terms of qualitative data analysis should be informed, disciplined, and principled (Coffey & Atkinson, 1996; B. Smith, 2016; Sparkes & Smith, 2014). That is to say, analytic decisions, regardless of the nature of analysis being quantitative or qualitative, should not be made solely based on trends or fashion at the moment, frequency of use within the field, or likelihood of publication. Instead, analytic decisions should be made considering the data and its suitability, the research question(s), what resources are available (e.g., time, money, members of a research team, etc.), and the theoretical framework(s) underlying different analytic approaches. Indeed, as Denzin and Lincoln (1994) highlight, researchers undertaking qualitative analysis should aim to be "methodological bricoleurs": "well informed about a range of different analyses, skilled at using them, and adept at adapting analytic tools or inventing new kinds of analysis whilst maintaining an epistemologically and ontologically coherent position" (Sparkes & Smith, 2014, p. 115). Endeavouring to make methodological decisions as such an analytical "bricoleur", five qualitative research methods were considered for the qualitative stream of this PhD, outlined below: Discourse Analysis, Narrative Analysis, Interpretative Phenomenological Analysis, Grounded Theory Analysis, and Thematic Analysis.

2.3.1.1 Discourse Analysis

Discourse Analysis (DA) is founded on epistemological social constructionism and flourished from the so-called "turn to language" in psychology in the 1980s, which viewed language as a social action (Coyle, 2016a). Potter and Wetherell (1987)'s *Discourse and Social Psychology: Beyond Attitudes and Behaviour* provided a cornerstone for the development of discourse analysis within social psychology (though the authors preferred to

term ‘interpretative stances’ over ‘discourse’). Coyle (2016a) puts forward that “any speech or text that critically considers the implications of language use for socially important matters can be seen as doing a basic form of discourse analysis” (p. 161). Due to its social constructionist commitments, DA can be thought of as a research perspective as well as a methodology: researchers undertaking a DA project are inherently investigating how participants use language to construct their social world and perform actions, and what they gain from such constructions.

DA research questions, therefore, focus on rhetoric, construction, ideology, and action (Coyle, 2016a; Potter, 2004). In fact, Potter (2004) outlines four foci of research questions within the discourse framework: (a) “actions and practice in settings”, (b) “fact construction”, (c) “psychology in practice”, and (d) “exploitation, prejudice, and ideology” (p. 614-15). That is, DA focuses on research questions which ask “how”. In terms of data, any discourse, written or spoken, can be transcribed or considered text-based and is therefore suitable for analysis (Walton, 2016).

Given the in-depth focus on language and the social constructionist perspective, which is difficult to align with the post-positivist hypothetico-deductive assumptions of quantitative approaches, this method was not selected for the present research.

2.3.1.2 Narrative Analysis

Narrative Analysis (NA) encompasses a group of methods which focus on stories, adopting ontological relativism and epistemological constructionism (Riessman, 2008; B. Smith, 2013, 2016). Therefore, NA accompanies research questions which ask “what” and “how”. Within this framework, “narrative” and “stories” take on different meanings than used in

colloquial speech. A "story" is a specific tale that people tell while a "narrative" is a resource "that provides people with a template – a scaffolding of sorts – from which to build and construct their own stories as well as understand the stories they hear or see in action" (B. Smith, 2016, p. 204). Therefore, a narrative itself is not a story, though stories together can form types of narratives (Frank, 2010). The data within this framework are the stories themselves. As a researcher under this paradigm, there are two roles that be undertaken: story analyst or storyteller. As a story analyst, the researcher puts the "narratives under analysis", meaning the narratives are objects of study, and produces an analytic account of the narratives under study (B. Smith, 2016). Alternatively, as a storyteller the "analysis *is* the story and the story is communicated in the form of a creative analytical practice to produce a tale *as* a story" (B. Smith, 2016, p. 209). Namely, the researcher as a storyteller retells the story rather than writing about them analytically. Notably, NA as a research method is not prescriptive, meaning it does not carry a linear, prescribed, step-by-step procedure, implying the analysis may take various forms (B. Smith, 2016). That is, Narrative Analysis is a methodology or heuristic guide rather than a method.

As the story is the focus of Narrative Analysis, it was not deemed the most appropriate analytic approach for the range of qualitative data in the present research (e.g., free response questions, focus groups, etc.).

2.3.1.3 Interpretative Phenomenological Analysis

Ontologically, Interpretative Phenomenological Analysis (IPA) is based in critical realism and epistemologically it is underpinned by contextualism (J. A. Smith & Eatough, 2016). Thus, the aim of IPA is to examine the lived experiences of participants (i.e., phenomenology) and how they make sense of their personal and social worlds. IPA provides a theoretically informed

framework, and thus, can be considered a methodology rather than a method (i.e., technique for collecting and analysing data). Theoretically speaking, IPA commits for the individual participant as a linguistic, cognitive, physical, and effective entity and assumes a connection, though nuanced, between speech, thought, and emotional states (J. A. Smith & Eatough, 2016; J. A. Smith & Osborn, 2015). As such, the researcher plays an active role in the research and analysis process, which is conceptualised as a double hermeneutic (i.e., two-stage process of interpretation). "The participants are trying to make sense of their world; the researcher is trying to make sense of the participants trying to make sense of their world" (J. A. Smith & Osborn, 2015, p. 53). In this way, IPA includes both empathetic and questioning or critical hermeneutics though beginning with an empathetic stance and taking a critical view in later stages of analysis (Ricoeur, 1996; J. A. Smith & Eatough, 2016). Indeed, the research process in an IPA study is best considered as an "interpretative cycle" exploring transcripts in a detailed case-by-case analysis (J. A. Smith & Osborn, 2015, p. 29). The researcher begins at "home-base" by exploring the topic of investigation, enters the world of the participant during the interview process as a "naïve listener", and then returns to "home-base" to formally analyse and interpret the participants' data (J. A. Smith & Eatough, 2016; J. A. Smith & Osborn, 2015)

In terms of design, IPA is best suited for small sample sizes, ideally single-case studies. In fact, for an undergraduate study, Smith and Osborn (2015) recommend a sample size of three. For a professional doctorate, between six to eight participants is appropriate (J. A. Smith & Eatough, 2016). Typically, data for an IPA study would be collected through semi-structured interviews, following either a cross-sectional or longitudinal design (L. Storey, 2016). Research questions often focus around an in-depth exploration of

identity and big questions such as "'hot cognition', engaging with issues that are current, emotive, and sometimes dilemmatic" (J. A. Smith & Eatough, 2016, p. 53).

Given the method is most suitable for in-depth analysis of small sample sizes and detailed examination of the personal and social experiences of individual participants, IPA was not selected for the present research.

2.3.1.4 Grounded Theory Analysis

Grounded Theory (GT) is one of the most popular methods in qualitative research, across a wide variety of fields including social work, psychology, nursing, and education (Charmaz, 2015; Payne, 2016). GT was founded by Glaser and Strauss (1965, 1967) with the publication of *Awareness of Dying* followed two years later by *The Discovery of Grounded Theory: Strategies for Qualitative Research*. As evidenced by the 1965 title, GT grew out of research exploring the experience of terminally ill patients in American hospitals in the 1950s and 1960s. Within the larger academic context of the time which was dominated by the hypothetic-deductive approach, Glaser and Strauss (1967) put forth a radical proposal of "discovering" or developing inductive theory from the close reading and examination of qualitative data. Further, they aimed to exhibit qualitative methods as encompassing explanatory and analytic power in response to the view of qualitative research as only descriptive (Bartlett & Payne, 1997; Payne, 2016). Strauss came from the "Chicago school" of sociology, which "saw human behaviour as determined by social structures and physical environmental factors rather than by genetic and personal characteristics" (Payne, 2016, p. 120). As such, it is not surprising that GT is widely believed to have been derived from symbolic interactionism and pragmatism (Bartlett & Payne, 1997). Symbolic interactionalism posits that

individuals order and make sense of their realities according to shared meanings developed through social interactions and expressed via the reflexive use of symbols (Bartlett & Payne, 1997). Therefore, GT "provides a means for eliciting these meanings and for describing the psychosocial and social processes that have been developed to assist people makes sense of their world" (Morse, 1992, p. 257).

Following the publication of their seminal work in the 1960s, Glaser and Strauss spearheaded divergent approaches to GT after a falling out between the two, so-called Glaserian and Straussian methods (Bartlett & Payne, 1997; Charmaz, 2015; Glaser, 1992; Payne, 2016). In their original work, Glaser and Strauss (1967) proposed that inductively derived hypotheses were to be tested empirically using external verification methods, aligning with epistemological positivism (Payne, 2016). Later versions of GT underscored more interpretative stances. Indeed, Henwood and Pidgeon (1994) called attention to the tensions between epistemological realism and constructivism in GT literature at the time. GT literature emphasising inductively reflecting participant accounts aligns with realism where the literature occupied with recognising the researcher as creating and interpreting data to generate new theory falls within constructivism.

Straussian GT (e.g., Strauss & Corbin, 1990) more explicitly outlines the analytic procedures for a GT analysis compared to Glaser and Strauss's (1967) original description of the approach (for a detailed outline of the different approaches see Payne, 2016). Glaser (1992) argued that Strauss's approach to GT was too prescriptive, with the primary areas of deviation being between the role of induction in the analytic process, the level to which theory 'emerges' from the data, and subsequent variations in procedure (Payne, 2016). This is evidenced not only by the title of Glaser's (1992) independent work: *Basics of*

Grounded Theory Analysis: Emergence vs. Forcing but also by the fact that he suggested the two approaches be referred to by different names. Numerous other researchers have since contributed GT as a method, resulting in various other varieties as well such as the social constructivist account (Charmaz, 1990, 2015), methodical hermeneutics account (Rennie, 2000), and postmodern account of GT (A. E. Clarke, 2003). Given this, it is not surprising that there is debate among GT researchers about what constitutes a "correct" grounded theory analysis.

As a method, GT is an iterative, recursive, comparative, and interactive methodology, guiding the means by which the researcher should collect, analyse, and report data and resulting theories. There are numerous distinguishing characteristics of GT, as outlined by Charmaz (2015):

- "collecting and analysing data simultaneously
- developing analytic codes and categories from the data, not from preconceived hypotheses
- constructing middle-range theories to understand and explain behaviour and processes
- memo-writing – that is, analytic notes to explicate and fill out categories
- making comparisons between data and data, data and concept, and concept and concept
- theoretical sampling – that is, sampling for theory construction to check and refine conceptual categories, not for representativeness of a given population
- delaying the literature review until after forming the analysis" (p. 54).

In terms of design, GT studies are best suited for research questions which are open-ended and exploratory. A fundamental element unique to this methodology is the use of purposeful samples within theoretical sampling as opposed to convenience sampling with the larger aim of representative groups (Charmaz, 2015; Payne, 2016). Furthermore, unlike linear experimental designs common in empirical quantitative methods, a grounded theory analysis views data collection and data analysis as overlapping, concurrent activities.

Grounded Theory was not selected for the present research as integrating reciprocal periods of data collection, data analysis, reflection, theory development and theory testing of GT within the empirical quantitative streams of the present research was deemed difficult, and in some cases impossible.

2.3.1.5 Thematic Analysis

While IPA and GT were founded by specific researchers, Thematic Analysis (TA) does not have as clear cut of an origin. There are in fact numerous different versions of TA; however, the term TA refers to a qualitative analysis which identifies patterns of meaning across a dataset via coding and theme development in order to answer a research question(s) (Braun & Clarke, 2021a; Clarke & Braun, 2016). Indeed, Fugard and Potts (2020) describe TA as "a family of qualitative social research methods that formalize, to varying degrees, the process of developing themes" (p. 2).

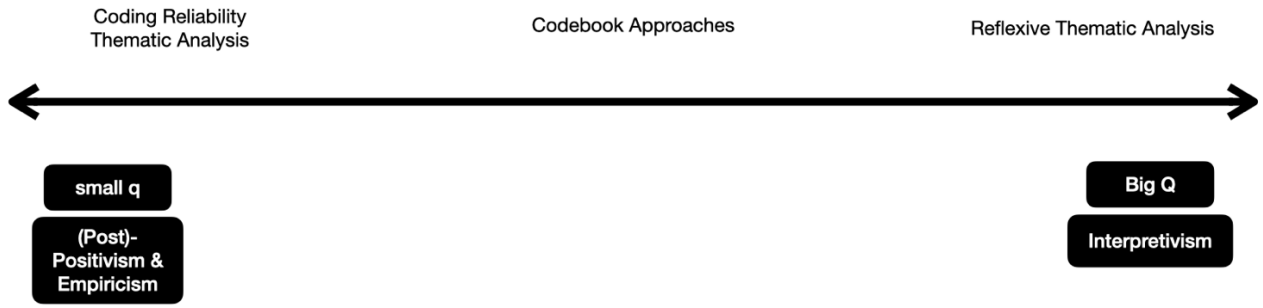
In the 1960s and 1970s when systematic processes for qualitative data analyses began being widely developed, there were already a variety of TA approaches used in different fields (Clarke & Braun, 2016). Compounding the complexities in identifying an origin of TA is the fact that many researchers used TA interchangeably with "content analysis" generally (e.g., Christ, 1970)

while others used TA to refer specifically to interpretative, qualitative content analysis (e.g., Woodrum, 1984). Adding to the confusion are methods which share the term "thematic", such as "thematic coding" (Gibbs, 2007; Rivas, 2018). Thematic coding can be conceptualized as a version of TA, though in practice falls somewhere between GT and TA as it adopts GT techniques to develop themes from data rather than the concepts and categories typical of a GT analysis. Furthermore, some researchers have, and indeed continue to, explain their analysis using TA by simply claiming that themes "emerged" from the data without any discussion to analytic procedure before or after this "emergence" or reference to methodological literature. The concept of "emergence" is often an indirect way of referring to the idea of themes being generated inductively as opposed to deductively. While such language is used in IPA and some versions of GT, it is not appropriate for TA, specifically "Big Q" TA. Braun and Clarke (2006, 2013, 2021a, 2021b, 2021c) in particular have taken issue with this blasé, passive explanation of the methodology: "If only it were possible to collect out data, then leave it somewhere to self-analyse! Analysis requires the active and systematic engagement of the researcher with the data and this is where analytic procedures come in." (Clarke & Braun, 2016, p. 85). While the terms TA began being actively used in academic literature in the 1970s, it was not until the 1990s (e.g., Aronson, 1995) that systematic procedures for conducting TA started being put forward (Braun & Clarke, 2021a; Clarke & Braun, 2016).

Of the qualitative approaches described here, IPA, NA and DA can be thought of as methodologies as they bring with them theoretical underpinnings and bases, a framework of appropriate research questions, and preferred methods for data collection. TA, like GT, is better conceptualised as a method or technique for collecting and analysing data. Both IPA and GT also

aim to identify patterns in data, though TA carries fewer theoretical assumptions, allowing it to be employed within numerous theoretical frameworks (Braun & Clarke, 2006). TA can be applied within the majority of the primary ontological, epistemological, and theoretical frameworks within qualitative approaches (Braun & Clarke, 2021c). Aligning with its theoretical flexibility as a method, there are not any inherent restrictions within TA around the size of data sets: it is suitable for smaller interview-based studies as well as larger qualitative surveys (Braun & Clarke, 2013; Terry, 2016). Similarly, there are not any particular sampling requirements, as in GT, or specified ideal data types. Braun and Clarke (2021a, 2021b)'s tripartite typology for classifying the family of TA methods integrates the varied philosophical underpinnings and accompanying analytic procedures, which can be thought of as existing on a spectrum from "Big Q" to "small q" qualitative paradigms (see Figure 2.2).

Figure 2.2 Visualisation of three primary TA approaches as a continuum of qualitative paradigms



2.3.1.5.1 Coding Reliability Thematic Analysis

Coding reliability TA include those which fit within the "small q" framework. That is, approaches to TA which concern themselves with positivist elements such as reliability and measurement accuracy by employing multiple independent coders and a structured codebook (e.g., Boyatzis, 1998; Guest et al., 2012; Joffe, 2012). Indeed, coding reliability approaches to TA are characterised by the use of a coding frame (or codebook) to facilitate calculating inter-reliability scores between coders (e.g., Cohen's Kappa), which assumes epistemologically positivist and ontologically realist perspectives of a single reality in the data that can be captured. Code books list and define codes and themes, including instructions for the independent raters as to how to quantify and apply them (Boyatzis, 1998). Themes, therefore, are analytically both an input and an output, typically taking the form of topic summaries. In line with the flexibility across different forms of TA, codebooks can be developed either inductively via data familiarisation or deductively based on existing theories or literature (Braun & Clarke, 2021a). Deductive codebook approaches can be conducted with little or no engagement with the dataset itself prior to the coding process, as the codebook is created independently from the data. Alternatively, in inductive codebook approaches the researcher must first become familiarised with some of the data in order to develop the codebook which will then be applied to the whole dataset. In both instances, there is fundamental concern for accuracy in coding driven by the developed codebook. Coding under this approach is viewed as a process of searching for evidence for the themes (Braun & Clarke, 2021b). Indeed, codes themselves are not considered to be analytic units but rather analytic output. Highlighting the blurry line between code and themes under this approach, they are often used interchangeably. Guest and colleagues (2012) define a code as "a textual

description of the semantic boundaries of a theme or component of a theme" (p. 279).

Crucially, the perception of researcher subjectivity as a source of bias to be managed through consensus coding across multiple coders and inter-rater reliability scores make coding reliability TA fundamentally incompatible with Big Q values. Boyatzis (1998) labelled TA as "a translator of those speaking the language of qualitative analysis and those speaking the language of quantitative analysis" (p. vii). While it has been suggested that this approach can function as a "translator" or to "bridge the divide" between quantitative and qualitative research (Boyatzis, 1998; Guest et al., 2012), such a view implicitly diminishes qualitative approaches to a singular, cohesive research framework. As evidenced by this chapter in itself, this is not a view that I personally subscribe to, instead advancing the Big Q perspective of qualitative research as offering both a philosophy and set of tools for research.

2.3.1.5.2 Codebook approaches to Thematic Analysis

Codebook approaches can be considered a "medium q", falling somewhere between coding reliability and reflexive thematic analysis (Braun & Clarke, 2021a). As with coding reliability approaches, the focus of the analytic process is the development of a codebook. Similarly, themes are conceptualised as domain or topic summaries functioning as both analytic inputs and outputs. Where this approach diverges from coding reliability TA, however, is in the recognition and value of researcher subjectivity (Braun & Clarke, 2021b). Therefore, coding reliability is neither encouraged nor viewed as the primary analytic goal. The codebook may be developed inductively or deductively, though functions as a guide or means of mapping the coded data, rather than the analytic result in need of evidence. Also, unlike coding reliability

TA which often adopts the TA name, the codebook cluster of approaches include methods which use other names: matrix analysis (e.g., Nadin & Cassell, 2004), framework analysis (e.g., Ritchie & Spencer, 1994), or template analysis (e.g., King, 1998).

2.3.1.5.3 Reflexive Thematic Analysis

Reflexive Thematic Analysis is a "Big Q" approach to TA, pioneered by Braun and Clarke (2006). As per the name, reflexivity is fundamental to this approach to TA which reflects the value of recognising the researcher as "subjective, situated, aware, and questioning" (Braun & Clarke, 2021a, p. 5). RTA involves a six phrase, recursive analytic process: "(1) data familiarisation and writing familiarisation notes; (2) systematic data coding; (3) generating initial themes from coded and collated data; (4) developing and reviewing themes; (5) refining, defining, and naming themes; and (6) writing up" (Braun & Clarke, 2021c, p. 331). Notably, this process is neither intended to be prescriptive nor to be conflated with the method itself. "The process is not the method. The process applies the method to work with and makes sense of data, but is embedded in, and surrounded by, a bigger set of values, assumptions, practices, which collectively make up the method" (Braun & Clarke, 2021a, p. 7). Indeed, design coherence is an important principle in "Big Q" qualitative approaches, in particular a flexible method like Thematic Analysis (Braun & Clarke, 2021a; Coyle, 2016b). Also called "fit" or "methodological integrity" (Levitt et al., 2017), design coherence refers to the use of harmonious and consistent philosophical and theoretical assumptions (i.e., ontological and epistemological decisions). Along these lines, Table 2.1 outlines the primary sources of variations within RTA.

Table 2.1 Sources of variation in Reflexive Thematic Analysis

Orientation to data	**Inductive**: where the analysis is located within, and coding and theme development are driven by, the data content.	⟷	**Deductive**: where the analysis is shaped by existing theoretical constructs, which provide the "lens" through which to read and code the data and develop themes.
Focus of meaning	**Semantic**: where the analysis explores meaning at the more surface, explicit, or manifest level.	⟷	**Latent**: where the analysis explores meaning at the more underlying or implicit level.
Qualitative conceptual camp	**Experiential**: where the analysis aims to capture and explore people's own perspectives and understandings.	⟷	**Critical**: where the analysis focusses on interrogating and unpacking meaning around the topic or issue.
Theoretical framework (epistemology and ontology)	**Realist, essentialist**: where analysis aims to capture truth and reality, as expressed within the dataset.	⟷	**Relativist, constructionist**: where analysis aims to interrogate and unpack the realities that are expressed within the dataset.

Note: Adapted from "Thematic analysis: a practical guide to understanding and doing" by V. Braun and V. Clarke, 2021, p. 10. Copyright 2021 by Sage Publications, Ltd.

A purely realist orientation in RTA is not compatible with the approaches "Big Q" orientation. A researcher adopting a "naïve" realist orientation would be better suited using a coding-reliability approach to TA rather than RTA. However, the other ontologies discussed in this chapter are

feasible within an RTA analysis. Orienting RTA around a critical realist ontology, for example, allows for a focus on the 'lived experiences' of the participants while also highlighting the cultural and social resources and understandings that underpin their data (Braun & Clarke, 2021a). A critical realist ontology aligns with an experiential conceptual approach as well. Adopting a relativist ontology in RTA necessitates a critical conceptual camp (see Figure 2.2 and Table 2.1) as the analytic task is more subjective, situated, and anti-foundational than a critical realist or experiential orientation. Under such a framing, the researcher's account of the data and analysis is not "true" or "accurate" but rather is convincing or not. The aim is to provide a convincing account of the meanings within the dataset and explain why such meanings matter (Braun & Clarke, 2021a). Similarly, an epistemologically constructionist approach to RTA focuses on language and exploring participant's constructions of reality and subsequent implications.

Both coding reliability and codebook approaches crucially adopt more structured approaches to coding, through the reliance on codebooks, than RTA. Codes under an RTA approach are considered distinct analytic units which form the basis, or act as the building blocks, for themes (Braun & Clarke, 2021a). Themes are analytic outputs built from codes, which cannot be identified or demarcated ahead of the analytic process. Crucially, themes are not topic or domain summaries, but rather patterns rooted by a shared idea which are actively produced by the researcher (i.e., not passively "emerging" from the data) (Braun & Clarke, 2021a).

Given the theoretical, epistemological, and ontological flexibility of TA as well as the fact that is a method rather than a methodology, it was selected as the primary qualitative analysis for the current research. Further, due to the variety of data types and research questions, this flexibility not only allows for

a mixed methods approach, but also for varied framing suited for each individual study. In particular, RTA aligns with not only the theoretical perspectives of qualitative research as "Big Q" but also the aim of this PhD to give balanced weighting to the quantitative and qualitative designs and data. Furthermore, due the fact that RTA does not concern itself with coding reliability, it does not require a team of researchers (or more than one coder) as in coding reliability and some codebook approaches. The suitability of this approach for a single researcher taken together with its flexibility and "Big Q" orientation made RTA the best choice for analysing the qualitative data in the present research.

2.4 Overview of quantitative methods

The following discussion of quantitative approaches is neither intended to be exhaustive nor to discuss in depth the mathematics behind the techniques mentioned, but rather to contextualise popular methods and inference criteria commonly used within psychology and the wider sciences and to explain and justify the choice of the quantitative methods used in this thesis.

2.4.1 Philosophical foundations of quantitative methods currently in use

(Post)-positivism and Popper's hypothetico-deductivism form the ontological and epistemological foundation of statistical inference. The hypothetico-deductivist model, as per the name, contrasts with inductive approaches most prevalent in qualitative methods. Generally speaking, qualitative research is more actively aware and engaged with the

epistemological and ontological assumptions of their methods than quantitative research. Indeed, a fundamental aspect of robust qualitative research is reflexivity and actively outlining the epistemological and ontological stances and assumptions being made (Braun & Clarke, 2021a; Patton, 2015). Reflexivity describes the process of the researcher actively and critically evaluating their role in the research process and decisions made in research practice (e.g., what, how, and why research is done) (Braun & Clarke, 2021a). Alternatively, many researchers employ statistical approaches without an active reflection on the conceptual origin or theoretical background to the tools they are using.

The prescriptive, Popperian philosophy of science that has dominated quantitative research, however, is not without issue. The infamous replication crisis in psychology in particular can be linked to the hypothetico-deductive method and deduction generally taken to the extreme (Derksen, 2019; Scheel et al., 2021). Indeed, Scheel and colleagues (2021) outline that psychological research has developed into a field that largely consists of hypothesis testing, specifically significance tests (e.g., null hypothesis significance testing, discussed below), however, psychologists often lack sufficient background knowledge in order to effectively develop hypotheses.

2.4.2 Frequentist statistics

The dominant method of statistical inference in the 20th century as a tool for empirical knowledge in science is frequentist statistics. Notably, the modern NHST (Null Hypothesis Significance Testing) approach to frequentist statistics, which is ubiquitous standard practice today, is in fact a hybrid of the ideas and practices of different schools of thought: Fisher's significant testing and resulting *p*-value and Neyman-Pearson hypothesis testing (Biau et al.,

2010; Fisher, 1925; Neyman & Pearson, 1933). Of note, while in practice it is frequently the case, the term null hypothesis does not necessitate a hypothesis that the value of interest equals 0. Rather, it implies that the hypothesis can be nullified or refuted (Gigerenzer, 2004).

Both Fisher and Neyman-Pearson developed and popularised their approaches in the early 20th century. Under Fisher's (1925) approach, the *p*-value quantifies the strength of evidence (from the observed data) against the null hypothesis in a descriptive measure ranging from 0 (completely incompatible with the null hypothesis) to 1 (perfectly compatible with the null hypothesis) (Greenland et al., 2016). It is up to the researcher to interpret the *p*-value with "statistical thoughtfulness" (Fisher, 1925). Under the Neyman-Pearson's (1933) theory of hypothesis testing, there are two hypotheses: the null hypothesis and the alternative hypothesis, and therefore two types of errors. The researcher can never be certain of the true outcome, regardless of the result of a specific experiment. As such, the aim is to limit the risk of what they call Type I error (i.e., false positive; denoted α) and Type II error (i.e., false negative; denoted *β*) in the long run (i.e., over numerous experiments). To appropriately use this approach, a precise alternative hypothesis must be specified in advance of the experiment taking place, which is often not done as it is difficult to implement (Sterne, 2001).

As proposed by Fisher, the *p*-value functions as an objective aid in assessing the plausibility of a hypothesis, given the observed data. It is up to the researcher to assess the plausibility of the hypotheses under investigation and draw conclusions. Neyman and Pearson took issue with the subjectivity in Fisher's *p*-value interpretation and sole focus on Type I error, instead proposing that both Type I and Type II error rates should be minimised through hypothesis testing. As such, under the Neyman-Pearson approach, the decision

rule for result interpretation is decided in advance (typically a Type I error rate of 5%) and the resulting analysis is either a rejection or acceptance of the null hypothesis. Fisher's approach can be thought of as testing for (subjectively defined) significance, while Neyman-Pearson's approach is focussed on accepting (and therefore also rejecting) the null hypothesis. Notably, Fisher strongly disagreed with Neyman and Pearson's approach (Fisher, 1956), although the implementation of frequentist statistics today (NHST) combines them, often without the conscious awareness of many of the researchers using these tools. Indeed, NHST frequentist statistics is the incongruent combination of these approaches by adopting Fisher's significant testing (and *p*-values) within a Neyman-Pearson hypothesis testing framework.

As such, it is not surprising that under the NHST framework *p*-values have been consistently misinterpreted and misused as providing evidence for a theory, when in fact they provide the probability of the observed data, or more extreme data, assuming the null hypothesis is true. Lakens (2022) outlines common misunderstandings and misinterpretations of p-values, including that (a) "a non-significant *p*-value means the null hypothesis is true", (b) "a significant *p*-value means the null hypothesis is false", and (c) "a significant *p*-value means that a practically important effect has been discovered". The inference criteria of a *p*-value must be set up the researcher in advance (i.e., there is not a universal, meaningful rule), though often *p*-values under 0.05 are considered "significant". Fisher proposed "If p is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05" (Fisher, 1950, p. 80). This means that values under 0.05 are considered "statistically significant". When under this threshold in NHST, the null hypothesis is then

rejected, and the alternative hypothesis accepted (which is a Neyman-Pearson principle applied to Fisher's *p*-value). Confusing the matter and giving a deceptive impression that the Neyman-Pearson approach is similar or the same to Fisher's, is the fact that Neyman-Pearson also outlined that the null hypothesis can be rejected if $p < 0.05$ when the alpha level (Type I error) rate is set to 5% before the experiment.

When employing a Neyman-Pearson hypothesis test, the specific value of the observed *p*-value is only used to determine if it is smaller than the chosen alpha level, how much smaller it is neither important nor meaningful (Lakens, 2022). However, the smaller the observed value of a Fisherian *p*-value, the more reluctant the research should be in accepting the null hypothesis. In practice, the use of frequentist statistics has been dominated by a dichotomisation of results as significant or non-significant based on *p*-values, with minimal consideration of the Type II error rate. The Neyman-Pearson approach to hypothesis testing (i.e., accepting or rejecting) is effectively what Popper (1959) describes as a basic statement or observation of the world, based on a prespecified decision rule. This statement is not about the theory used to make predictions, but rather the data that has been observed (Lakens, 2022). In fact, many of the suggested means of avoiding the "null ritual" (Gigerenzer, 2004) and introducing more rigour into conscious statistical analysis include Neyman-Pearson framework concepts such as power analyses and reporting confidence intervals.

2.4.3 Bayesian statistics

The predominance of frequentist statistics has been recently challenged by the rising popularity of Bayesian statistics. Bayesian statistics' name sake and basis is Bayes' theorem, which describes the probability of an event

occurring based on previous knowledge (priors) of the conditions of the event (Bayes, 1763). While Bayesian inference dates back to the 18th century when Thomas Bayes first outlined his theorem, Bayesian computation is a much more recent field. The MCMC (Markov chain Monte Carlo) algorithm was invented in the mid-20th century (Metropolis et al., 1953) and general access to computers with the computational power to run these algorithms was not available until the early 2000s (Franke, 2021). The fundamental distinction between frequentist and Bayesian statistics is in how probability is used. Frequentist approaches take an objective view on probability whereas Bayesian approaches take a subjective view on probability by dealing with beliefs (priors) which can be constantly updated. For a recent discussion on the differences between these two approaches to statistics, see Fornacon-Wood et al. (2022) and for an example of a principled Bayesian workflow in cognitive science, see Schad et al. (2021).

2.4.4 Practical implementation/applications of quantitative statistics

Moving to the practical implementation of these statistical approaches, computers have allowed for increasingly complex frequentist models to be fitted and widely used as well. Often the goal of statistical test in psychological research is to determine if there is a difference in a measured outcome of 2 or more groups. William Sealy Gosset, publishing under the pseudonym "Student", famously developed the *t*-statistic to monitor the quality of stout produced at the Guinness Brewery (Student, 1908). The *t*-test has been a popularly adopted means of comparing two groups since its introduction as the *t*-statistic can be calculated by hand. Analysis of Variance (ANOVA) developed by Fisher (1921, 1925) also compared the differences in means,

though among more than 2 groups (using balanced data where there are an equal number of observations for each level of a factor), generalising the *t*-test beyond 2 sets of means. ANOVA models are a specific implementation of linear regression wherein a continuous outcome is predicted by one or more categorical variables. Within behavioural sciences in particular, ANOVAs are so common historically that much of the terminology used in experimental design in psychology is based on the logic of this method (e.g., within group and between group variables).

Notably, the tests mentioned thus far are parametric, meaning the outcome variable must be continuous, the distribution of which is assumed to be parametric (i.e., normally distributed). In the 1940s and 1950s non-parametric alternatives to *t*-tests (e.g., Mann-Whitney U test in place of an independent samples *t*-test and Wilcoxon sign-ranked test in place of a paired-sample *t*-test) and ANOVAs (i.e., Kruskal-Wallis test) were developed as well and are widely adopted when the assumptions of parametric tests are violated (Kruskal & Wallis, 1952; Mann & Whitney, 1947; Wilcoxon, 1945).

An extension of linear regression which has become increasingly wide-spread due to advances in computing power on personal devices is the linear mixed model, which was adopted as the primary statistical method used throughout this thesis. The linear mixed model (LMM) is referred to by many different names including multilevel models, hierarchical models, or mixed effects models. LMM include multiple levels of effects: fixed effects and random effects. Random variables are so-called as they are theoretically random samples from a larger population, as such should the experiment be repeated, different levels may be used (e.g., different individuals from the population of interest). Alternatively, should the experiment be repeated, the levels of fixed effects would be the same, that is the possible values of the

variables are fixed. When data is clustered (i.e., grouped) such that the individual observations are no longer independent, but instead belong to a grouping structure such as multiple observations from the same participant, there are multiple sources of random variability.

A simple linear regression is unable to capture multiple sources of random variability, whereas LMM can effectively deal with not only repeated measures but also increasingly complex hierarchical structures within the data within the random effects structure. As the studies throughout this thesis are repeated measures, include a wide variety of variables, and clustered (often hierarchical) study designs, LMMs were selected as the best suited quantitative approach. Further, another benefit of LMM is shrinkage (or partial-pooling) which limits over-fitting, where in cluster-level estimates are shrunk towards the average depend on (a) the number of datapoints within clusters and (b) the level of across-cluster variation.

2.5 Mixed Methods

Within experimental psychology, from its advent founding figures such as Wundt whilst gather quantitative information such as physiological responses or reaction times, also asked participants to note down their experiences (so-called introspective observations) (Martin et al., 2010). In its infancy, both quantitative and qualitative data were often employed in psychology. Across the 20th century, quantitative approaches became increasingly dominant in psychology, which may have been influenced by several factors. For example, behaviourism, which is rooted in epistemological empiricism, becoming increasingly popular (Holt et al., 2019; Martin et al., 2010). In conjunction with this, from the 1950s in particular the Popperian

hypothetico-deductive model (as discussed in Section 2.1.2) took a stronghold over the field, which largely remains today.

Although the qualitative and quantitative approaches discussed above are often seen as exclusionary and in opposition and this used in isolation, they can, in fact, be used in conjunction in mixed methods approaches. Mixed methods allow for an investigation of research question adopting both nomothetic and idiographic as well as deductive and inductive approaches by combining (or mixing) qualitative and quantitative approaches. Indeed, the notion of "**critical multiplism**" as a research strategy forms the foundation for a dominant argument in favour of adopting mixed methods designs (Patton, 2015; Shadish, 1993). Critical multiplism acknowledges that any given research method is biased and thus only by thinking critically regarding the strengths and weaknesses of numerous methods can rigorous results be found (Shadish, 1993). Therefore, by adopting a mixed methods framework, individual biases or shortcomings of a single quantitative or qualitative method are minimised. A mixed methods researcher is able to investigate a research question with "an arsenal of methods that have non-overlapping weaknesses in addition to their complementary strengths" (J. Brewer & Hunter, 1993, p. 17). Beyond rigour, some research questions are better suited by mixed methods approaches as specific elements may be best answered by a qualitative or quantitative line of enquiry, as is the case in the present body of research (Tashakkori & Creswell, 2007). In particular, as Creswell and Plano Clark (2017) outline qualitative research questions are open-ended exploring "what" and "how" whereas quantitative research questions aim to address "what" and "why". Further, mixed methods approaches bring flexibility to study designs due to the variety of implementations available (Creswell & Plano Clark, 2011).

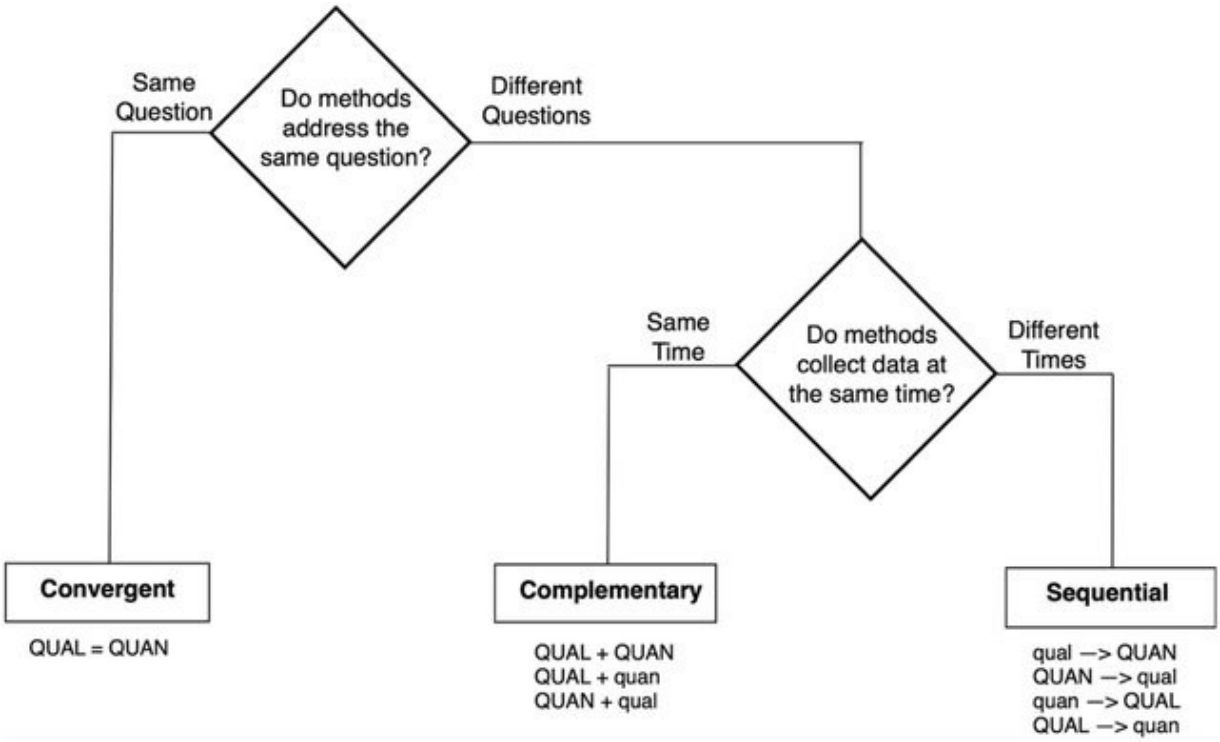
2.5.1 Mixed methods design frameworks

Within the overall mixed methods design framework, various designs exist (Leech & Onwuegbuzie, 2009; D. L. Morgan, 1998). As visualized by Morgan and Hoffman (2021) in Figure 2.3, and first described in Morgan (1998), these numerous designs can be based on two key considerations: whether data are to be collected at the same time and whether the same question is addressed by both methods. Within Morgan (1998; 2021)'s framework, the primary study or dominant method is written in capital letters (e.g., QUANT), whereas those used as supplementary are written in lower-case letters (e.g., quant). The three basic design categories are **convergent designs**, where the same question is answered using both methods; **complementary designs**, where different questions are answered using the different methods within a single study; and **sequential designs**, where one study functions as either an input or follow-up to another. Leech and Onwuegbuzie (2009) outline a theory-based typology of mixed methods, also developing an eight-design framework though differing from Morgan (1998) in distinguishing first a so-called "mixing dimension" which can be either partially or fully mixed. In reviewing the literature in health using mixed methods, Green et al. (1989) refined their conceptual theoretical framework and outlined five purposes for mixed method studies: triangulation, complementarity, development, initiation, and expansion. On the other hand, Bryman (2006) outlined 17 reasons why researchers decided to implement mixed methods after conducting a larger analysis of social science articles (232 studies as opposed to 57 studies in Greene et al., 1989).

In a more recent examination of research designs in mixed methods literature based on citations of Morgan (1998) and therefore from a variety of disciplines, Morgan and Hoffman (2021) found a very similar set of

implemented mixed methods designs to Greene and colleagues (1989). In particular, what Morgan and Hoffman (2021) refer to as "**convergence**" resembles Green et al. (1989)'s label "**triangulation**", where independently conducted studies on the same topic, but using different methods, are compared to see if there are similar results. Within the field of mixed methods research, terms often carry multiple meanings, due in part of the application of a given term by various researchers from different traditions and sub-disciplines. "Triangulation", a term originated by Campbell and colleagues based on navigation and surveying procedures (e.g., Campbell & Fiske, 1959; Webb et al., 1999), and "validity" in particular have been identified as problematic by researchers (e.g., Onwuegbuzie & Johnson, 2006; Teddlie & Tashakkori, 2009) and editors (e.g., Fetters & Molina-Azorin, 2017) alike. Validity, for example, has been found to have at least 35 different meanings within quantitative and qualitative traditions (Teddlie & Tashakkori, 2009).

Figure 2.3 Flowchart for research designs in mixed methods research



Note: Reprinted from "Searching for qualitatively driven mixed methods research: a citation analysis" by D.L. Morgan & K. Hoffman, 2021, *Quality & Quantity, 55*(2), p. 734. Copyright 2020 by Springer Nature B.V.

Following a pragmatic rather than theory driven typology, as visualized in Figure 2.3, there are at least three fundamental purposes for adopting a mixed methods design: convergent findings, complementary coverage, and sequential contributions (D. L. Morgan & Hoffman, 2021). The same research question is answered using both qualitative and quantitative methods under a convergent design. Convergent designs give equal priority to qualitative and quantitative approaches as they are used independent of each other, allowing for more certainty in results as methods with different strengths can produce similar conclusions. In the larger context of mixed methods research, convergent designs are the most well-established stemming from the work of Campbell and colleagues (e.g., Campbell & Fiske, 1959; Webb et al., 1999) and

more widely popularized by Denzin (1970). In contrast, a complementary design utilizes various methods to answer different research questions (e.g., "how" and "why" questions) within a single study whereas sequential design employs one study as an input or follow-up to another. Complementary designs, also referred to as additional coverage (D. L. Morgan, 2014), capitalize on the different strengths of and data produced by qualitative and quantitative methods to address different goals in the larger project. Most commonly, complementary designs are operationalized as parallel strands where the collection and often the analysis of the different data types run separately but concurrently (Leech & Onwuegbuzie, 2009). Interestingly, in a later edition of *The Research Act*, Denzin (1989) moved more towards the complementary design motivation acknowledging that various data types "reveal differences aspects of what is being studied" (p. 245) and that different methods allow researchers to "broaden, thicken, and deepen the interpretive base of any study" (p. 247). Complementary designs promote a more holistic understanding in integrating findings from different methods as opposed to validity-oriented convergent designs.

2.5.1.1 Sequential mixed methods

Within the context of the present body of work, sequential designs best fit the kind of research questions under investigation, particularly in the language learning studies presented in Chapter 4, Chapter 5, and Chapter 6. Sequential mixed methods, or so-called sequential priorities, typically follow a two-step process, through three-step sequential studies are found in the larger literature (e.g., Flynn, 2020). Within sequential mixed methods designs, there are two key considerations or factors – priority and sequence, creating a 2x2 design matrix which results in four possible methodological designs, see Table

2.2. The priority is the principal or core method of the research project (i.e., quantitative or qualitative), denoted in capital letters (e.g., QUAL) whereas the sequence is the supplementary or supporting method (i.e., preliminary/input or follow-up/extension), denoted in lower-case letters (e.g., qual).

Table 2.2 Sequential priorities designs for integrating qualitative and quantitative methods

		Priority	
Sequence		Quantitative	Qualitative
	Preliminary	qual → QUANT Exploratory Sequential	quant → QUAL Sampling Sequential
	Follow-up	QUANT → qual Explanatory Sequential	QUAL → quant Operationalizing Sequential

Note: Adapted from "Integrating qualitative and quantitative methods" by D.L. Morgan, 2014, p. 85-102. Copyright 2014 by Sage Publications, Ltd.

Unlike convergent or complimentary designs where the integration of qualitative and quantitative methods relies on comparison at some level of the results, under sequential designs, what is learned from one method informs how the other method should be used, therefore allowing for more to be accomplished in answering the research question than if either method were used in isolation (D. L. Morgan, 2014). **Exploratory sequential designs** use a preliminary qualitative study to create content for a primary quantitative study, where the latter benefits from the exploratory strengths of the former. **Explanatory sequential designs** use a follow-up qualitative study to understand or explain results from a primary quantitative study. Less common are studies with a qualitative priority (e.g., Sampling Sequential and

Operationalizing Sequential in Table 2.2). In an examination of 228 studies over 20 years, only 18% of mixed methods designs used a qualitative priority (D. L. Morgan & Hoffman, 2021). These findings support the suggestion by some researchers that a quantitative dominance exists in mixed methods research (e.g., Giddings, 2006; Hesse-Biber, 2015). In fact, this preference/bias for quantitative approaches is so prevalent that Guest et al. (2012) suggest "to gauge what bias a particular journal might have, and if you're still in doubt, lead with the quantitative findings" (p. 269). Indeed, in the quantitative versus qualitative debate (or the "paradigm wars"), Onwuegbuzie (2012) argues for the mixed methods paradigm to be "the radical middle":

> "It is not enough for it is not enough for mixed methodology researchers to exist in an epistemological space that lies somewhere between the quantitative and qualitative epistemological spaces. Rather, mixed researchers should strive for what is the radical middle, which should not be a passive and comfortable middle space wherein the status quo among quantitative and qualitative epistemologies is maintained, but rather a new theoretical and methodological space in which a socially just and productive coexistence among all research traditions is actively promoted, and in which mixed research is consciously local, dynamic, interactive, situated, contingent, fluid, strategic, and generative." (p. 192)

Sampling sequential designs use a preliminary quantitative study as input for the primary qualitative study, usually to locate cases for focus groups or interviews (e.g., purposive sampling).

Finally, **operationalizing sequential designs** use a follow-up quantitative study to measure aspects of a primary qualitative study, such as operationalizing a set of survey questions from the hypotheses generated from interviews or applying interventions in other cases from case study results. Crucially, in such a study design, the quantitative follow-up is neither "validating" nor "verifying" the primary qualitative study but rather, in typical

mixed-methods fashion, extending and adding-on to the qualitative study through operationalization of the hypothesis via generating quantitative measures. That is, the quantitative follow-up includes goals outwith the characteristic research questions of stand-alone qualitative study designs (e.g., operationalizing what is learned from focus groups into a series of survey questions). Two common such cases are to (a) increase credibility by demonstrating findings which are converted into tangible measures that preform in predictable ways and (b) provide generality to other similar groups. As Morgan (2014) describes:

> "the issue here is the 'politically charged' relationship between the larger fields of qualitative and quantitative research… It is important to understand that this is not merely a question of vocabulary. Instead, it reflects political, historical, and cultural issues that continue to resonate with members of the qualitative community" (p. 180).

A key design not described explicitly in Figure 2.3 or Table 2.2 is the so-called **embedded design**, typically employed in intervention studies such as randomized controlled trials (Creswell & Plano Clark, 2011). Embedded designs commonly employ qualitative studies conducted, or embedded, in larger quantitative intervention programs, often utilized in terms of a follow-up. According to Morgan's (1998, 2014) typology, embedded designs whereby the qualitative study serves to follow-up the primary quantitative study, could be considered as a sequential design, specifically Exploratory sequential, where temporally the two components of the larger project might occur simultaneously. The pragmatic use or function of the mixed method results outweighs the temporal consideration of when the studies occurred, or the data was collected.

When conducting research, decisions around analysis of data must be made from the initial stages, informing study design and data collection

methods. This is even more relevant when conducting mixed-methods research as generally speaking purely qualitative and quantitative researchers typically do not align in fundamental assumptions around their research practice (e.g., what 'doing analysis' means or acceptable forms of explanation) (Madill et al., 2000). As such, prior to formalizing the study design, epistemological and ontological implications of different qualitative and quantitative approaches should be considered, particularly in terms of integration.

2.6 Data collection methods and measures

Across the present body of research, there were a variety of design and methodological decisions as well as measures relevant to multiple studies, which are outlined below.

2.6.1 Qualitative data collection methods

The primary means of collecting qualitative data throughout the current research was via focus groups, which the secondary being free-response questionnaire items. Focus groups were selected as a research method as they are not linked to a specific theoretical framework (Wilkinson, 2015), aligning with the theoretical flexibility of the selected qualitative analysis, reflexive thematic analysis. Furthermore, focus groups mirror the group setting found in language learning classrooms.

Focus group can be defined generally as "any interview in which the researcher simultaneously gathers data from more than one participant" (J. Green & Thorogood, 2014, p. 126). Typically, focus groups can be characterised by five key features: "(1) a small group of people, who (2) possess certain

characteristics, (3) provide qualitative data (4) in a focused discussion (5) to help understand the topic of interest" (Krueger & Casey, 2015, p. 6). Focus groups as a data collection method date back to the 1920s, though did not gain wider popularity until the 1970s (Wilkinson, 2015). Sociologist Robert Merton and colleagues Patricia Kendall and Marjorie Fiske are often credited with inventing the data collection method in the 1940s (Merton & Kendall, 1946). These 'focused group interviews' were used to gather audience responses to radio programs. Incorporating their experience exploring morale in the US Military during World War II, Merton and colleagues (1956) later outlined procedures for conducting focus groups in their classic work *The Focused Interview.* However, the use of focus groups can be traced back further to the 1920s when psychologists Walter Thurstone and Emory Bogardus used them to create surveys. Prior to the 1970s, focus groups were largely used as a market research tool, as developed by Merton and colleagues (J. Green & Thorogood, 2014). Health researchers began using focus groups for social action research in the 1980s, leading to a wider adoption of the method (Wilkinson, 2015). Within psychology, focus groups have been increasingly employed as a qualitative data collection method since the 1990s and early 2000s.

Prototypically, focus groups include 5 to 8 people, though can range from 2 to 12, after which it is difficult to gather input from all participants or avoid groups fragmentation (Krueger & Casey, 2015; D. L. Morgan, 1998). Through piloting focus group discussion online, it was determined that the ideal group size was a bit smaller than the traditional in-person focus group: around 5 rather than around 8 participants. Participants in a focus group may be sampled from pre-existing group structures (e.g., a language class, members of a team or club) or brought together for the research context

specifically (Wilkinson, 2015). The group discussion is 'focused' on a particular topic and typically based around an interview schedule (i.e., series of questions) (Krueger & Casey, 2015; Wilkinson, 2015). The role of the researcher within the focus group context is to act as a moderator of the group by "posing the questions, keeping the discussions flowing, and encouraging people to participate fully… the moderator does not ask questions of each focus group participant in term, but rather facilitates group discussion, actively encouraging group members to interact with each other" (Wilkinson, 2015, p. 199). This interaction is a key strength of the focus group format which distinguishes it from one-to-one interviews (J. Green & Thorogood, 2014; Krueger & Casey, 2015; D. Morgan, 1997).

The data that is produced from a focus group is the sum of what has been constructed via in the group discussion, including the interactions between participants. For this reason, it is inappropriate to consider focus groups as a collection of individual interviews (Braun & Clarke, 2013; D. L. Morgan, 1998). Indeed, as Krueger and Casey (2015) outline, "the focus group presents a more natural environment than that of an individual interview because participants are influencing and influenced by others – just as they are in real life" (p. 7). The interviewer, thus, can have less of an influence than in one-to-one interviews as the participants themselves assume greater authority of the content and direction of discussion in a focus group. These group dynamics are a key strength of focus groups, though can also be a disadvantage in the case of individual participants dominating the discussion.

Across present work, focus groups were transcribed orthographically/verbatim. Participant names and identifiable referents (e.g., mention of a specific city or location, name of another focus group member, etc.) were anonymised after/during transcription and prior to analysis. The

anonymisation procedure involved looking across all participant names including all studies with qualitative data to ensure that a selected anonymised name did not happen to be a participant's genuine name (even in a different study). Using lists of common baby names in the UK, an anonymised name was then select for each participant. I not only conducted the focus groups, but also this anonymisation procedure, as well as transcription. This meant that any other member of the research team to view the transcriptions did not encounter identifiable information about participants in the focus groups, or indeed in any of the other data collected throughout the studies presented.

2.6.2 Questionnaire design

2.6.2.1 Likert Scales

Likert style scales, developed by the American social psychologist Rensis Likert in the 1930s are ubiquitous in structured survey questionnaire research, particularly in psychology to measure attitudes and opinions (Chyung et al., 2018; Likert, 1932). A Likert scale is comprised of numerous Likert items (i.e., questions). Each item is comprised of a stem (i.e., phrase or statement) and a scale (i.e., response options). Typically, Likert scales use a bipolar, symmetrical layout with verbal descriptors of agreement on one side, a neutral option in the middle, and disagree on the other side (e.g., "Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree") (Chyung et al., 2018; Maeda, 2015). Implementations of Likert scales can vary from this design by (a) not including a neutral option, (b) being unipolar, measuring the strength of a variable in a single direction (e.g., "Never, Sometimes, About half the time, Most of the time, Always"), or (c) including verbal ending anchors and numerical points between (e.g., "Strongly disagree, 2, 3, 4, Strongly agree"). Regardless of the specific design elements, Likert scales are discrete rating scales as they provide the

respondent with a discrete number of options to choose between, typically varying from 2 to 11 (Chyung et al., 2018; Svensson, 2000). Though some argue that under certain conditions Likert scale response data can be pragmatically analysed with parametric tests, the data is inherently ordinal (i.e., comprised of ordered categories).

There are a variety of important design and format decisions when using Likert scales in order to minimise potential response bias. Within the literature, there are 5 key types of response bias:

1. **Net acquiescence**: the tendency for respondents to just 'agree' or 'disagree' regardless of the item's content, affecting the central tendency of the observed data (Maeda, 2015; Weijters et al., 2010)
2. **Extreme response**: the tendency for respondents to disproportionately employ the extreme response categories within the scale, affecting the spread of the observed data (Weijters et al., 2010)
3. **Mis-response to reversed items**: the tendency for respondents to respond similarly (i.e., in the same direction) to items which are opposite in meaning (Weijters et al., 2010). So-called balanced scales containing reversed coded items are often employed to counteract net acquiescence (e.g., "I feel sad" is a reversed item to measure happiness) (Paulhus, 1991).
4. **Left-side response or primacy effect**: the tendency for participants to select response options on the left-side of a horizontally presented Likert scale regardless of the item's content (Barnette, 2000; Chan, 1991; Friedman et al., 1993; Maeda, 2015; Nicholls et al., 2006)

5. **Recency effect**: the tendency for respondents to select response options given last (typically on the right side) or near the bottom of the list (Chan, 1991).

The recency effect is most prevalent in orally presented Likert scales, likely due to limitations of verbal working memory (Chan, 1991). However, when presented in written form, there is less evidence for the recency effect bias (Chan, 1991). Give this, the recency effect as described in the literature is not a primary concern for the present research. In a study using written 5-point Likert style scales, Chan (1991) found the means of the observed data to be significantly higher in descending order (i.e., agreement on the left and disagreement on the right) than ascending (i.e., disagreement on the left and agreement on the right), concluding there is a left-side or primacy effect bias. Friedman et al. (1993) also found evidence for the primacy or left side bias on favourably worded items among a student population. Overall, the content of the stem (that is the item itself) seems to be agreed upon within the literature as the most important factor (Maeda, 2015; Swain et al., 2008). Indeed, Barnette (2000) found that directly worded stems and bipolar scales more effectively combat net acquiescence and left-side response bias than negatively worded or reversed items.

Moreover, in an online study of 1693 adult respondents across the life span (aged 18 to 86), Maeda (2015) found vertical unidirectional response options were most effective when measuring absolute judgements, while horizontal unidirectional response options better suited relative judgements. Additionally, horizontal ascending and descending configurations took the shortest time to complete. While the results of Maeda (2015) provide evidence of a left-side selection bias, the author concluded in agreement with Nicholls

and colleagues (2006) that this is not a major issue if the scales are used to make relative judgments as all individuals will be influenced by this bias. Furthermore, in two online studies of 1207 and 226 adult respondents, Weijters and colleagues (2010) experimentally manipulated the labelling of response categories (i.e., labelling the endpoints only versus all response categories), the number of response categories (i.e., 4, 5, 6, or 7-point scales), and the inclusion of a midpoint. When all response categories were labelled, there was a lower rate of extreme responses as well as mis-response to reversed items, though higher net acquiescence. The inclusions of a midpoint response increased net acquiescence, which was suggested to be due to otherwise negative responses being treated as neutral. Indeed, Gilljam and Granberg (1993) found that when forced to choose a negative or positive response, ambivalent respondents tend to react negatively. Mis-response to reversed items was lowest when the scales included a midpoint and were fully labelled (Weijters et al., 2010). From these findings, Weijters and colleagues (2010) proposed a framework whereby studies, such as in the present research, seeking to measure opinions in non-student populations should use 5-point, fully labelled Likert style scales.

Approaches to eliminate selection bias include randomly assigning half of the participants to complete the study in ascending order and the other half descending (Nicholls et al., 2006), using bidirectional response options (Barnette, 2000), or vertically oriented responses (Maeda, 2015). Bidirectional response options, however, have been shown to have poor internal consistency and take longer for respondents to complete (Maeda, 2015). Randomising the scale response order may address the concern of group means differing but fails to address issues with variant standard deviations or individual scores based on the response order. Additionally, such randomisation creates another variable that must be taken into consideration in analysis. In light of the

presented literature, Likert scales as applied through the current research, outwith pre-existing and validated scales, take the form of fully labelled, ascending 6-point scales with a vertical orientation for absolute judgements and horizontal for relative judgements.

2.6.2.2 Visual Analogue Scales

Originally the question of VAS-style questions was explored when research in care-homes with people living with dementia or MCI was a possibility. Given the population of interest, a paper-and-pencil administration of the research battery was planned, with the researcher verbally administering the battery one-to-one. However, in the adaptation to online research, online VAS questions were included as part of the research battery for various studies to allow for a more fine-grained analysis of opinions/responses.

Visual Analogue Scales (VAS) are a continuous rating scale alternative to Likert-style scales which can be more intuitive and less cognitively demanding for the respondent, particularly if they are experiencing cognitive decline (Arons et al., 2013). VAS are more typically used in medical settings for this reason, such as asking MCI or dementia patients to rate their experience of pain or to measure health-related quality of life. In general, VAS is frequently used as a self-assessment tool to measure subjective symptoms and experiences (Chyung et al., 2018). In a study investigating patient satisfaction upon exiting the hospital, Voutilainen and colleagues (2016) found that not only were VAS questions faster (28%) for participants to complete than Likert-scales, but they were also less likely to produce ceiling effects. Perhaps most relevant to the present body of research was the additional finding that compared to a 5-point Likert scale, VAS were less vulnerable to age-related bias (Voutilainen et al., 2016). VAS questions prototypically include a

continuous 100mm horizontal line with anchor descriptors such as "no pain" and "worst pain imaginable", where the respondent places a mark on the line reflecting their response (Heller et al., 2016; Wewers & Lowe, 1990). The mark is then measured using a ruler from the left endpoint/anchor. This "graphic rating method" was first described scientifically by Hayes and Patterson (1921) and further characterized two years later by Freyd (1923). Freyd (1923) attributed the creation of the method to Scott Co. Laboratory and the Bureau of Personnel Research at Carnegie Institute of Technology in 1920, describing it to be "the latest development in rating methods and promises to be most popular. Its only original feature is the combination of the methods of rating on a line and by checking descriptive terms" (p. 92). It was not until the 1960s, however, that VAS style questions began consistently appearing in academic literature (Heller et al., 2016; Wewers & Lowe, 1990).

The Graphic Rating Scale (GRS), which is now considered a variation on VAS, has several descriptors along the line whereas VAS have two ending anchors only. The Numeric Rating Scale (NRS) is another variation where a VAS line is calibrated with a range of numbers rather than verbal descriptors (Chyung et al., 2018; Wewers & Lowe, 1990). Within VAS design, there are three fundamental methodical decisions: line length, line orientation, and direction. As mentioned, the prototypical implementation of VAS in 100mm or 10cm, though various sizes ranging from 5cm to 20cm can be seen in the literature (Wewers & Lowe, 1990). Direction refers to where the positive and negative labels appear on the scale: left/top or right/bottom. Liu (2017) found that the direction of slider style questions did not significantly impact mean scores. In terms of line orientation, either a horizontal or vertical (AKA "thermometer") orientation are commonly used. Sriwatanakul et al. (1983) found that participants significantly preferred a horizonal orientation as opposed to a

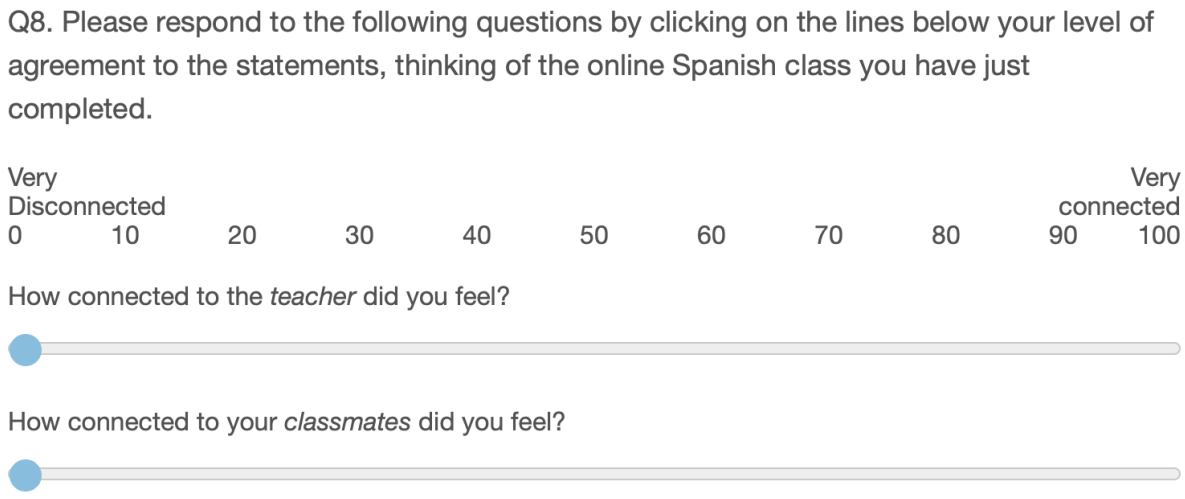
vertical one, further concluding that graded linear horizontal VAS scales were more reliable. For the present research, an in-person pilot study not reported as part of this PhD confirmed the horizontal orientation to be more intuitive, perhaps it is similar to reading. Outside of the benefit of providing a cognitively less demanding measure, VAS also allow for a more fine-grained analysis of responses compared to traditional Likert-style scales. In particular, VAS questions can avoid the issue of shared ranks (Funke, 2016). The primary drawback of VAS relates to paper-and-pencil implementation of the scale where it can be time-consuming for the researcher to measure the respondent's mark and convert it to numeric scores. When employed in an online medium, this concern is obsolete.

In an online application, there are often two styles of response questions: VAS and slider questions. Slider questions "displays a horizontal or vertical line with a bar on the line. Respondents need to click and drag the bar to the desired position on the line in order to register their answers" (M. Liu, 2017, p. 601). Online VAS questions are different only in the fact that they do not include a bar to be dragged and dropped, but rather respondents point and click (Chyung et al., 2018; M. Liu & Conrad, 2019). In many web-based survey platforms, including Qualtrics, slider questions are answered as VAS questions via pointing and clicking (though the question type is named "slider"). In an online administration, there are additional considerations over and above the traditional paper-and-pencil questionnaires. For example, literature suggests that the starting point of the slider can be a source of bias (M. Liu & Conrad, 2019; Roster et al., 2015). Seller (2013) found their data to be strongly biased by the starting point of the sliders with starting at the midpoint resulting in increased midpoint responses and starting at the higher end of the scale producing systematically higher responses. Across four web

based survey experiments, Liu and Conrad (2019) found that for 101-point VAS or slider questions (where the questions ranges from 0-100), the default value of 25, 50, 75, or 100 significantly biased responses, where participants were more likely to choose the default values over any other answers. These findings support Funke (2016)'s proposal that participants may use the starting point of the slider as an anchoring point, leading to biases.

Given these findings, across all reported studies using online VAS questions, the starting point or default value was set to the lower end of the scale (i.e., 0), see Figure 2.4. This required participants to actively express their level of agreement, rather than move the slider to express their level of disagreement. Additionally, research comparing traditional Likert style scales (sometimes referred to as radio button scales) with online VAS suggests that VAS questions may have higher answer precision given participants are more likely to edit their answers after initially responding (Funke & Reips, 2012).

Figure 2.4 Example of VAS Style Question from Chapter 4



The level of measurement and thus how to analyse and treat the outcome data from VAS questions is a point of much debate within the literature (Chyung et al., 2018; Heller et al., 2016; Roster et al., 2015; Svensson, 2000; Wewers & Lowe, 1990). VAS data can be treated as either a continuous scale or discrete n-point scale (Chyung et al., 2018; Funke, 2016; Wewers & Lowe, 1990). When treated as continuous, the resulting data is considered to be an interval-level measurement, as opposed to an ordinal-level measurement which result from Likert scales (Reips & Funke, 2008). Other authors, particularly in clinical applications, however, take issue with the implications of treating VAS measurements as interval data as that means differences along the scale carry the same meaning. For example, considering a pain scale, a difference of 3 would represent the same change in pain at any part of the scale (e.g., 12 to 9 or 3 to 1). Philip (1990) advanced that VAS "lies somewhere between ordinal and interval values… though are most appropriately analysed by parametric techniques" (p. 170). Alternatively, other authors argue for an ordinal interpretation (Svensson, 2000; Wewers & Lowe, 1990). Heller and colleagues (2016) conducted a review of analysis strategies of VAS in the published pain literature (analysing 80 journal articles) as well as a simulation, ultimately arguing for a more nuanced analytic approach depending on the role of covariates. When covariates are not present, the authors put forward distribution-free methods (e.g., Wilcoxon or Mann-Whitney tests) as the most reliable. However, when covariates are present, continuous ordinal regression is the optimal analysis strategy as neither Wilcoxon nor Mann-Whitney tests are able to consider the role of covariates (Heller et al., 2016). Further, if the distribution of VAS responses is not skewed to either extreme of the scale, parametric tests (e.g., *t*-test, linear regression) are as powerful, and therefore the pragmatic choice.

2.6.3 Scales adopted from existing research

2.6.3.1 Social isolation scales

Within intervention or training research generally, the lived experience of the individual is critical. While it is valid to say that a given intervention or training results in an improvement over another intervention or acticitiy in a given cognitive domain, a crucial aspect missed in this narrow approach (if this is taken as the only marker of success) is the experience of the individual. An intervention or training may be beneficial or effective on paper or in terms of the numbers, but if the participant/individual does not find purpose in the intervention or does not enjoy it, the numerical efficacy is somewhat voided as they are unlikely to continue with it outside of the research context. If the goal of research is to contribute to knowledge and impact the lives and ways of thinking of people beyond/outside the scope of academia, surely the lived experiences of participants in intervention research carries some, if not equal weight to the numerical findings. Indeed, qualitative data outlining the lived experience of participants can inform new variables or avenues for quantitative research questions. For example, the social impact of language learning classes became evidence through qualitative data collected during a 2017 study in collaboration with Lingo Flamingo and Alzheimer Scotland (Vega Mendoza et al., 2021). Based on this experience, in all subsequent studies, whether with healthy ageing adults across the lifespan or people living with dementia, I have included quantitative scales to measure the social impact of the language learning experience.

Following our findings from Vega Mendoza (2021), I conducted a review of existing scales measuring loneliness and isolation. Loneliness is not a phenomenon unique to a specific age group or demographic: it is experienced

by children, adolescents, adults, and older adults alike (Perlman & Landolt, 1999). However, there is evidence to suggest a general trend of increasing experiences of loneliness as individuals age due to decreasing social integration and increasing rates of poor health or disability (Jylhä, 2004). Furthermore, loneliness is both a state of being and a feeling, meaning it can be a subjective phenomenon. For example, objectively stating that an individual interacts with ten people a day neither illuminates the experience of the individual nor their satisfaction with those interactions. One person may feel satisfied and content with those ten interactions, while another may feel lonely and isolated. Indeed, within the literature there are related though distinct terms to refer to these two contexts: social isolation and loneliness. **Social isolation** "concerns the objective characteristics of a situation and refers to a small network of kin and non-kin relationships" and can be thought of as one end of a continuum with social participation on the other (de Jong-Gierveld et al., 2018, p. 391). Individuals with a small number or non-existent number of meaningful relationships are considered to be socially isolated. Social isolation is a risk factor for loneliness though does not necessitate it. An individual can be socially isolated and not lonely or lonely and not necessarily socially isolated (de Jong-Gierveld et al., 2018; Victor et al., 2000). **Loneliness**, on the other hand, is "the unpleasant experience that occurs when a person's network of social relations is deficient in some important way, either quantitatively or qualitatively" (Perlman & Peplau, 1981, p. 31). Quantitative loneliness refers to the number of existing relationships being less than desired whereas qualitative loneliness entails the quality of those relationships being poorer than desired. Individual differences in expectations or standards make loneliness a more subjective state, and therefore more difficult to measure, especially when considering the compounding negative connotation or

societal stigma around loneliness. Within the literature, loneliness is also referred to as perceived social isolation, distinguishing it from the more objective definition of social isolation (Cacioppo et al., 2009).

There are a variety of scales aimed at measuring social isolation and loneliness, including the UCLA Loneliness Scale (Russell et al., 1978), the de Jong-Giervald loneliness scale (de Jong-Gierveld & Kamphuls, 1985), and the Social Connectedness Scale (R. M. Lee & Robbins, 1995). The UCLA Loneliness Scale is a unidimensional scale developed and validated on a college student population consisting of 20 items rated on a 4-point, fully labelled Likert scale where higher scores represented greater feelings of loneliness (Russell et al., 1978). The de Jong-Gierveld loneliness scale, on the other hand, is based on a multidimensional approach to loneliness (i.e., type of deprivation, time perspective, and emotional characteristics) and contains 11 items rated on a 5-point, fully labelled, ascending horizontal Likert scale (de Jong-Gierveld & Kamphuls, 1985). In their original versions, both scales refrained from using the term "loneliness" explicitly, though all items in the UCLA Loneliness scale were worded negatively while the de Jong-Gierveld scale contained 5 positive and 6 negative items. The Social Connectedness Scale was validated using a college student sample and is comprised solely of negatively worded items. There are 8-items rated on a 6-point, descending horizontal Likert scale measuring unidimensional feelings of social connectedness (R. M. Lee & Robbins, 1995). Of these, the UCLA Loneliness Scale is most commonly used (Perlman & Peplau, 1981). In fact, in a systematic review of health promotion interventions preventing social isolation and loneliness among older adults published between 1970 and 2002, Cattan and colleagues (2005) found that the UCLA loneliness scale was used in 35% of studies. The authors noted that the de Jong-Gierveld loneliness scale, which was developed for adults specifically, was

only used in 6% of studies. Notably, 17% of the studies included in the review did not use a loneliness scale, but rather included an item such as "I feel lonely" to an existing scale (Cattan et al., 2005).

While these scales would be suitable to address the question of the social impact of language learning experiences, the length of the UCLA Loneliness Scale (20 items) and de Jong-Gierveld loneliness scale (11 items) was a point of concern in terms of unnecessarily fatiguing participants as they would be integrated into the larger research battery. The Social Connectedness Scale, while short, raised concern as all items were negative, which could put off participants or invoke response resistance. Additionally, as the question of social impact was developed in the context of language learning classes for older adults and people living with dementia, the fact that none of these scales was developed with this population in mind was not ideal. On the other hand, the Friendship scale (Hawthorne, 2006) was created and validated using a sample of 829 older adults (defined as 60+) from the World Health Organization Quality of Life Group's larger study measuring quality of life of older adults. It is a unidimensional scale comprised of six items rated on a 5-point, fully labelled, descending, vertical Likert scale. Three of the six items are reversed scored.

After piloting the four reported scales in studies not reported as part of the current thesis, the Friendship Scale was determined to be not only the most suitable due to the content of the items but also due to its function as a standalone measure of perceived social isolation, ease of administration and the short time to complete it.

2.6.4 Scales developed for the present research

2.6.4.1 Language History Questionnaire

Across the majority of studies presented as part of the current research, participants completed a version of a background questionnaire to gather information on their language history and general demographic variables. After identifying their age, gender, education level, and country of residence, participants rated their expression, comprehension, reading, and writing skills for all languages they have at least a basic knowledge of. These language command skills are rated on a 5-point, fully labelled, ascending horizontal Likert scale ranging from Beginner (1) to Fluent (5) (i.e., None, Beginner, Elementary, Intermediate, Advanced, Fluent). Following the procedure outlined in Vega-Mendoza and colleagues (2015) and Bak and colleagues (2016b), these ratings across all four domains were summed into a composite language background proficiency score for each participant (with a maximum score of 20 per language). In the sequencing empirical language learning chapters (Chapter 4, Chapter 5, Chapter 6), this score is referred to as language knowledge before the class. The frequency of use of the language in childhood, adulthood, and the last three years was also rated on a 6-point, fully labelled, ascending horizontal Likert scale ranging from Less than yearly (1) to Daily (5) (i.e., Never, Less than yearly, Yearly, Monthly, Weekly, Daily).

2.6.4.2 Language Attitudes

Within each language learning study, participants were also a series of questions about their attitudes towards the language. The language attitude questions varied for each individual language (see Chapter 4, Chapter 5, Chapter 6), though two questions were presented across all studies: "*This*

language is difficult to learn" and "*It's really food for your brain to learn a language like this one*". As originally this was presented in a paper-and-pencil questionnaire, in order to save space on the page the responses were presented in an ascending, horizontal 6-point Likert scale ranging from Disagree Strongly (1) to Agree Strongly (6). This presentation was also maintained in the online format.

2.6.5 Cognitive measures

2.6.5.1 Test of Everyday Attention

The primary outcome measure of the language learning studies reported as part of this thesis are scores in the cognitive domain of attention as measured by the Test of Everyday Attention (TEA). The Test of Everyday Attention is a normed and standardised clinical battery for adults measuring different attentional capacities (Robertson et al., 1994). Originally designed to measure the effects of neurorehabilitation in patients with traumatic brain injury and brain damage, the test includes three parallel versions to allow for sensitive and fine-grained analysis of attentional changes over time. The test itself, given the name, uses ecologically plausible (i.e., real life) activities and tasks to measure attention, such as searching on a map. The battery is based on Posner and Peterson's (1990) neuro-anatomical multi-system attentional model of orienting, detecting, and alerting via orientation, vigilance and selecting systems. The selection system selects relevant stimuli or processes while inhibiting those that are irrelevant whereas the vigilance system is responsible for sustaining alertness or readiness to respond when external cues are lacking (Posner & Petersen, 1990; Robertson et al., 1996). The orientation system, on the other hand, moves, engages, and disengages attention in space, and thus is difficult to assess via a paper-and-pencil test such as the TEA.

The TEA comprises eight subtests meant to represent everyday tasks, assessing sustained attention (i.e., the ability to focus on repetitive stimuli without external motivation), selective attention (i.e., the ability to select and attend to relevant target items whilst ignoring distractor or irrelevant stimuli), and attentional switching (i.e., the ability to flexibly switch attention from one framework to another) (Robertson et al., 1994, 1996). The subtests include: (a) map search, (b) elevator counting, (c) elevator counting with distraction, (d) visual elevator, (e) elevator counting with reversal, (f) telephone search, and (g) lottery. Taken together, administering the full battery can take between 60 and 90 minutes depending on the participant. The TEA was normed using a pool of health ageing participants aged 18-80 as well as unilateral stroke patients two months post-stroke (A. S. Evans & Preston, 2011; Robertson et al., 1996). More recently, the TEA has been validated as a useful measure of attentional resources in community-dwelling older adults aged 80 to 101 (G. van der Leeuw et al., 2017). In validation, a four-factor structure was revealed containing sustained attention, selective attention, attentional switching, and auditory-verbal working memory, which map onto Posner and Peterson's (1990) multi-system attentional model (Robertson et al., 1996). Sustained attention parallels the vigilance system while selective attention, attentional switching, and visual-auditory working memory equates to the selection system (A. S. Evans & Preston, 2011).

The TEA and the auditory subtests in particular were selected as the primary cognitive outcome measure due to it (a) providing a fine-grained methods of assessing theoretically distinct aspects of attention, (b) including auditory measures which are appropriate due to the auditory nature of language learning and language use, (c) being a non-verbal measure, (d) being an ecologically valid test, and (e) including three parallel versions to allow for

multiple testing time points. Typically, the TEA is administered in a one-to-one setting with a participant and test administrator or researcher. However, as part of the current PhD research, the auditory elevator counting subscales (discussed below) were piloted and validated when administered in-person in a group setting with healthy ageing adults across the lifespan. In such cases, the audio files were played out loud via speakers and participants were asked to record their answers on a sheet of paper. Feedback to the practice items was given verbally, as when the test is administered one-to-one. This approach necessitates that participants can hold a writing utensil and record their responses independently, and thus is not suitable for older adults with motor issues or those living with dementia.

The three auditory elevator counting subtests were used in the present research. In the elevator counting subtest, participants are asked to imagine they have entered into an elevator or lift but the floor light indictor is out, so they cannot visually see which floor of the building they are on. They are asked to listen to a series of tones, counting them in different ways for the three subtests, and indicate which floor of the building the lift has stopped on when cued by the voice on the recording. The elevator counting subtests include Elevator Counting (ET), Elevator Counting with Distraction (ETD), and Elevator Counting with Reversal (ETR), which take 15-20 minutes to complete including instruction time. Following the instruction manual, the practice trials are not included in scoring (Robertson et al., 1994). For each subtest, the percentage of correct trails is calculated, ranging from 0 to 100. Participants are instructed to count in their head, without the use of their fingers or tick marks on the page to keep track of the number of tones. Save for the ET wherein healthy ageing adults typically make few errors, normative data shows that participants

vary across the range of scores in the ETD and ETR tasks (Robertson et al., 1994, 1996).

2.6.5.1.1 Elevator Counting

The Elevator Counting task involves a single tone which is repeated at varying intervals. Participants must keep track of these tones without losing focus. There are two practice trails and seven test trials. This subtest is a measure of sustained attention and is the least computationally difficult of the three measures. However, ET can be used to discriminate between patients with early-stages versus mild or more progressed Alzheimer's Dementia (J. D. W. Greene et al., 1995; Robertson et al., 1996).

2.6.5.1.2 Elevator Counting with Distraction

The Elevator Counting with Distraction task involves two tones: the same tone as in the ET task as well as a new higher pitched tone. Participants are asked to count the familiar low tone while ignoring the new higher pitched tone. There are two practice trials and ten test trials. This subtest is a measure of selection attention or inhibitory control as well as auditory-verbal working memory.

2.6.5.1.3 Elevator Counting with Reversal

The Elevator Counting with Reversal task is the most complicated of three elevator counting subtests. Participants are presented with three tones: a high pitch tone, a middle tone (the low tone in the ETD task and the only tone in the ET task), and a new low tone in random order. Across all three elevator counting subtests, participants are counting the same tone – in this case the middle tone (the high and low tones are not counted). The high tone

and the low tone are used to indicate the direction in which in count the middle tone. That is, the low tone indicates that the lift is moving down, so subsequent middle tones should be counted by subtraction. The high tone, by contrast, indicates that the lift will be moving up, so the middle tones which follow should be counted by addition. As this task is more demanding, there are three practice trials and ten test trials. This subtest is a measure of attentional switching or mental flexibility and auditory-verbal working memory.

2.6.5.2 Raven's Advanced Progressive Matrices

General non-verbal intelligence, as measured by Raven's Advanced Progressive Matrices (APM) (Raven & Foulds, 1962), was included as part of the language learning studies research battery to act as a control variable for the primary outcome of attention scores. Raven's APM is a non-verbal measure of abstract reasoning skills which is progressive, in the sense that the test becomes increasingly difficult as it progresses. This visuospatial problem solving task is thought to access general intelligence (so called 'g') which is independent of education or culture (Hall et al., 2010). Participants are presented with various patterns or puzzles in the form of a matrix with a piece missing. They are asked to find the puzzle piece, from a selection of 8 choices, which solves the pattern in add directions (e.g., up and down, side to side, as well as along any diagonals, etc.). There is only one correct answer for each matrix. Raven's APM is comprised of two sets: Set I and Set II. Set I functions as (a) a practice set, (b) a short version of the test, or (c) a means of verifying if the Standard Progressive Matrices would be more appropriate with 12 items while Set II has 36 items (Raven et al., 1998). All items are presented in black and white, minimising additional considerations of participants with variations of colour-blindness. APM were selected above the Standard Progressive

Matrices in an effort to minimise ceiling effects, as the majority of participants within the current research had completed a University Degree or higher.

In the effort of avoiding an overly complicated and length testing battery, the full APM (both Set I and Set II) were not administered to participants. Instead, following Ooi and colleagues (2018), who found a ceiling effect when using Set I alone, three items from Set I were administered as practice trials with feedback and participants completed items in Set II with a time limit of 10 minutes. Participant scores remain valid and reliable even when a subset of the items are attempted, if they are presented in order. The 36 items in Set II "are arranged in order according to the frequency with which they are solved, the validity of the total scores does not depend upon everyone attempting all items in the Set" (Raven et al., 1998, p. APM 3).

2.7 Influence of the CoViD-19 pandemic and resulting lockdowns

In the initial conceptualisation of the current research, the vast majority of studies were to take place in-person either with group or one-to-one testing of healthy ageing older adults and older adults living with dementia, respectively. Unforeseen hurdles, namely the COVID-19 pandemic in March 2020 meant that any patient-facing research was no longer feasible within my PhD research portfolio (though the first care home visit was scheduled for 17 March 2020 after a year of ethics reviews and discussions with collaborators). The first set of empirical studies presented in Chapter 3 were initially developed and designed to be conducted online. The language learning studies in Chapter 4, Chapter 5, Chapter 6, were either adapted to an online delivery after data collected had already begun or were reconceptualised to be suitable for an online delivery.

While some considerations were unique to individual study design, discussed further in the relevant chapters, there were a handful of decisions which affected all language learning studies. The decision of which online platform was best suited for the delivery of the language learning classes was largely left to the individual language teachers, while I decided on which platform would be best for the research sessions. However, in all cases I discussed the various options with the teachers keeping in mind the demographic of students (see Chapter 7). As the target sample was adults across the lifespan or mixed age classes, older adults not familiar with online video platforms would be included. While some considerations such as avoiding platforms which require numerous applications to be installed, additional downloads, or creating a user account were recognised in the initial transition to online, others were learned through experience (see Chapter 7).

For example, Zoom, which became incredibly popular during the pandemic, is developed with meetings and speaking particularly in mind (*Zoom*, n.d.). Normal speaking voices are typically very clear on Zoom, though music or tones are difficult if not impossible to correctly transmit. The application, as I now know, obtains high-quality audio of speaking voices by eliminating and minimising background feedback and high-frequency sounds during calls. This means that auditory stimuli will be unreliable and, in some cases, entirely supressed. For this reason, Zoom was quickly rejected as a platform for the individual testing or research sessions, though was a popular platform for the language classes themselves. Other positives of Zoom include the fact that it is designed to work in low-bandwidth situations, various viewing options to allow for many participants to be visible at one time on the screen, and collaborative learning functionalities like breakout rooms, further

discussed in the context of Chinese language learning, presented in Chapter 5, in Zhao et al. (2020).

Given the auditory limitations, and at the time the 40-minute call limit due to the University of Edinburgh having not purchased a Zoom license until August of 2020, another platform was sought for the research sessions. Skype was suitable in some cases, though requires the app to the downloaded and for users to create an account which made it unsuitable as a long-term solution (*Skype*, n.d.). Instead, Whereby was used for research sessions as, at the time, a free account included a personalised meeting room and (at the time of data collection) unlimited call times (*Whereby*, n.d.). Furthermore, Whereby does not require any installations or downloads for users, instead working through an internet browser, or logins for guests and can be accessed via computer, tablet, or mobile phone. Additionally, the feature "knock to enter" (i.e., a waiting room) allowed for research sessions with different participants to be scheduled right after each other.

In the context of qualitative data collection such as focus groups, Zoom was the clear choice not only as participants were familiar with the platform due to the language classes, but also due to the ability to securely record then store sessions without a third-party software (Archibald et al., 2019).

Outside of the question of platform for the research sessions, the primary element of concern in adaption to online research was the administration and suitability of the tests themselves. The questionnaire elements were easily modified to be online via Qualtrics (*Qualtrics*, n.d.). The primary difficulty was adapting cognitive tests measuring accuracy and reaction time to an online delivery. This was particularly challenging as a solution needed to be in place not only determining online delivery of the language classes and materials, but also the research elements in under two-

weeks for the scheduled start of the next round of classes (see Chapter 5). Having only taken a short afternoon workshop introducing JavaScript and the jsPsych framework, I had to learn the coding system, design and code an experiment from scratch, as well as communicate and troubleshoot with the IT department about hosting the experiment (J. R. de Leeuw, 2015).

This unpredicted move to online research brought with it new and unanticipated research questions and avenues of research, discussed where applicable in the subsequent chapters.

Chapter 3 Language and Memory: Are Memories Language Specific?

The title of the present PhD work is twofold: "natural history and effects of language learning." This chapter explores the role of memory and language across the lifespan without any experimental intervention into participants lives through language learning, that is through a natural history approach. At the end of this chapter, a study in progress examining the question of the patterns of language changes in older age among healthy ageing and pathologically ageing multilinguals.

3.1 Introduction

3.1.1 Two approaches to language processing

While there are numerous specific models within cognitive science and psycholinguistics around bilingual language processing, one means of broadly categorising the overarching principles behind these models is a distinction of static versus dynamic language processing. Many of the traditional, more static models are based on a metaphor of information processing in the human brain being similar to that of a computer (de Bot, 2012). However, advances in various fields, most notably including neuroscientific evidence of neuroplasticity in the human brain, have led to more dynamic models of language and cognition more generally. Experience-related plasticity, or adaptation, dates back to pioneering work with rats in the 1960s (Diamond et al., 1964; Rosenzweig et al., 1962, 1964) and has been one of the most influential scientific findings in cognitive science and related fields in the 21st century. Both behaviourally and structurally, the fundamental principle guiding dynamic models comes from Dynamic Systems Theory, stating that a complex, open system (i.e., the human brain) continuously interacts with and adapts to

its environment, thus continuously changing over time (de Bot, 2010, 2012). As Van Gelder (1998), an influential advocate of the "dynamical hypothesis" in cognitive science describes there are conceptual faults with considering the mind to be in one to one correspondence with a computer:

> "the cognitive system is not a discrete sequential manipulator of static representational structures; rather, it is a structure of mutually and simultaneously influencing change. Its processes do not take place in the arbitrary, discrete time of computer steps; rather, they unfold in real time of ongoing change in the environment, the body, and the nervous system. The cognitive system does not interact with other aspects of the world by passing messages and commands; rather, it continuously coevolved with them" (p. 3).

A static model of bilingual language would hold that the first language input (i.e., acquisition of L1), is the most important and once in place, does not change. Returning to the metaphor of the brain as a computer, the driving force of processing and system functioning is that which was first input. Additional languages may be added, but the information is stored and accessed via the original language of input. Such an approach can be referred to as the Mother Tongue Hypothesis, which takes a deterministic view of language wherein one's first language is the most influential. A more dynamic approach to language and memory recognises individual differences (e.g., varied proficiency, linguistic environments, language switching, etc.) and the role that varied environments and experiences play in information processing. Indeed, at the structural level, dynamic approaches to bilingual language processing can account for the influence of multilingualism on structural neuroplasticity more effectively than static models, integrating and explaining the varied findings in the literature (e.g., Plaitsikas's (2020) three-stage Dynamic Restructuring Model). An interesting means of exploring these

models of language and how language is processed is through memory, one of the fundamental functions of our cognitive system.

3.1.2 Divisions of memory

Long-term memory is a complex system which can be broken down in terms of (a) the type of information (e.g., semantic memory and episodic memory) or (b) longitudinal processes (e.g., encoding, storage, and retrieval). These two means of dividing memory are orthogonal, meaning they can be thought of as different dimensions rather than mutually exclusive. **Semantic memory** is knowledge of facts and information, that is general cultural and world knowledge (e.g., ideas, meanings, concepts, facts) (McRae & Jones, 2013). **Episodic memory**, first coined and distinguished from semantic memory by Tulving (1972) is the memory of "episodes" or events (Wheeler & Ploran, 2009). For example, semantic memory encapsulates knowledge of what a dog is (i.e., semantic retrieval), whereas episodic memory might entail a specific memory of taking your childhood dog to the beach (i.e., episodic retrieval). Unlike episodic memory, semantic memories do not contain contextualized information (Gillund, 2012). That is, you "know" a fact (i.e., **noetic awareness**, meaning "knowing") rather than remember the experience(s) of acquiring that knowledge (i.e., **autonoetic awareness**, meaning "self-knowing"). For example, you may know that 'the mitochondria is the powerhouse of the cell' (semantic memory), which is distinct from the various episodes or experiences in grade school sciences classes being taught this information (episodic memory).

Both semantic and episodic memory are types of declarative or explicit memory, that is "memories which are accessible to consciousness and which can be readily communicated or declared to others" (Wheeler & Ploran, 2009, p. 1167). Declarative memory is distinct from nondeclarative or implicit

memory, which entails the memory of how to perform specific behaviours or skills often without an awareness of retrieving that information (e.g., riding a bike) (Gillund, 2012; Wheeler & Ploran, 2009). Another way of thinking about these distinctions is by considering the type of consciousness related to these kinds of memory. Tulving (1985) delineated the terms anoetic, noetic, and autonoetic consciousness. **Anoetic consciousness**, meaning "not knowing" corresponds to procedural or implicit memory, noetic to semantic memory, and autonoetic to episodic memory.

Overlapping with, though not identical to, episodic memory is **autobiographical memory** or the memory for and about the self, which can also encapsulate personal knowledge and skills outwith episodic memory (Gillund, 2012). For example, remembering your childhood best friend's home phone number is not only a part of your personal history and may be associated with a specific event or memory, but is also a fact. Autobiographical memory is accompanied by conscious recall and a sense of reliving, "a sense that the remembered event actually occurred to you at specific time and place" (Rubin, 1998, p. 49; Tulving, 1989). Rubin (1998) outlines the minimum five key elements of autobiographical memory: cueing, retrieval, language, imagery, and affect. Some stimuli (e.g., word, sound, smell, etc.) instigates a set of cognitive processes that retrieves a narrative of a lived experience accompanied by visual imagery and emotional content (or affect). Brewer (1996) defines autobiographical memory, which he refers to as recollective memory, as:

> "memory for a specific episode from an individual's past. It typically appears to be a 'reliving' of the individual's phenomenal experience during the earlier moment. Thus, these memories typically contain information about place, actions, persons, objects, thoughts, and affect…The information in this form of memory is expressed as a mental image…They are accompanied by a belief that the remembered episode

> was personally experienced... Recollective memories give rise to high confidence in the accuracy of their content" (pp. 60-61).

Indeed, recollecting (i.e., episodic retrieval), or having a sense of reliving the original experience, is a fundamental and distinguishing feature of autobiographical memory and episodic memory more generally (Rubin, 1998; Rubin et al., 2003). Recollection and autonoetic consciousness distinguish episodic autobiographical memory from more factual semantic knowledge about oneself (e.g., date of birth). When accessing factual information such as your birthday or name, there is not an associated sense of reliving or mental time travel (corresponding to noetic consciousness). You "know" when your birthday is, you do not relive being born to retrieve this information.

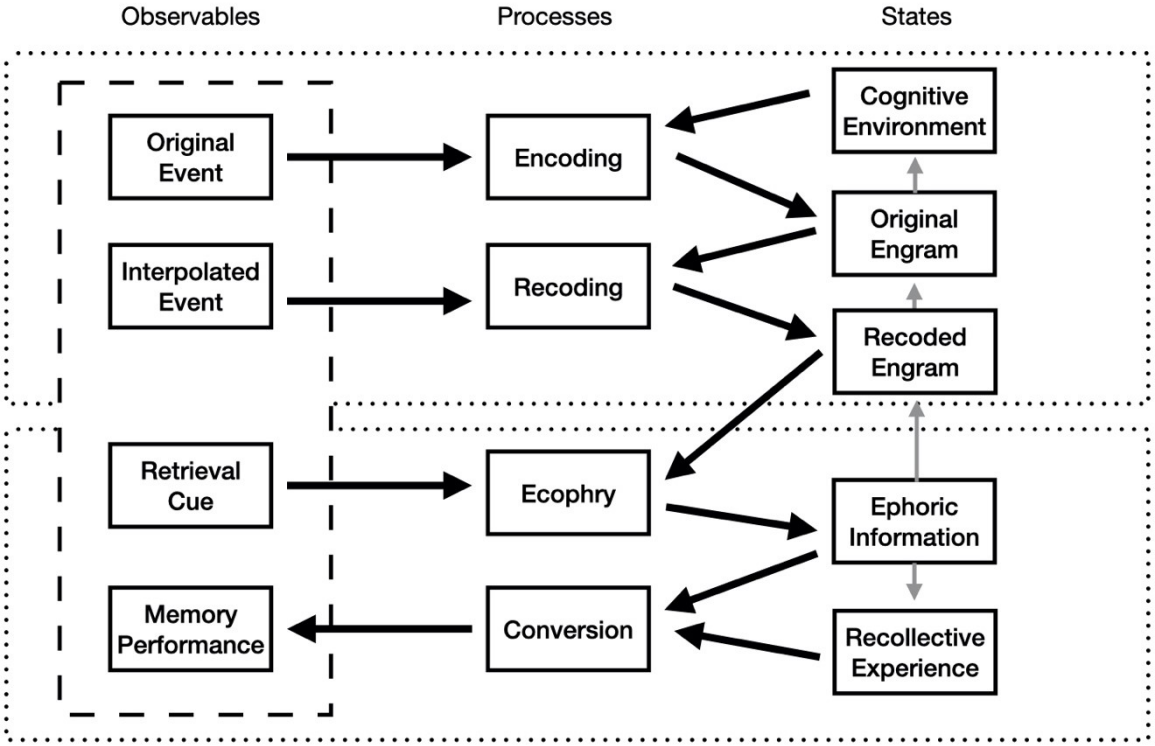
Much of the research on language and recollection in autobiographical memory focuses on narrative structure in the retelling of the memory, not the role that knowledge and use of multiple languages may play in recollection. As a fundamental element of autobiographical memory, the question then arises of what role does language play in recollection among multilinguals.

The predominant theory of autobiographical memory posits that these memories are reconstructions of the original event, rather than a copy of an immutable record of the event (W. F. Brewer, 1996; Schrauf & Rubin, 2004). The mechanism of encoding and retrieval of episodic memory can be understood though the General Abstract Processing System (GAPS, see Figure 3.1) (Tulving, 1983, 1984).

> "It is general in that it is meant to apply to remembering of events of all sorts; it is abstract in that the specific nature of its components is not specified; it is a processing system since its major components have to do with the activity and the functioning of the system rather than its structure; and it is a system in the sense of an ordered and reasonably comprehensive collection of interacting components whose assemblage constitutes an integrated whole" (Tulving, 1984, p. 229).

Under this model, observable events, underlying mental processes, and unobservable mental states are distinguished from each other (Tulving, 1983, 1984). In a research context, as in the present study, observable states are limited to the moment of memory retrieval (Schrauf & Rubin, 2004). The original events, called engrams (i.e., mental traces), are theoretically observable though often not in practice as this would require intense longitudinal studies controlling the encoding processes. Further, this original engram can be altered via reconstructions or subsequent experiences and be recoded. In the research context, retrieval is simulated by giving the participant a retrieval cue (such as a word or picture) and subsequently asking them to share the related story or memory. The retrieval process itself, called ecophry, is complex, integrating the information from the recoded engram and the retrieval cue. The result is an observable memory (i.e., memory performance), typically shared via language. Indeed, in the case of a multilingual individual, the specific language for all observable conceptual elements in the GAPS model could theoretically vary. As described by Tulving (1984), the basic unit of study in episodic memory is the act of remembering or recollection. As such, the present study focuses on the role that language plays in remembering or retrieval of autobiographical memories/events specifically. Further, remembering can be thought of as an expression of autonoetic consciousness, which entails mental time travel (remembering), self-projection, and notably episodic future thinking (Tulving, 1985). By investigating not only memories but also imagined future events, the present study is aims to illuminate a more complete picture of the role language plays in the bilingual autonoetic consciousness.

Figure 3.1 GAPS model



Note: The 13 boxes represent conceptual elements, organized in three groups. The first dotted box represents elements of encoding and the second elements of retrieval. The large, black arrows represent relations which can be read as "influences", "brings about", or "has an effect on". The smaller, grey arrows "represent relations that do not affect the ongoing act of remembering but may influence the outcome of a subsequent one" (Tulving, 1984, p. 229).

Adapted from "Elements of Episodic Memory and Their Relations" by E. Tulving, 1983, p. 135. Copyright 1983 by Oxford University Press.

3.1.3 Galton-Crovitz cueing technique

The empirical study of autobiographical memory was pioneered by Sir Francis Galton (1879a, 1879b) in the late nineteenth century. Galton set out to study his own memory by developing an unconstrained method to sample his memory store by reflecting on a word or object until he made an association with a memory from his life. In his original experiment, Galton describes walking "leisurely along Pall Mall, a distance of 450 yards, during which time I scrutinised with attention every successive object that caught my eyes, and I allowed my attention to rest on it until one or two thoughts had arisen through direct association with that object" (Galton, 1879a, p. 151). In later studies, Galton (1879b) focused on words specifically rather than objects. He produced a list of 75 memory probes, or cue words, recording and dating the associations he made once every month for four months. After forming associations with the words, he then made a note of the age of each association with as much accuracy as possible.

Galton's cueing technique was further developed and modified by Herbert Crovitz (1976; 1974) in the 1970s, when studying the distribution of memories among college students. Crovitz and Schiffman (1974) constrained Galton's approach by using common English words likely to refer to all periods in life, that is "to develop Galton's methods as an unbiased probe into the store of episodic memory and to treat the reported ages of memories in an unbiased and meaningful way" (p. 517). This method of cueing autobiographical memory has the advantage of providing a random sample of memories (Crovitz & Schiffman, 1974; Schrauf & Rubin, 2001). The general principle of this method is to present, often verbally, participants with words which function as cues to retrieve autobiographical memories. In other words, the Galton-Crovitz technique is a type of cued recall task. In the decades following

Crovitz and Schiffman (1974)'s reintroduction of Galton's technique into experimental psychology as an empirical means of sampling autobiographical memories, the modified Galton-Crovitz cueing technique has been used with a variety of memory probes or cues in the form of words (Rubin et al., 2003), music (Platz et al., 2015), odors (Chu & Downes, 2002), pictures (El Haj et al., 2020), or indeed a combination of cue types (Cady et al., 2008; Greenberg et al., 2011; Rubin et al., 1984). Further, this cueing technique has been employed to elicit both autobiographical memories and imaged possible future events (Spreng & Levine, 2006) as well as adapted for online implementation (Janssen et al., 2006, 2011; Spreng & Levine, 2006) with various rating scales and questionnaires (Janssen et al., 2011; Rubin et al., 1984, 2003). Therefore, this technique has been validated with a variety of stimuli and presentation methods. The present study integrates several empirically established applications of the Galton-Crovitz cueing technique in being conducted online, using written words as cues, and including past as well as prospective autobiographical events.

3.1.4 Bilingual autobiographical memory

Thus far, the studies cited assume, in some cases not explicitly, that these memory processes (e.g., encoding, storage, and retrieval) happen in one language. In the early 1980s, Grosjean (1984) estimated that at least half of the world was bilingual, given the number of recognized languages outnumbered countries thirty to one. Given the frequency of bilingualism across the world, the question arises of what role language might play in the processes of memory encoding, storage, and retrieval. It is obvious that language plays an important role in communicating memories with others, be it writing or speaking, though the function of language in the encoding, storage, and initial

internal retrieval of memories is less clear. Theoretically, someone who speaks three languages could experience each of these three processes in a different language. As multilinguals can communicate a memory in a different language than it was encoded or retrieved, the question arises of to what extent memories are language specific.

A subset of the larger literature employing the Galton-Crovitz cueing technique to study autobiographical memory has looked at bilingual speakers, largely immigrants to Western, English-speaking countries. Bilingual groups studied in this literature include Spanish-English (J. G. Cox & Zlupko, 2019; Esposito & Baker-Ward, 2016; Marsh et al., 2015; Schrauf, 2003; Schrauf & Rubin, 1998, 2000, 2001, 2004), Polish-Danish (Larsen et al., 2002), Danish-English (Mortensen et al., 2015), Russian-English (Marian & Kaushanskaya, 2004, 2008; Marian & Neisser, 2000) and Japanese-English speakers (Matsumoto & Stanny, 2006). The vast majority of studies include LX (additional language) speakers of English. Notably, only Larsen and colleagues (2002) included non-English speaking participants. Most of the work with this specific population (i.e., bilinguals) in mind has been conducted by Robert Schrauf and David Rubin with Spanish-English bilinguals. Studying autobiographical memories in a multilingual sample allows for the investigation of questions around the role of language(s) in encoding, storage, and retrieval of memories. The primary theoretical question in bilingual autobiographical memory studies is whether people who speak more than one language can equally access memories in both (or all) languages. Indeed, at the lexical level, the Common Store Model of bilingual memory (or the successive Revised Hierarchy Model which is dependent on levels of fluency) suggest that bilinguals access conceptual knowledge (e.g., the concept of summer) in both mental lexicons (e.g., *summer* in English and *été* in French)

(Kroll et al., 2010, 2014). Further, the current theory is one of parallel activation, meaning that all languages are active in the multilingual speaker's mind regardless of their intention to use a single language in their repertoire (Kroll et al., 2014).

Many studies of bilingual autobiographical memory explore a phenomenon referred to as language dependent recall (LDR), which is an application of the overarching principle of encoding specificity (Tulving & Thomson, 1973). The encoding specificity principle holds that the degree of similarity among the encoding context (properties of the memory) and retrieval context (cueing environment) facilitates recall (Marsh et al., 2015; Tulving & Thomson, 1973). LDR further holds that language specifically is the most pertinent memory property between encoding and retrieval experiences. Language can function as both an external context (e.g., words spoken out loud) and an internal context (e.g., self-talk) for memories (Marian & Neisser, 2000). Applied to the Galton-Crovitz technique, this suggests that when the encoding language and cue language match, memories should be more easily and therefore frequently recalled in that language. Further, the Mother Tongue Hypothesis would suggest that L1 in particular, as it is the first language and dominant in formative years of life (in the case of consecutive bilinguals) would more strongly elicit L1 memories.

Larsen and colleagues (2002) found that 20 Polish immigrants to Denmark tested 30 years after seeking political asylum (so-called "Early Immigrators" M_{age} = 51.30, SD = 2.55; so-called "Late Immigrators" M_{age} = 61.40, SD = 3.34) were more likely to produce memories in Polish prior to immigration when cued in Polish (L1) and Danish cues produced memories after arriving to Denmark. The authors conclude that memories are tagged with language at encoding, which remains stable and effects later retrieval

processes (Larsen et al., 2002; Schrauf & Durazo-Arvizu, 2006). This pattern was also found among 24 Russian-English bilinguals who had immigrated to the United States (M_{age} = 21.8, SD = 2.9) where Russian-encoded (L1) memories were more likely to be recalled with Russian cues than English cues, and vice versa (Marian & Neisser, 2000). Similarly, Matsumoto and Stanny (2006)'s study supported encoding specificity and LDR among 18 Japanese-English bilinguals going to college in the United States (M_{age} = 2.7, SD = 2.65). Supporting both LDR and the Mother Tongue Hypothesis, Mortensen and colleagues (2015) found that among 24 Danish-English bilingual college-aged students (M_{age} = 22.5, SD = 4.7) more memories were retrieved (which the authors conclude translates to ease of retrieval) when cue words were presented in L1 (Danish). Further, the participants reported that regardless of cue language (L1 or L2), more memories were recalled (internal language) in L1 (Danish).

However, not all studies so clearly support the dominance of L1 in the retrieval of autobiographical memories, particularly among older bilinguals. Among 12 Spanish-English immigrants to the United States (M_{age} = 61.58, SD = 2.93) cues in L1 (Spanish) did not produce earlier memories (Schrauf & Rubin, 1998). Further, 20% of memories recalled internally (i.e., language of retrieval) did not match the language of the cue. Similarly, Schrauf and Rubin (2000) also found the cue language failed to influence the mean age of memories among 8 Spanish-English bilinguals (M_{age} = 65.63, SD = 2.77) who had immigrated to the United States. However, regardless of the cue language, memories recalled in L1 (inner speech being in Spanish) were from earlier periods of life. This may suggest that among older bilinguals with high levels of proficiency in both languages, access to memories in both languages is relatively equal. Notably, as highlighted by Marian and Neisser (2000), generalizing the age of memories to mean ages is not sufficiently discriminatory among this population in

particular, who has lived in the United States from age 28 to 65. Given this, the participants likely have memories from the nearly four decades spent in the United States in English, Spanish, and both languages, which would result in a similar mean age of memories by language. Particularly relevant to the current study, Schrauf and Rubin (2004) found the vividness (referred to as intensity of imagery) and emotional intensity of memories to be stronger when recalled in L2 (English) or both L1 and L2 (both Spanish and English) among 30 older adult Spanish-English bilingual Puerto Rican immigrants to the United States (M_{age} 69.35, SD = 6.90). The authors suggest that this may be due to L2 memories being more recent and novel than L1 memories.

Indeed, most of the literature around bilingual autobiographical memory has focused on the relationship and match between the cue word and language of encoding or the cue word and language of recall or retrieval of the memory. However, language specificity may manifest itself in other ways beyond frequency differences in the number of memories retrieved or the age/time in life in which the events occurred. A more nuanced investigation of this idea, as adopted in the present study as well as Schrauf and Rubin (2004), is not only to note the number of memories retrieved in L1 or LX, but also to explore the quality of these memories in terms of vividness and emotional intensity.

3.1.5 Present study

The present study builds upon previous literature to analyse both memories and imagined future events in a large, diverse (varying in age, contexts of language acquisition, and language combinations) group of bilinguals to further investigate the role of language in autonoetic consciousness, leading to two research questions:

- Are memories more vivid and intense when recalled in L1 in multilinguals across the lifespan?
- Are future events more vivid and intense when recalled in L1 in multilinguals across the lifespan?

According to the LDR and particularly the Mother Tongue Hypothesis, which takes a static view of language, there will be an invariable dominance of L1 in the recall of autobiographical memories and as well as imagined future events with these memories and future events being more vivid and intense.

3.2 Methods

3.2.1 Materials

The Galton-Crovitz test of autobiographical memory was adapted for online implementation in 6 languages: English, Spanish, French, German, Polish, and Greek. These languages were selected as they are among most commonly spoken languages within Europe (Emery, 2022b), as well as being languages members of the research team at the time had knowledge of. Indeed, five of the six languages are in the top 10 spoken languages in Europe, all six are within the top 12, and three are within the top 10 most spoken languages in the world as per the Ethnologue (Emery, 2022a, 2022b; 'Ethnologue: Languages of the World.', 2022). Further, German, French, and Spanish are amongst the most common languages taught at school, while Polish, Spanish, and Greek are common diaspora languages (and therefore common languages of immigrants). All materials were adapted from earlier studies which either (a) included bilingual participants (Marian & Neisser, 2000; Schrauf & Rubin, 1998, 2000) or (b) online implementations of the Galton-Crovitz cueing technique (Janssen et al., 2006, 2011; Spreng & Levine, 2006).

Though not included in the current analysis, after the development of the European experiment, the principles and procedures of the current study were also translated into Tamil and Mandarin Chinese using both traditional and simplified characters. Furthermore, the online implementation was extended beyond the written medium to include an auditory presentation of cues, which allowed for Cantonese to be included.

The survey was open from June 2019 until June 2022 with the last response included in the present analysis collected on June 6, 2022. As the various language are included on the same link for ease of dissemination of the experiment, the survey remains open until there are more comparable response numbers across languages. Throughout the three years that the experiment was running, various additions and edits were made. Initially, the experiment included English and five language combinations (Spanish, French, German, and Polish). English was originally selected as a common comparison language based on previous literature (Marian & Neisser, 2000; Schrauf & Rubin, 1998) as well as pragmatically being the language shared amongst all members of the research team. In this initial stage, participants had to be English speakers as English was the language in which consent was gathered, after which participants could select to continue the experiment in English only or to include one of the other four languages. In January 2021 the experiment was expanded to include non-English speaking participants as well by allowing for the survey to be conducted in German only or Polish only, or with four possible language combinations. In January 2022 Greek was also added, creating five possible language combinations per primary language (English, German, Polish, or Greek) as well as primary language only responses (of which there are four).

3.2.1.1 Cue words

A total of 16 cues were presented to participants, separated into two matched blocks (see Table 3.1; see Appendix A for the full list of cues in all five languages). Sixteen cues were used in line with previous literature (Marian & Kaushanskaya, 2004; Marian & Neisser, 2000; Schrauf & Rubin, 2000) as well as in line with literature suggesting 10-20 cues are required in order to produce temporal features of autobiographical memory (e.g., the reminiscence bump, a temporal feature in recollection of autobiographic events, well established in previous studies, which is not under investigation in the current study).

Table 3.1 Cue words by block and source

	Cue Word	Source
Block A		
1	Summer	Marian & Neisser, 2000
2	Neighbours	Marian & Neisser, 2000
3	Birthday[1]	Marian & Neisser, 2000
4	Cat[1]	Marian & Neisser, 2000
5	Mountain	Rubin et al., 2003
6	Getting lost	Marian & Neisser, 2000
7	Frightened	Marian & Neisser, 2000
8	Water	Rubin et al., 2003
Block B		
9	Winter	Marian & Neisser, 2000
10	Friends	Marian & Neisser, 2000
11	Holiday[1]	Marian & Neisser, 2000
12	Dog[1]	Marian & Neisser, 2000
13	River	Rubin et al., 2003
14	Competition	Marian & Neisser, 2000
15	Laughing	Marian & Neisser, 2000
16	Fire	Rubin et al., 2003

[1] Imagined future event cue

The majority of cues (i.e., 12 of 16) were adapted from Marian and Neisser (2000)'s study, which forms the basis for the current study. In the original study, "snow" and "contest" were used, though pilot work using the language combinations in the current study showed that these words were not well understood across cultures (particularly "contest"). "Snow" and "contest" were therefore changed to "winter" to match "summer" and "competition", respectively. The remaining cues used by Marian and Neisser (2000) (i.e., "doctor", "blood", "bride", and "newborn") were not selected due to their potential of evoking strong negative memories. As the overarching aim of the study was to investigate multilingual autobiographical memory generally, and not specifically traumatic events, it was deemed unethical to increase the probability of evoking highly negative memories by using such cue words. This is in line with the reduction methods employed by Rubin and colleagues (2003). Instead, these four words were replaced with "mountain", "river", "water" and "fire" from the list of 33 words in Rubin, Schrauf, and Greenberg (2003) which produced the smallest amount of incomplete responses or missing data from Rubin's (1980) study, where 125 words were normed on 51 rating scales. These cue words have since been employed in autobiographical memory studies in both English and Spanish (Rubin & Schulkind, 1997; Schrauf & Rubin, 1998, 2000). Furthermore, "mountain", "river" and "fire" have also been used in online autobiographical memory investigations (Janssen et al., 2006, 2011).

Four cue words, "birthday" and "cat" in Block A and "holiday" and "dog" in Block B, were chosen to illicit imagined future events. These cues were selected due to frequently occurring in everyday life (e.g., celebrating your birthday once a year). Imagined future events were included in the study as an avenue of exploratory research into the question of language specific memory

consolidation and internal thought. Indeed, as discussed above, the Galton Crovitz cueing technique has been used in previous studies to look at episodic future thinking or foreseeing (Neroni et al., 2016; Spreng & Levine, 2006).

The 16 cues as shown in Table 3.1 were divided into two broad categories of the *natural world* and *social events* and thematically matched across blocks (e.g., "water" and "fire" or "neighbours" and "friends"). This matching and thematic categorisation was an attempt to ensure respondents produced a variety of autobiographical memories while controlling for the cue content across blocks/languages. Bilingual respondents received Block A in one language and Block B in another (e.g., in counterbalanced order), therefore, the content of memories generated in both languages would be similar/controlled for.

Language combinations were limited to two to minimise participant fatigue as well as incomplete responses due to the experiment length. Furthermore, including additional languages would have complicated the strenuous cue matching procedure and subsequent pilot studies. Additionally, employing a combination of two languages is not uncommon in the literature using the Galton Crovitz technique or indeed the wider applied literature, such as the Bilingual Aphasia Test (BAT) (Gómez, 2008; Paradis, 2004).

3.2.1.2 Test instructions

Instructions for the experiment were adapted from Janssen, Rubin, and St. Jacques (2011) and in fact written in consultation with Steve Janssen, who wrote the original online Galton-Crovitz test (Meeter et al., 2005). For memories, participants were instructed that they would see six words in the language that the instructions were written in (e.g., instructions in English say they will see six English words). They were asked to write a memory from their

life that first comes to mind, which should be a specific personal experience (i.e., not something they have only read or heard about in the news). This memory was further described as needing to have occurred in a particular place and point in time but did not need to be important. Regarding the time in which the event took place, participants were told that the memory could be of any moment or event in their life (e.g., yesterday, five years ago, childhood). Finally, participants were asked to write their responses in the language of the text (both cue and instructions). In terms of imagined future events, instead of recalling a memory from their life, participants were asked to imagine a possible future event.

3.2.1.3 Rating

After the cueing stage, participants were again shown their responses and asked to answer a series of future questions about their respective memories. The rating questions used in this study were adapted from Marian and Neisser (2000) and Schrauf and Rubin (2000). Participants were asked their approximate age at the time of the event, the city/country and language in which it took place (i.e., encoding language, which Marian and Neisser refer to as "external language"), the language in which they recalled the memory (i.e., which Marian and Neisser refer to as "inner language", which is distinct from the cue language, see Schrauf and Rubin (1998)). For the language-based questions, four options were presented: the language of the cue, the other language in which participants were completing the experiment, both languages, and other (with the prompt to further specify). For example, if a participant completed the experiment in German and Polish and was rating a memory cued in German the options would be German, Polish, both languages, and other. Furthermore, expanding upon Marian and Kaushanskaya

(2004), participants were then asked to rate the memory in terms of vividness, emotional intensity, and valence. Adapted from Rasmussen and Berntsen (2009), valence was rated on a 7-point fully-labelled, horizontal, ascending Likert scale from "very negative" to "very positive". Vividness and emotional intensity were rated on a 7-point partially labelled (midpoint and endpoints), horizontal, ascending Likert scale from "not at all" to "very much so". For imagined future events, participants were asked which language they imagined the event in as well as to rate the event in terms of vividness, emotional intensity, and valence.

Following the rating section, participants were asked demographic information including their year of birth, gender, highest completed level of education, and countries they had lived in for more than one year. Participants were also asked to provide information on the languages in which they completed the experiment, specifically the age in which they first contacted the language, contexts and frequency of use across the lifespan, and self-rated proficiency. These questions were abridged from the Language History Questionnaire described in Section 2.6.4.1. Lastly, both monolingual and bilingual respondent types were asked if they had knowledge of any other languages, and if so to list them.

3.2.1.4 Translations

The cues, instructions, rating questions, and background questions were originally developed in English. The aim in translating the experiment into Spanish, French, German, Polish, and Greek was not to obtain literal one-to-one translations of English into the target language, but rather to create conceptually equivalent versions of the experiment in each language. As such, detailed consideration was given to each translation procedure in order to

retain ecologically valid language-specific elements (as opposed to conducting a one-for-one translation which clearly reads as English translated into another language). For example, just translating the word "memory" itself presented a variety of linguistic decisions. In English, memory is a single word encompassing a variety of concepts. However, in German, Polish, Spanish, and French, memory is categorised by distinct words for the faculty of memory and memory of one's personal past (Spanish: memoria vs recuerdo, French: memoire vs souvenir, German: Gedächtnis vs. Erinnerung, Polish: pamięć vs wspomnienie, Greek: μνημονικό vs ανάμνηση). In these five languages, the latter word was used in translation. Furthermore, English was the only language within the experiment without a formal and informal second person. All translated text using the English second-person "you" adopted the formal pronoun in the target language and subsequent conjugation standards. More generally, as German, Spanish, French, and Greek, in particular, have a variety of dialects, the standard form (e.g., High German) was adopted in order to increase readability for participants from various regions or countries.

The translation procedures followed a four step process of forward translation from English, back-translation to English from the target language, pre-testing, and final translation based on Brislin (1970) and the World Health Organization (2016)'s guidance on adapting research for cross-cultural contexts. Forward and backward translation was conducted by independent native speakers of each target language. Back translation was done by speakers without any knowledge of the study or of the original English version. These back translations were then compared to the original English versions, discussed with the translators and/or other speakers of the language, and modified as necessary for the final translation.

3.2.2 Procedure

The experiment took place online via Qualtrics (*Qualtrics*, n.d.). Participants were invited via an anonymous link to complete the experiment and similar to Schrauf and Rubin (1998) were told they were being asked to participant in a study about language and memory. No further specification about the type of memory was made prior to the presentation of the cueing procedure. In the effort of inclusivity, and with particular consideration of targeting adults across the lifespan, the survey presentation was altered from the Qualtrics defaults (i.e., Arial size-12) in line with readability recommendations around font color and size (Bernard et al., 2003). Specifically, the font was changed from grey to black on a white background and increased to size-16 Arial.

Participants were initially asked to select a language (English, German, Polish, or Greek) in order to proceed to the consent form, this was deemed crucial to allow for informed consent to be gathered. After providing informed consent, participants were asked to select a language combination to continue. For example, if they had chosen English to read the information sheet and provide their consent, participants were then presented with six options:

1. Spanish & English
2. French & English
3. German & English
4. Polish & English
5. Greek & English
6. English only

The experiment included four sections (described above): past events, future events, rating/dating, and a short demographics questionnaire. The specific procedure varied based on the version of the experiment, described below.

3.2.2.1 Monolingual procedure

The monolingual version of the experiment could be conducted in either English, German, Polish, or Greek. Block presentation was evenly randomised (i.e., every second participant saw Block A first). Participants were shown one block of cue words and asked to record six memories and two future events. On the next page, they were presented with the second block of cues and asked to do the same. The order of cue presentation was randomised within each block by memory and future event. Following the recording of memories and future events, participants were asked to rate each event. The first block of events was rated in a randomised order, followed by the second block (also randomised). That is to say, participants did not rate/date memories or events in the same order in which they were cued. For each event, the text a participant had written was shown at the top of the page. All cues were presented prior to any rating, as opposed to rating each memory after every cue, in order to avoid knowledge of the questions biasing the content of subsequent memories or events (Janssen et al., 2011). Lastly, participants completed the background demographic and language questions.

3.2.2.2 Bilingual procedure

The bilingual version of the experiment follows a similar procedure to the monolingual version, though the design considerations around the presentation of cues within blocks were more involved. After selecting a language combination, participants were shown a block of cues in one of the two languages. The presentation order of language and blocks was randomised and counterbalanced following a 2x2 Latin-square design. This was necessary to prevent participants randomly receiving the same block of

cues in both languages. If a participant had selected English and Spanish and they were shown Block A cues in Spanish, they were then shown Block B cues in English. In this way, all participants, regardless of the experiment version, were cued by the same 16 words. After recording memories and future events, as above, participants were asked to rate each event followed by the demographics questionnaire. As in the monolingual version, the order presentation of cues as well as rating was randomised within each block by memory or future event.

The study was given ethical approval by the University of Edinburgh PPLS Research Ethics Committee (ref: 353-1819/7).

3.2.3 Participants

Participants were recruited online via social media channels such as Twitter, Facebook, Instagram as well as language specific Reddit groups. A total of 911 participants completed the entire experiment, with 1740 total clicks or partially completed responses. Through the experiment, participants were asked to report 12 memories and 4 imagined future events, resulting in 16 total data points per individual respondent. There are 911 participants who completed all 16 cues and 173 partially completed responses, resulting in 15318 total datapoints including both memories (n = 11585) and imagined future events (n = 3733). As the experiment was conducted online, the participant pool was not geographically limited. Indeed, participants from across the globe completed the survey, increasing the representativeness of the sample compared to what could have been possible with in-person data collection in Scotland, see Figure 3.2 and Figure 3.3.

Figure 3.2 World map of respondents

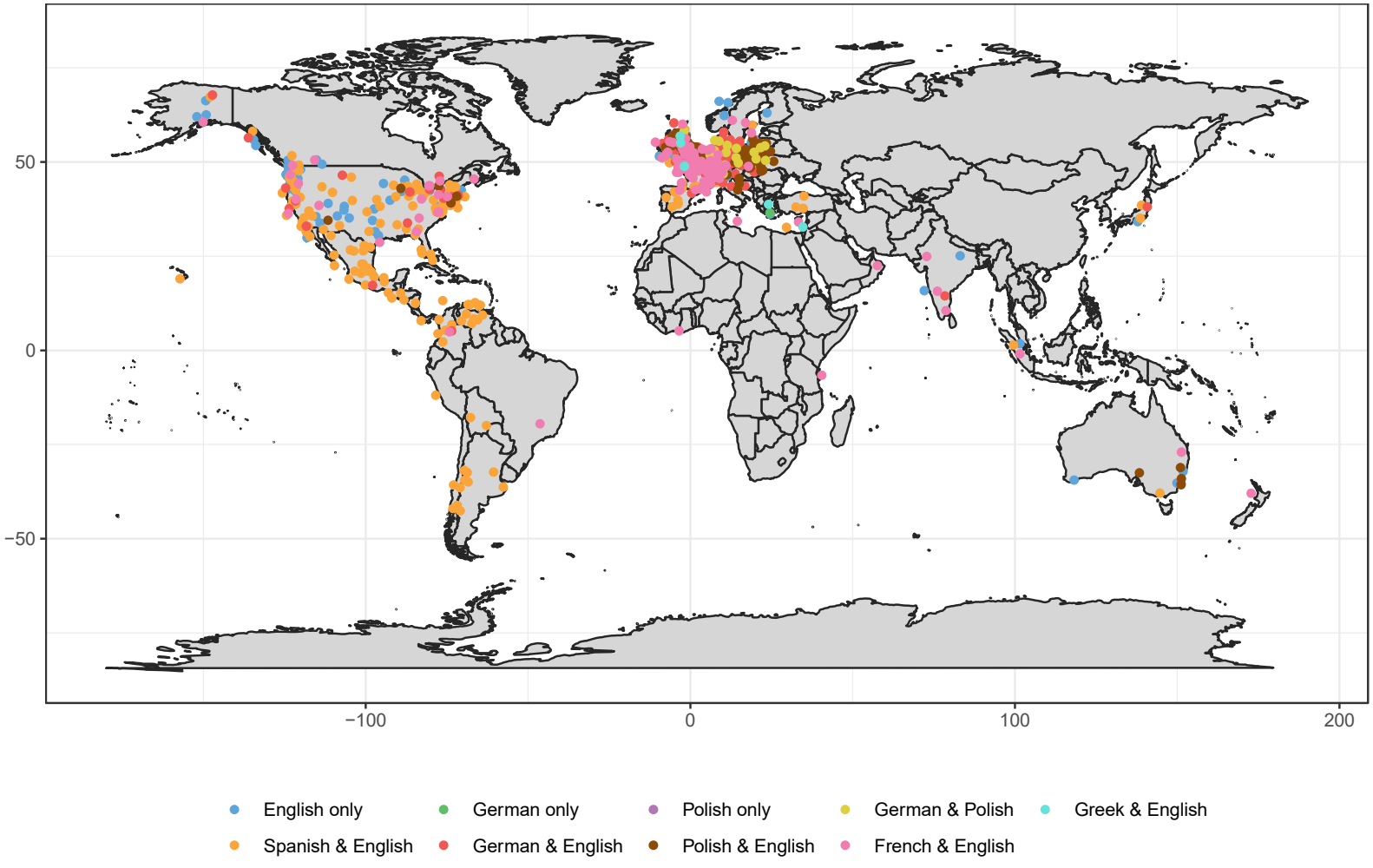
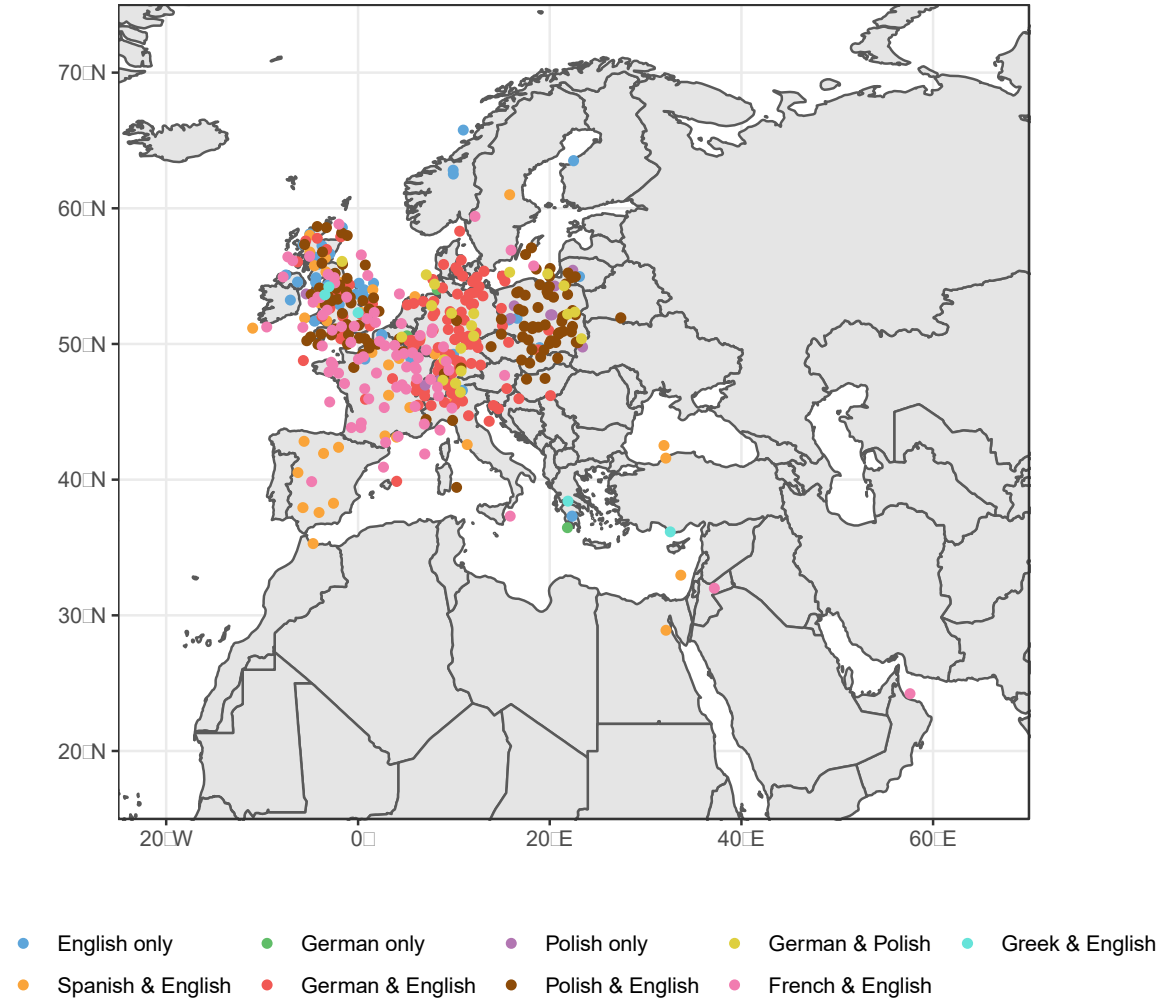


Figure 3.3 European map of respondents



As the research questions investigated as part of the present PhD work focus the experience of bilinguals, and older adults in particular, monolingual respondents (*n* = 146) were removed from the data for analysis. Demographics of the bilingual dataset are shown in Table 3.2. The data provides a mine of information into possible research questions such as: what role does cue language, encoding language and retrieval language play?; Are negative memories more likely to be recalled in LX?; Do migration patterns influence the

language memories are more likely to be recalled in?, etc. However, these lines of inquiry are beyond the current scope.

Table 3.2 Demographics of the bilingual dataset

Characteristic	**Bilingual respondents** N = 743[1]
Participant Age	34 (18, 85)
Gender	
Female	409 (55%)
Male	329 (44%)
Prefer not to answer	1 (0.1%)
Prefer to self-describe	3 (0.4%)
(Missing)	1
Education Level	
Primary school	1 (0.1%)
High school	159 (21%)
Vocational training/professional degree	50 (6.7%)
University degree	429 (58%)
Master's degree	31 (4.2%)
Doctorate	71 (9.6%)
(Missing)	2
Language Combination	
French & English	121 (16%)

German & English	161 (22%)
German & Polish	27 (3.6%)
Greek & English	9 (1.2%)
Polish & English	141 (19%)
Spanish & English	284 (38%)

[1] n (%); Mean (Range)

3.2.4 Analysis

As the outcome variables of interest, namely vividity and intensity ratings are ordinal in nature, metric analyses are not appropriate. Although in Psychology in particular, metric statistical models (e.g. *t*-tests, ANOVAs) are ubiquitous in analysing ordinal variables resulting from Likert-style scales, this practice is not recommended (Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018). Metric statistical methods assume that the dependent variables (or data generally) are on an interval or ratio scale, not an ordinal scale (Stevens, 1946). Not only can this distort effect-size estimates and inflate error rates, but it can also result in serious errors in inference (Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018). Liddell and Kruschke (2018) elegantly outline the various issues with such mistreatment of ordinal data, highlighting that statistical models which assume the use of metric outcome variables can result is distorted effect-size estimates, inflated Type I error rates, low rates of correct detection and even an inversion of difference between groups. This is due largely to the fact that response categories in ordinal data may not be equidistant (as in ratio or interval data). For example, a participant may find the difference between "very negative" and "negative" to be larger than that between "very positive" and "positive". Furthermore, the distribution of ordinal

data responses are often non-normal due to the extreme values (very high or very low) being frequently chosen (i.e., extreme response bias) (Bürkner & Vuorre, 2019).

Within the larger statistical literature, there are a variety of ordinal models each with their own names and conventions (Bürkner & Vuorre, 2019; Tutz, 2022). In Tutz (2022)'s taxonomy of ordinal regression models, basic ordinal models can be categorised within three distinct model classes: cumulative models, sequential models, and adjacent-category models. Cumulative models "assume that the observed ordinal variable, the opinion rating, originates from the categorization of a latent (not observable) continuous variable" (Bürkner & Vuorre, 2019, p. 78). As such, cumulative models are the most relevant class of ordinal models within psychological research wherein Likert items are used to gather discrete responses from an (assumed to be) underlying continuous psychological variable. Sequential models are more appropriate when assuming a latent continuous variable is not valid in the context of the ordinal variable in question (Tutz, 2022). For example, sequential models are better suited for a context where the ordinal responses are a result of a sequential or stepwise process wherein the higher response categories are feasible only after the lower categories are achieved or satisfied. Sequential models assume that for "for every category there is a latent continuous variable that determines the transition" between the successive categories (Bürkner & Vuorre, 2019, p. 81). Lastly, unlike sequential models, adjacent-category models (as per the name) are not hierarchical (Tutz, 2022). Adjacent-category models, common in item-response theory, "predict the decision between two adjacent categories using latent variables with thresholds" such that if the latent variable is less than its threshold, category k is selected, otherwise category $k+1$ (Bürkner & Vuorre, 2019, p. 82). In

summary, cumulative models compare groups of categories, sequential models compare categories and groups of categories, and adjacent-category models compare categories locally (Tutz, 2022).

As both vividness and intensity can be assumed to originate from latent continuous variables, cumulative ordinal models were selected. More specifically, due to the hierarchical or clustered nature of the data (thus creating more than one source of random error), hierarchical (or mixed) cumulative ordinal models with random intercepts by participant, cue word, and language were used to investigate the research questions. The use of mixed-models in addressing research questions in the study of autobiographical memory was first put forth by Wright (1998) and further advanced by Schrauf and Durazo-Arizu (2006) in order to best analyse the clustered data which typically results from the Galton-Crovits cueing technique. As Schrauf and Durazo-Arizu (2006) point out, by using other approaches such as linear models or ANOVAs, the researcher is "treating participants as the unit of analysis, the statistical procedure has subtly shifted the question from the emotional intensity of memories to the emotional intensity of participants" (p. 304). In the current study, the suggested random effects by participant were extended to encapsulate the more complicated grouping structure by language and cue as well. All CLMMs (cumulative link mixed model) were fitted with Laplace approximation, flexible thresholds, and a logit link function. Random intercepts by participant, language, and cue word were included in all models using data from the bilingual responses, with cue nested within language. All categorical covariates/predictors were included using treatment coding (also known as dummy coding). All numeric covariates (age-related variables) were mean-centered for interpretability (age of participant M = 33.61, age of memory M = 19.67). An interaction between age

of acquisition and recall language as well as age of the participant and recall language (fixed effects) was investigated for each model within the model building strategy. Overall, the approach to model building adopted follows Barr et al. (2013)'s suggestion of beginning with the maximal random effects structure justified for the study design, or "keeping it maximal". Assumptions and case diagnostics were checked for all reported models.

As multiple models were created for memories and imagined future events, a correction for multiple testing was needed to address the inflated alpha level. There are two primary means of making corrections for multiple testing to control for type I error rates (i.e., false positives) under a frequentist framework: family-wise error rates (FWER) and false discovery rates (FDR). FWER defined as the probability of incorrectly rejecting at least one null hypothesis (i.e., accepting the alternative hypothesis) when it is true (i.e., making a type I error) (Chen et al., 2017; Voelkl, 2019). Alternatively, FDR is concerned about the expected proportion of false positives (i.e., rejected true null hypotheses) amongst all of the rejected hypothesis (Benjamini et al., 2006; Benjamini & Hochberg, 1995; J. D. Storey, 2002). That is, FWER controls the probability of making any incorrect rejections of a null hypothesis at all while FDR controls the percentage of incorrect rejections. When the null hypothesis is true for all tested hypotheses, FWER and FDR are equivalent (Benjamini et al., 2006; Benjamini & Hochberg, 1995). FWER methods report corrected *p*-values (defined in Section 2.4.2) while FDR methods report *q*-values, which are is an analogue of the *p*-value incorporating corrections for multiple testing. A *p*-value provides the probability of the observed data, or more extreme data, assuming the null hypothesis is true (Lakens, 2022). Analogously, the *q*-value is the expected proportion of false positives among all results as or more extreme than what is observed (J. D. Storey, 2002). For example, when $\alpha = 0.05$,

a *p*-value of 5% (i.e., $p = 0.05$) means that 5% of all tests will result in false positives. That is, a *p*-value of 0.05 produces a 5% false positive rate among all truly null results. On the other hand, a *q*-value of 5% means that 5% of the significant results will be false positives.

FWER methods are widely used and suitable when "only a few or at most several tens of hypotheses are simultaneously tested" (Chen et al., 2017, p. 1727). When the number of tested hypotheses is in the several hundreds or thousands, FWER has been argued to be too conservative (Benjamini et al., 2006; Benjamini & Hochberg, 1995; Noble, 2009). As such, FDR more suitable for large-scale multiple testing such as in genomics research (Benjamini et al., 2006; Voelkl, 2019). In other words, FDR is better suited when testing many hypotheses in a similar way and the primary emphasis is to gather which is likely to be true, while FWER is preferable when testing the same hypothesis many times or in various ways (Voelkl, 2019). Therefore, for the present study the FWER approach was adopted.

The Holm (1979) sequential method for controlling family wise error rates was selected as it is more powerful the historically popular Bonferroni correction (Aickin & Gensler, 1996; Chen et al., 2017). Further, the Holm-Bonferroni method attempts to address criticisms the Bonferroni method has received for being too conservative and has comparatively lower increases in type II errors (i.e., false negatives) (Aickin & Gensler, 1996; Chen et al., 2017; Perneger, 1998). In fact, Aickin and Gensler (1996) definitively concluded that "there does not apparent to be any valid reason to continue using the Bonferroni procedure" (p. 728). Holm-Bonferroni corrected p-values are reported for all models.

Statistical analyses and data visualisations were conducted using R Version 4.1.3 (R Core Team, 2022), the ordinal (Christensen, 2019), ggplot2

(Wickham, 2016), patchwork (Pedersen, 2020), gtsummary (Sjoberg et al., 2021), and sjPlot (Lüdecke, 2021) packages.

3.3 Results

3.3.1 Memories

Descriptive data of the memory dataset are shown in Figure 3.4. Ratings of memory vividness and intensity are by the language of recall are shown in Figure 3.5. Overall, more memories were recalled in L1 (54%) than in LX (33.36%) or both languages (12.64%). However, the pattern of memory vividity and intensity ratings appears similar across recall language(s).

Figure 3.4 Descriptive data of the memory dataset

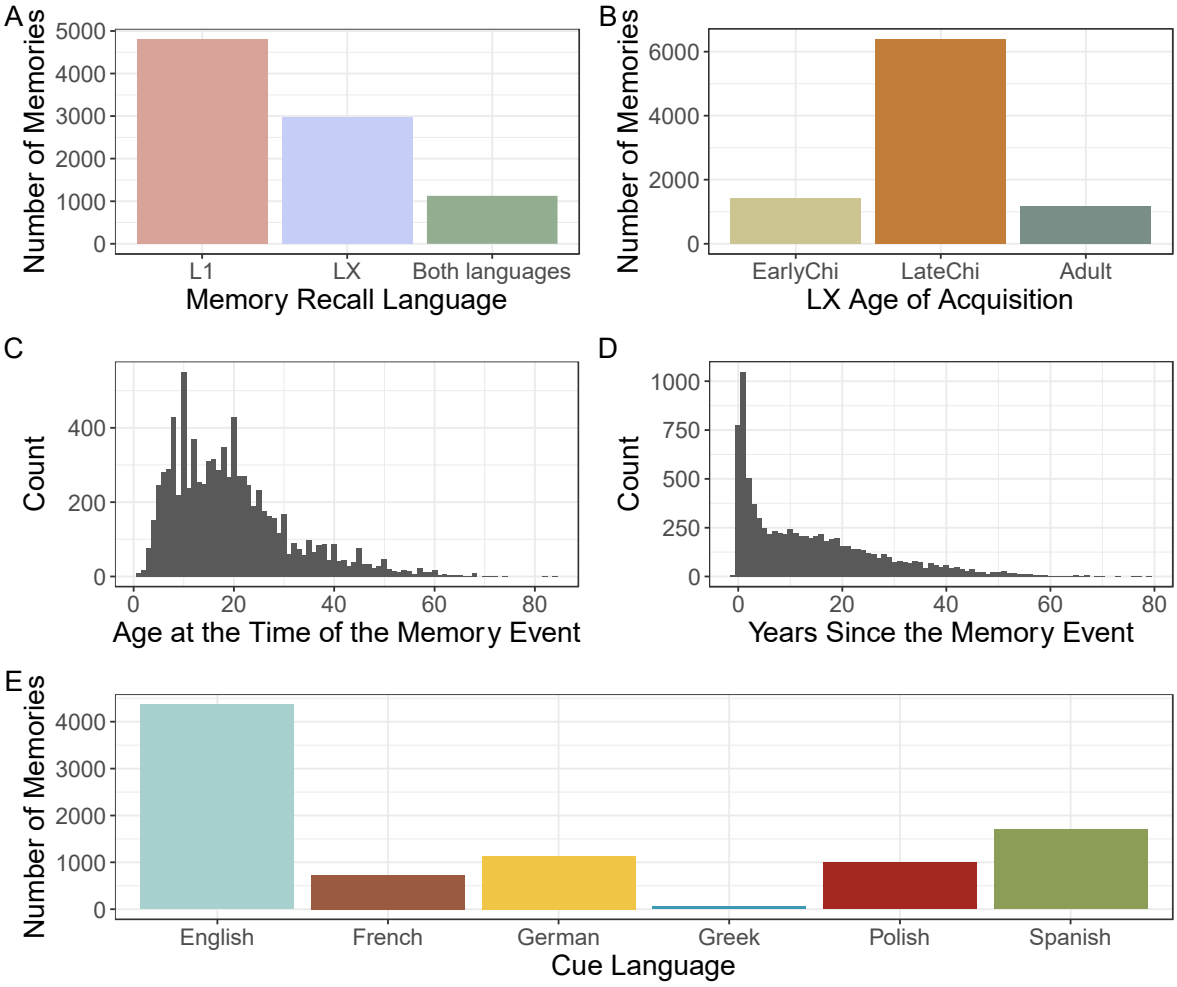
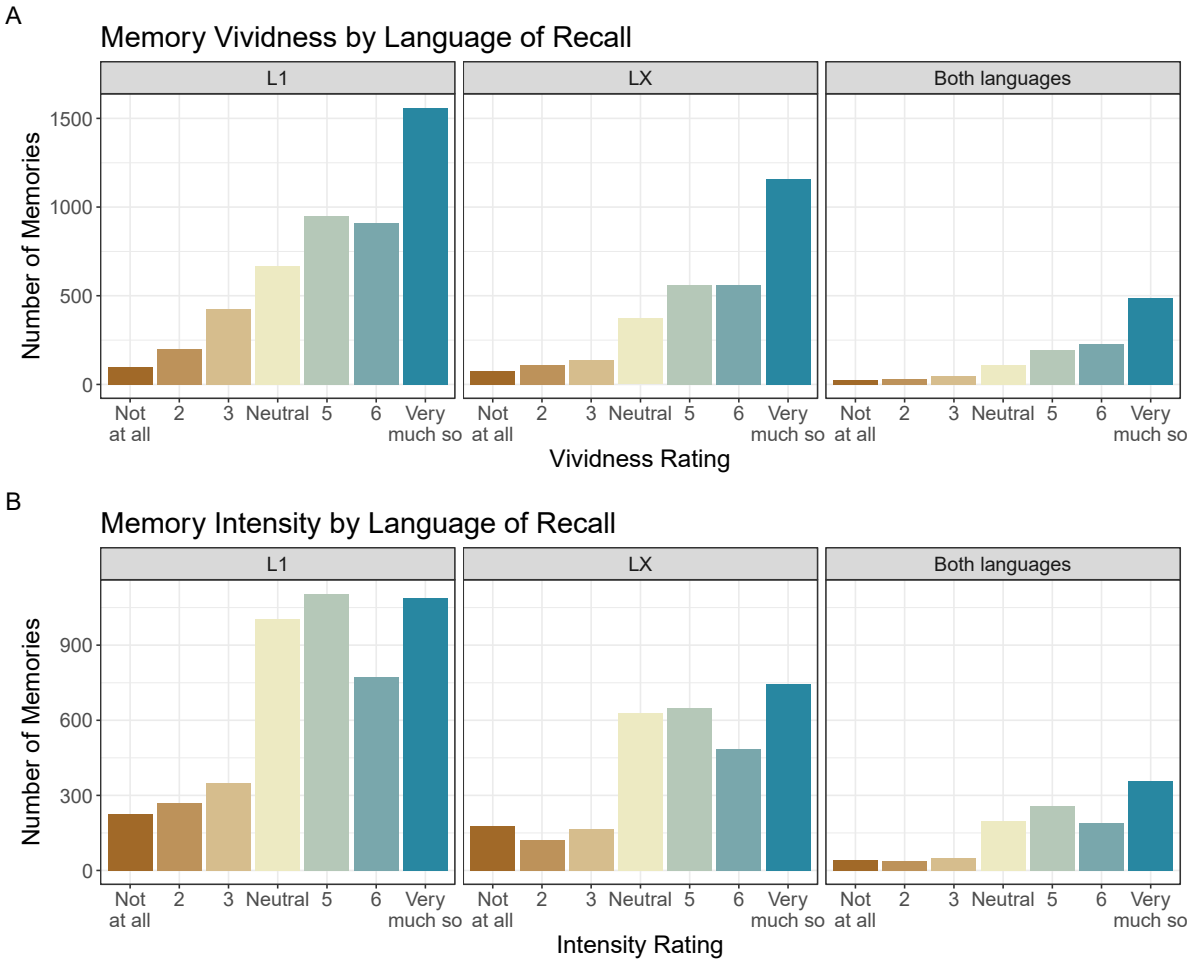


Figure 3.5. Memory vividness and intensity ratings by language of recall



3.3.1.1 Vividity ratings

Table 3.3 shows the model summary for the best fitting model of vividity ratings of memories. The language of recall of the memory was not a significant predictor (p = .30), nor was the age of acquisition of LX (p = .43) when controlling for the influence of the relative age of the memory encoding, participant age, and intensity ratings. However, the age of the memory (i.e., years since the memory event occurred; p = < .001), participant age (p = < .001), intensity ratings (p = < .001) did significantly predict vividity ratings, after

adjusting for the influence of age of acquisition and recall language. Older participants showed an increased odds (5%) of rating vividity of memories higher (see Appendix B for visualisations). Older memories have a decreased odds (6%) of rating the vividity of the memory higher (see Appendix B for visualisations). Similarly, those who rated the memory as intense have an increased odds of also rating the vividity higher (see Appendix B for visualisations).

The ICC (intraclass correlation coefficient, which can be used to understand how much variable is explained by the random effects) for this model is 0.32. The ICC gives information on the level of explained variance in the model, specifically as it relates to the random effects structure or the "the proportion of the variance explained by the grouping structure in the population... the proportion of group-level variance compared to the total variance. The intraclass correlation can also be interpreted as the expected correlation between two randomly drawn units that are in the same group" (Hox et al., 2017, p. 13).

Table 3.3 Memory vividity model

Characteristic	**OR[1]**	**95% CI[1]**	**p-value**	**adj. p-value[2]**
Recall Language				
L1	—	—	—	—
Both languages	1.15	0.99, 1.35	.072	.30
LX	1.12	0.99, 1.27	.072	.30
Age of Acquisition LX				
Early Childhood	—	—	—	—
Late Childhood	0.88	0.69, 1.13	.40	.40
Adulthood	0.80	0.57, 1.12	.20	.40
Age of Memory	0.94	0.94, 0.95	**<.001**	**<.001**
Participant Age	1.05	1.04, 1.06	**<.001**	**<.001**
Intensity Rating	2.35	2.27, 2.43	**<.001**	**<.001**
Observations	8,605			
ICC	.32			
Marginal R^2 / Conditional R^2	.37 / .57			

[1] OR = Odds Ratio, CI = Confidence Interval

[2] Holm-Bonferroni correction for multiple testing

3.3.1.2 Intensity ratings

Table 3.4 shows the model summary for the best fitting model of intensity ratings of memories. The language of recall of the memory was not a significant predictor (p = .60), nor was the age of the memory (p = .40) when controlling for the influence of age of LX acquisition, participant age, and vividity ratings. However, the age of acquisition of LX (p = .022), participant age (p = < .001), and vividity ratings (p = < .001) did significantly predict intensity ratings, after adjusting for the influence of age at the time of the memory and recall language. Compared to those who acquired LX in early childhood, acquiring LX in adulthood is associated with odds that are 58% higher of rating the memory as more intense (see Appendix C for visualisations). Older participants showed an increased odds (2%) of rating the intensity of memories higher (see Appendix C for visualisations). Similarly, those who rated the memory as intense have an increased odds of also rating the vividity higher (see Appendix C for visualisations). The ICC for this model is 0.25.

Table 3.4 Memory intensity model

Characteristic	**OR[1]**	**95% CI[1]**	**p-value**	**adj. p-value[2]**
Recall Language				
L1	—	—	—	—
Both languages	1.08	0.93, 1.26	.30	.60
LX	1.04	0.92, 1.18	.5	.60
Age of Acquisition LX				
Early Childhood	—	—	—	—
Late Childhood	1.18	0.93, 1.50	.20	.50
Adult	1.58	1.14, 2.19	**.006**	**.032**
Age of memory	1.00	1.00, 1.01	.11	.40
Participant Age	1.02	1.01, 1.02	**<.001**	**<.001**
Vividity Rating	2.42	2.34, 2.51	**<.001**	**<.001**
Observations	8,605			
ICC	.25			
Marginal R^2 / Conditional R^2	.32 / .50			

[1] OR = Odds Ratio, CI = Confidence Interval

[2] Holm-Bonferroni correction for multiple testing

3.3.2 Imagined future events

Descriptive data of the imagined future events dataset are shown in Figure 3.6. Ratings of imagined future event vividness and intensity are by the language of recall are shown in Figure 3.7. Overall, unlike memories, imagined future events were equally likely to be experienced in LX (40.46%) or L1 (40.38%), followed by both languages (19.15%). The pattern of imagined future event vividity and intensity ratings suggests that events experienced in L1 and LX are overall more vivid and intense than those experienced in both languages. Further, events imagined in LX appear to be rated more intense and vivid overall.

Figure 3.6 Descriptive data of imagined future events

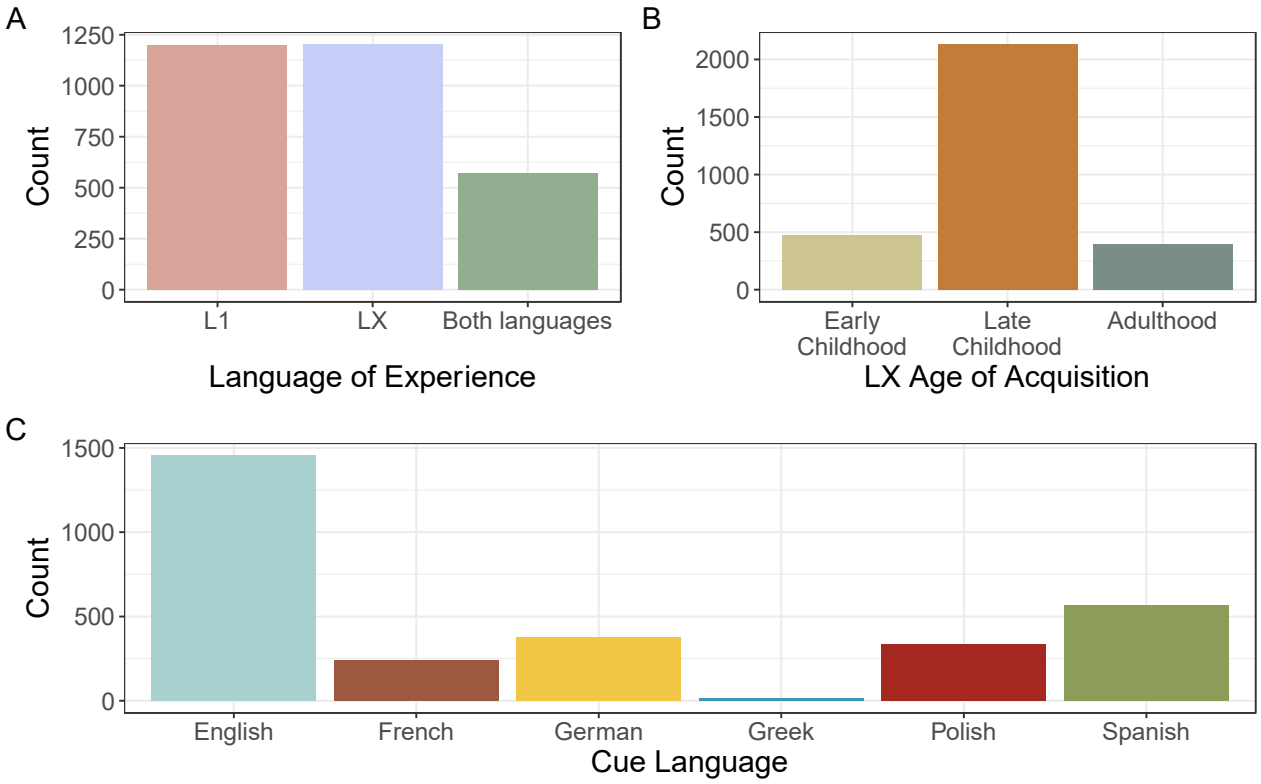
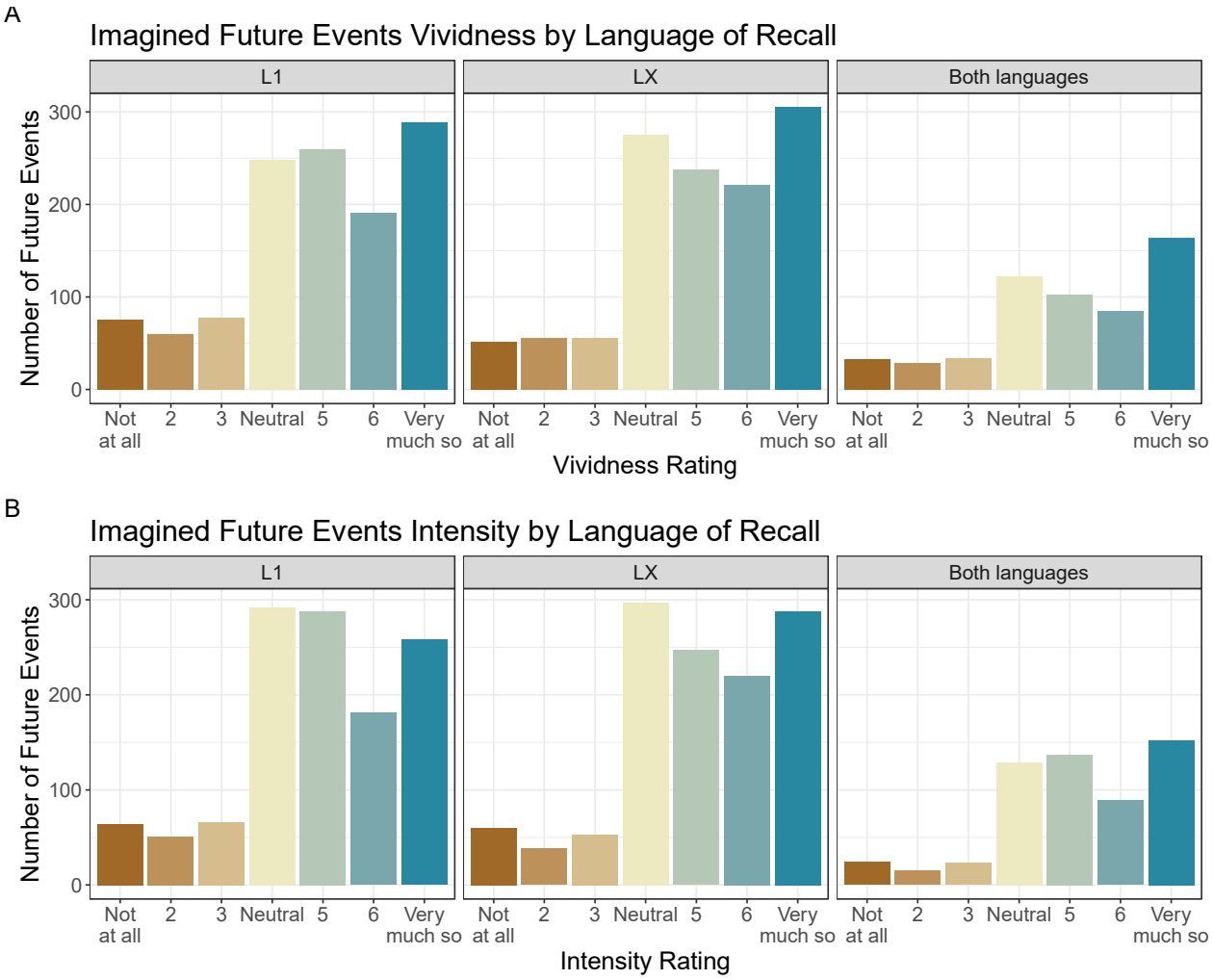


Figure 3.7. Imagined future event vividness and intensity ratings by language of recall



3.3.2.1 Vividity ratings

The cue nested within language random grouping variable did not explain any variance in the data and thus was removed in an iterative fashion from the random effects structure (first language and then cue). The final model includes a random intercept by participant only in the random effects structure.

Table 3.5 Future events vividity model

Characteristic	**OR[1]**	**95% CI[1]**	**p-value**	**adj. p-value[2]**
Recall Language				
L1	—	—	—	—
Both languages	0.85	0.68, 1.07	.20	.70
LX	1.00	0.82, 1.21	>.90	>.90
Age of Acquisition LX				
Early Childhood	—	—	—	—
Late Childhood	0.98	0.73, 1.33	>.90	>.90
Adulthood	1.09	0.72, 1.65	.70	>.90
Participant Age	1.01	1.00, 1.02	**.005**	**.025**
Intensity Rating	2.81	2.64, 2.99	**<.001**	**<.001**
Observations	2,972			
ICC	.26			
Marginal R^2 / Conditional R^2	.39 / .55			

[1] OR = Odds Ratio, CI = Confidence Interval

[2] Holm-Bonferroni correction for multiple testing

Table 3.5 shows the model summary for the best fitting model of vividity ratings of future events. The language of experience was not a significant predictor (p = .91), nor was the age of acquisition of LX (p = .97) when controlling for the influence of participant age and intensity ratings. However, participant age (p = .025) and intensity ratings (p < .001) did significantly predict vividity ratings, after adjusting for the influence of age of acquisition and recall language. Older participants showed an increased odds (1.2%) of rating vividity of imagined future events higher (see Appendix D for visualisations). Similarly, participants who rated the imagined future event as intense have an increased odds of also rating the vividity higher (see Appendix D for visualisations). The ICC for this model is 0.26.

3.3.2.2 Intensity ratings

Similar to the imagined future event vividity model, the maximal random effects structure resulted in a singular model. Upon inspection, language did not explain any variance and was removed, resulting in the fitting model with a random intercept by participant and by cue.

Table 3.6 Future events intensity model

Characteristic	OR[1]	95% CI[1]	p-value	adj. p-value[2]
Recall Language				
L1	—	—	—	—
Both languages	1.30	1.05, 1.62	**0.017**	**0.048**
LX	1.08	0.90, 1.30	0.4	>.90
Age of Acquisition LX				
Early Childhood	—	—	—	—
Late Childhood	1.03	0.79, 1.35	0.8	>.90
Adult	1.14	0.79, 1.64	0.5	>.90
Participant Age	1.00	0.99, 1.00	0.4	.70
Vividity Rating	2.62	2.47, 2.78	**<0.001**	**<0.001**
Observations	2,972			
ICC	.21			
Marginal R^2 / Conditional R^2	.40 / .52			

[1] OR = Odds Ratio, CI = Confidence Interval

[2] Holm-Bonferroni correction for multiple testing

Table 3.6 shows the model summary for the best fitting model of intensity ratings of future events. Participant age (p = .70) and age of acquisition of LX (p = .85) were not a significant predictor when controlling for the influence of language of experience, age of acquisition of LX, and vividity ratings. However, retrieval language (p = .048) and vividity ratings (p < .001) were significant predictors after adjusting for participant age and age of LX acquisition. Future events experienced in both languages had an increased

odds of 1.30 of rating vividity higher compared to L1 (see Appendix E for visualisations). Furthermore, participants who rated the imagined future event as more vivid have an increased odds of also rating the vividity higher (see Appendix E for visualisations). The ICC for this model is 0.21.

3.4 Discussion

"Autobiographical memory is saturated with language", even in the context of monolinguals, but particularly as it relates to multilinguals (Marian & Neisser, 2000, p. 366). Words are used to express the contents of memories, which themselves likely are engrossed in language through conversation, music, or other linguistic elements. This study set out to investigate if memories and imagined future events were more vivid when recalled in L1 over and above LX or both languages in bilingual adults across the lifespan. The fact that vividness and emotional intensity did not behave in the same way in our analysis validates not only asking about them separately, but also treating them independently to answer the research questions. Indeed, theoretically the two concepts map onto related but distinct elements identified by Rubin (1998) as fundamental elements of autobiographical memory: imagery (vividity) and affect (emotional intensity).

While more memories were recalled in L1 (51%), these memories were not more vivid than those recalled in L2 (33.36%) or both languages (12.64%), when controlling for the effect of the age of LX acquisition, age of the participant, age of the memory, or vividity ratings. However, more recent events (memories of events which occurred later in life) were more vivid. Similarly, memories which were rated higher in emotional intensity had an increased odds of being more vivid in rating. Interestingly, relating to the age of participants, older participants had higher odds of rating memories high in

vividity. The relationship between the age of the participant and ratings of memory emotion intensity followed the same pattern. Older participants had an increased odds of rating memories as more emotionally intense. Further, participants who acquired LX in adulthood had an increased odds of rating the recalled memories as more emotionally intense compared to those who acquired LX in early childhood. As with the vividity ratings of memories, the recall language (L1, LX, or both languages) was not a significant predictor, though memories which were more vivid had an increased odds of being rated as emotionally intense, and vice versa.

These results suggest that, especially as it relates to the vividness of memories, there are two opposing tendencies which are outweighing each other: an advantage in a favour of L1 wherein more memories are likely to be recalled and a recency effect favouring memories retrieved in LX and both languages. That is, while more memories were likely to be recalled in L1, the memories which were more recent were more vivid, and more likely to be recalled in LX or both languages. Indeed, descriptively memories retrieved in L1 were about events which took place an average of 15 years ago (*M* = 15.54, *SD* = 25, *Mdn* = 13), whereas the memories recalled in LX (*M* = 11.82, *SD* = 12.7, *Mdn* = 8) and both languages (*M* = 12.18, *SD* = 13.35, *Mdn* = 7) were of more recent events.

The vividity and emotional intensity of imagined future events present a slightly different picture. In terms of recall language, future autobiographical events were not experienced at a higher proportion in L1. Rather, imagined future events were equally likely to be experienced in L1 (40.38%) and LX (40.46%), followed by both languages (19.15%). Interestingly, the random effects structure for the models of imagined future events differed from that of memories. While both the vividity and intensity ratings models for memories

fit well with random intercepts by participants and by cue nested within language, the specific language (i.e., English, French, Spanish, Polish, German, or Greek) explained negligible random variance (i.e., the models were singular) in either future events model over and above that of the individual participant. The vividity of imagined future events was not significantly predicted by recall language, though, as with memories. Instead, the future events that were rated as more emotionally intense had an increased odds of being rated higher in vividity, independent of the language in which the future event was experiences (e.g., L1, LX, or both languages). This finding diverges from some literature which suggests that the use of LX reduces mental imagery. For example, Hayakawa and Keysar (2018) concluded that the use of a foreign language, or LX, significantly reduces the vividness of mental imagery (in this case specifically in the context of sensory experiences) as rating on a 7-point Likert scale compared to L1.

In the present study, relating to participant age, older participants had an increased odds of rating future events as more vivid. This is a similar pattern observed in the autobiographical memories data, wherein older participants had an increased odds of rating memories higher in vividity. While the temporal aspect of future events was not measured within this study, these findings offer an interesting point for future research. For example, do younger people imagine events in the more distant future while older adults imagine event in a more immediate future? Similarly, an interesting avenue of future research is if imagined future events follow a similar temporal gradient found within the memories collected in the current study wherein more recent events are more vivid. If so, this may explain the current findings that older participants experienced more vivid imagined future events, perhaps as the

events they imagined are in a more immediate future than those imagined by younger participants.

Emotional intensity of imagined future events presented another interesting set of findings. While the age of participants was not a significant predictor, this was the only model in which recall language significantly predicted ratings. Running contrary to the Mother Tongue Hypothesis, future autobiographical events which were recalled in both languages had an increased odds of being rated higher in emotional intensity than those recalled in L1.

The present study is novel in investigating both autobiographical memories and imagined future events in bilinguals across the lifespan, all within the same sample. Spreng and Levine (2006) gathered data on memories and future events in their study, their sample included (presumably) monolingual speakers. In fact, the language background of the sample was neither reported nor discussed, meaning language was not treated as a relevant modifying variable. Further, while some studies with monolingual participants have larger sample sizes (i.e., Janssen and colleagues (2006) with 1,579 Dutch speakers and (2011) with 2341 Dutch speakers), most bilingual autobiographical memory studies include a sample 30 or fewer bilinguals. Not only does the present study includes 765 bilingual participants, but also a large sample of diverse bilinguals with a variety of language backgrounds and language combinations. Additionally, the present study is novel in applying the Galton-Crovitz cueing technique in a bilingual sample online.

Partially replicating Shrauf and Rubin (2004) who found LX memories to be more intense and emotional among Spanish-English bilinguals aged 60 plus, the present study found, regardless of the language of recall, older adults had an increased odds of recalling more emotionally intense and vivid

memories. Not explored in Shrauf and Rubin (2004) were imagined future events, which showed a similar pattern to memories in terms of participant age. Older adults had an increased odds of experiencing more vivid imagined future events, regardless of the recall language. In terms of emotional intensity of imagined future events, participant age was not a significant predictor.

Relating to the research questions guiding the present work, the Mother Tongue Hypothesis within the measured bilingual autonoetic consciousness is largely unsupported. Controlling for the effect of the other variables of interest in both memories (i.e., the age of the memory, participant age, language of recall, age of acquisition of LX, and emotional intensity/vividity ratings) and imagined future events (i.e., participant age, language of recall, age of acquisition of LX, and emotional intensity/vividity ratings), the language of recall was a significant predictor only in relation to the emotional intensity of imagined future events wherein events internally experience in both languages were more emotionally intense than those experienced in L1. While indeed memories were mostly recalled in L1, the emotional intensity and vividity of these memories was not significantly different from those recalled in LX or both languages. Episodic future thinking, another expression of autonoetic consciousness, was not more likely to be experienced in L1 in this sample. In fact, in the data set, three more imagined future events were experienced in LX (n = 1203) above L1 (n = 1200).

In conclusion, in the context of autobiographical memory and imagined future events, the current study suggests there is not a systematic dominance of L1 in terms of vividness or emotional intensity in expressions of autonoetic consciousness. That is, a dynamic rather than static pattern of language across the multilingual lifespan.

3.5 Language changes across the lifespan in healthy and pathological ageing (work in progress)

Building from this, a further study was designed and initiated looking at the pattern of reported language changes in older age. In conjunction with the study of autobiographical memory, this study extends the means by which language use across the lifespan can be empirically explored by comparing patterns of change in healthy and pathologically ageing older adults. As data collection is still ongoing and the survey is actively being developed into additional languages, it is not presented as a complete empirical chapter within this thesis. Preliminary results were presented as a poster at the Conference of Multilingualism 2020 and a talk at The International Online Workshop on Language in Healthy and Pathological Aging 2021. Further, I was interviewed as part of the BBC 4 Radio programme "Lost for Words" around this study (*Lost for Words*, 2021)*.*

Language attrition, that is the loss of L1/shift to L2 in L2 environment, is a relatively well-studied phenomenon (de Bot & Weltens, 1995; Hermans, 2012; Schmid & Dusseldorp, 2010; Wei, 2014). In contrast, few studies have systematically studied the opposite process: so-called language reversion, that is loss of L2/shift to L1 by aging migrants in an L2 environment. The experience of reverting to L1 in an L2 environment is a widely accepted phenomenon among migrant communities and multilingual families, but the empirical backing and foundation is lacking and what does exist is inconclusive (de Bot & Weltens, 1995; Keijzer, 2011). Some research has suggested that the experience of language reversion is driven by social phenomena such as nostalgia, factors in the acquisition of L2, or is strongly linked to neuropathology such as dementia (A. Costa et al., 2012; de Bot & Clyne, 1989; Hyltenstam & Stroud, 1993; Schmid & Keijzer, 2009; Tipping & Whiteside, 2015).

3.5.1 Methods

In order to systematically investigate the (a) prevalence of language reversion and other language changes in ageing multilingual communities and (b) if reported changes are related to neuropsychological, social, or acquisitional factors, an international online survey was developed in June 2019. Originally in English, the survey was later translated into Spanish, Mandarin Chinese using simplified characters, Mandarin Chinese using traditional characters, and Tamil. The survey has recently also been developed in Greek in February 2022 and Italian in November 2022. As additional languages are still being included to allowing for an even greater range of participant responses, the survey remains open until the number of responses are comparable across languages. All translations are included on the same link for ease of dissemination of the survey and to prevent the link expiring from previous posts or shares. To allow for varied response types beyond the language of the survey, 4 versions of the survey were developed:

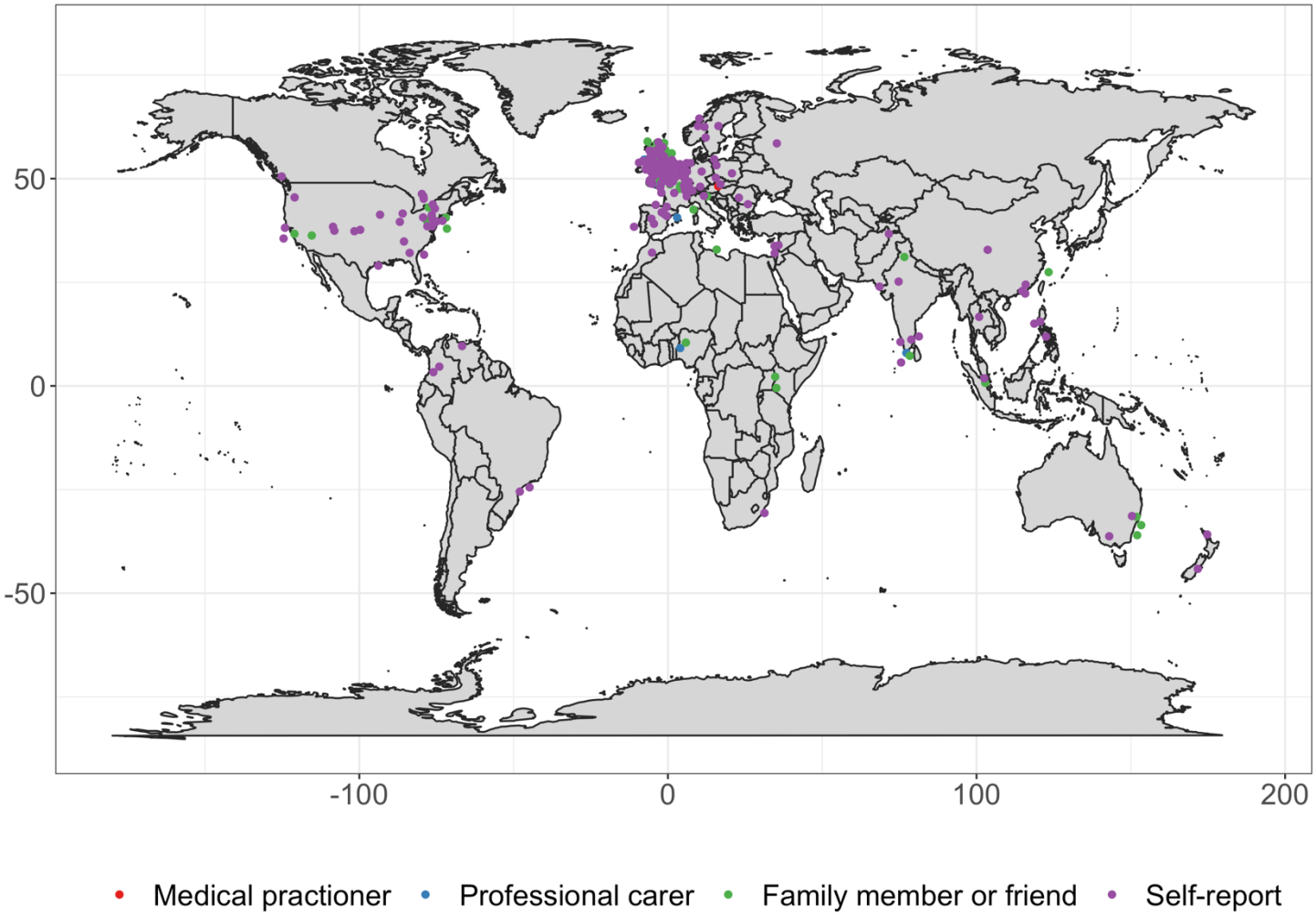
- self-report
- proxy report by a family member or friend
- proxy report by a professional carer
- report by a medical professional.

The medical practitioner version contains only 4 questions around the frequency in which they have experienced or heard about patients reverting to their first language while the self-report and proxy surveys are more detailed. Proxy respondents include carers reporting on behalf of a care-receiver and family/friends reporting on behalf of a family member or close friend who experience(d) language changes in later life. While there are limitations to self-report, in particular in the case of dementia (Kiyak et al., 1994), a survey methodology including proxy responses allows for a systematic exploration of

the prevalence of differing forms of language change in older age. Guardian Carers, an organisation based in London which specialises in providing at-home bilingual care for older adults (*Guardian Carers - Who Are We*, n.d.), advised throughout the survey development process. Ethics was approved by the University of Edinburgh PPLS Research Ethics Committee (ref: 236-1819/6).

As of August 2022, 342 completed responses have been collected (Figure 3.8). In terms of survey language, 89.5% of responses are in English, 4% in Spanish, 6% in Chinese (including both simplified and traditional characters), and 0.5% in Tamil. In terms of response types, the majority of responses to date are self-report (61.2%), followed by proxy reports (30.3% overall; 25% from family members or friends; 5.3% from professional carers), and medical practitioner responses (3.2%).

Figure 3.8 World map of survey respondents

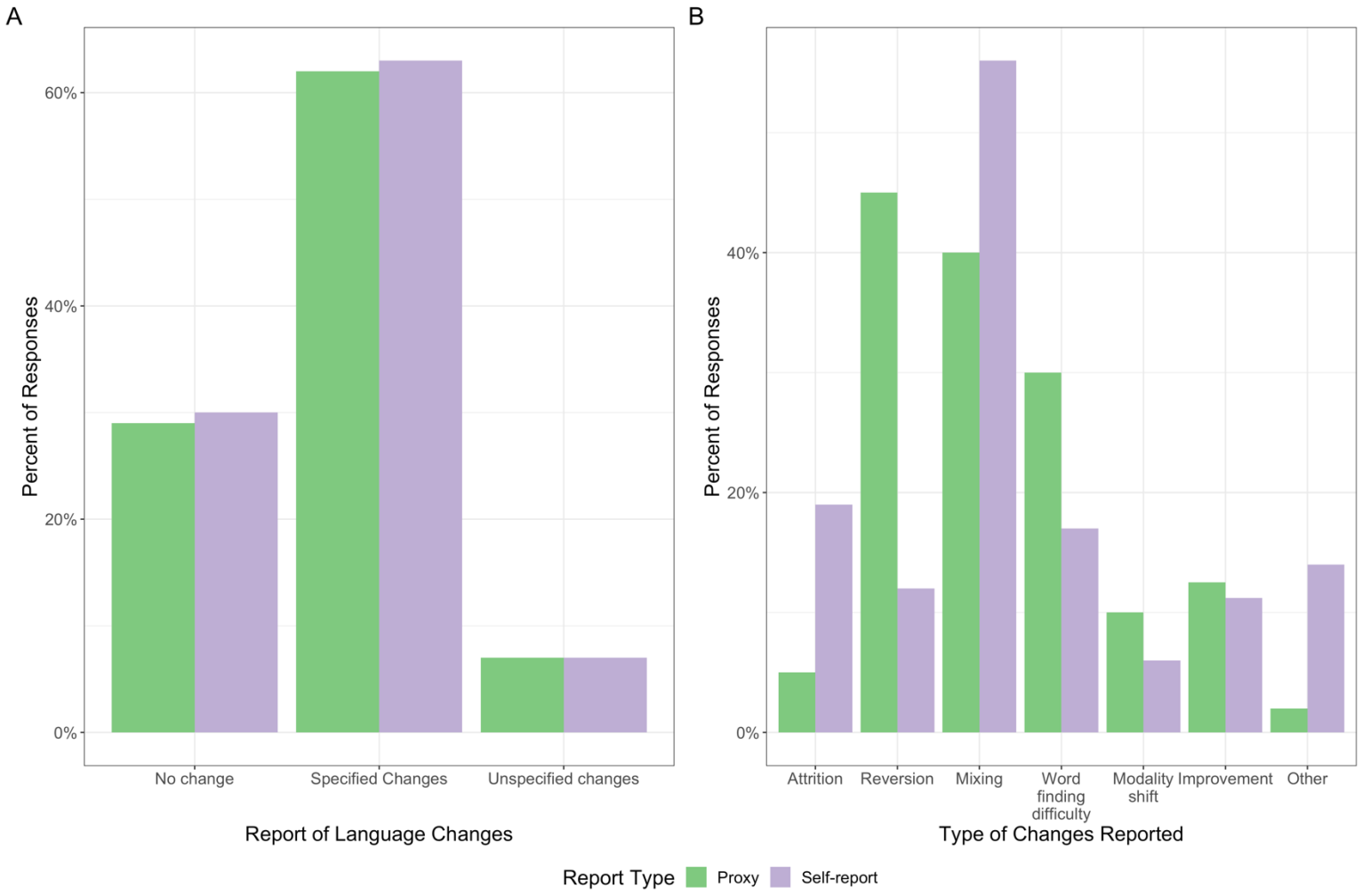


3.5.2 Preliminary results and discussion

Preliminary descriptive results show that a return to the use of the first language is indeed reported by proxy and self-report; though it is neither the only nor the most frequent pattern observed (Figure 3.9). In terms of the profile of proxy reporters, the majority (52%) are the children of the person they are reporting about. Other relations include grandchild, niece/nephew, spouse, cousin, son/daughter-in-law, friend, and neighbour. Fifty-six percent of responses can be categorised as healthy ageing adults as no diagnosis was reported. The most prevalent diagnosis reported was dementia (17%), followed by clinical depression (7%), stroke (3%) and Parkinson's Disease (3%).

Fourteen percent of responses reported another diagnosis, including cancer, hydrocephalus, diabetes, and hypertension. Observed changes in language use were reported in roughly 60% of responses by both self-report and proxy. Notably, within the present dataset there is also a portion of responses with no reported changes. As the study design relies on participant reporting, these responses reporting no change suggest that the sample collected thus far is not entirely biased towards reports of significant language changes.

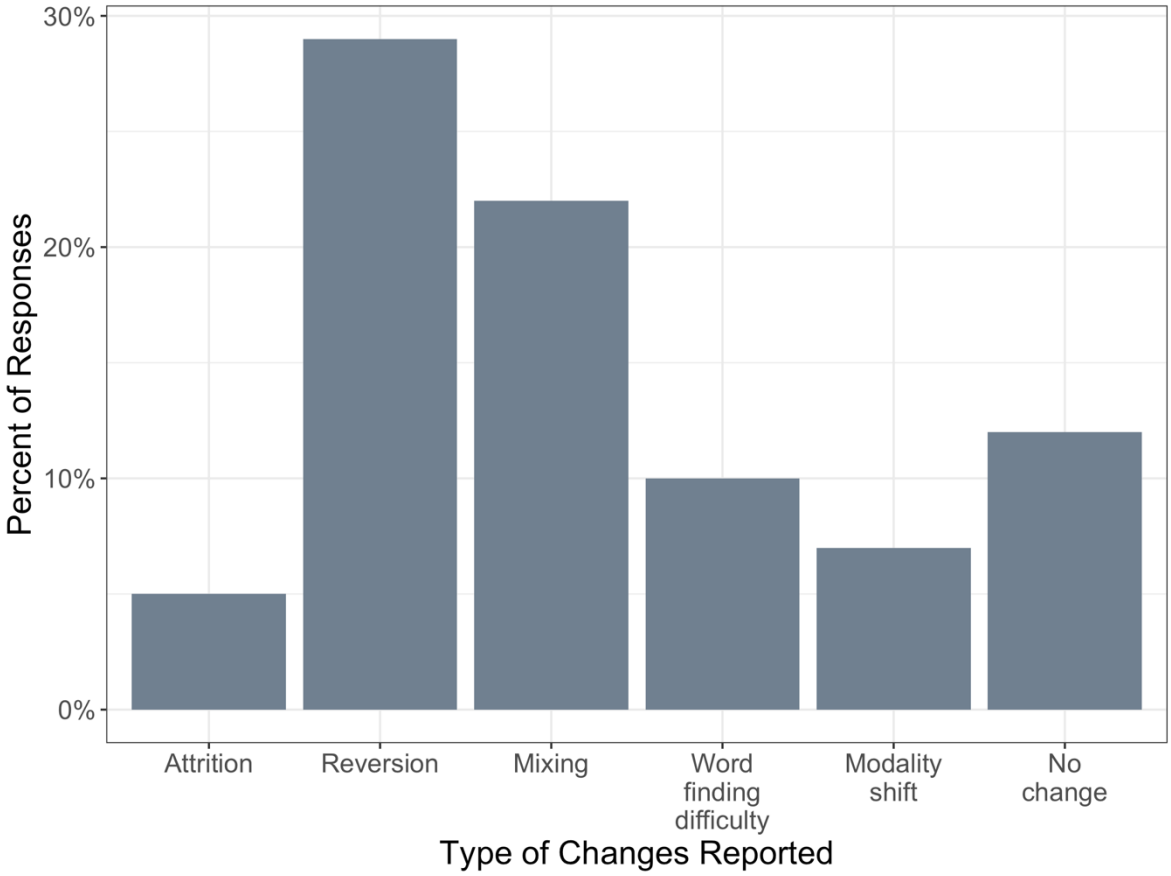
Figure 3.9 Language changes reported by survey type



Note: "Other" in Figure B includes reported changes of accent, environment changes leading to changes in language dominance, change in language preferences, and increased metalinguistic awareness.

As seen in Figure 3.9B, reports of language reversion were far more common by proxy while reports of language mixing were more common in the self-report data. Further, not all language changes were negative, for example some participants reported that with age they found an improvement in proficiency and language use or increased metalinguistic awareness. Looking further into the reports of language reversion, 40% were among people diagnosed with dementia. Indeed, within the subsample with a dementia diagnosis, reversion was the most common pattern of language change reported, followed by language mixing (Figure 3.10). Further, all cases of language mixing within the sample diagnoses with dementia were reported as unintentional mixing.

Figure 3.10 Language changes reported with dementia diagnosis



While the study is still ongoing, these preliminary results suggest that (a) language changes in older age are common, but not universal and (b) there are different patterns of change, with a reversion to L1 being a frequent but by no means the universal pattern. It is not the case within this sample that all participants revert to L1 while in an L2 environment in older age. These preliminary findings go against the assumption that L1 is always dominant.

Notably, reports of reversion to an earlier acquired language is the most common in dementia, wherein an increase in involuntary language mixing is also frequently reported. Interestingly, the majority of both reports of a clinical diagnosis of dementia and language reversion are found in the proxy responses. One of the factors influencing this discrepancy in proxy versus self-report responses could be that children and grandchildren often do not know the first language of their parents or grandparents. As noted, the 52% of proxy responses are from children reporting on behalf of their parents. It could be that the individual is aphasic or mixing in both languages, but if the proxy reporter only knows one of the languages, they would not be able to effectively assess the level of impairment in L1. Impairment in LX would be much more obvious and notable. That is to say, perhaps family members interpret language changes as reversal into languages they do not know themselves. This hypothesis is supported by the comments from many proxy respondents reporting that after their parents or grandparents immigrated to another country, they did not teach or pass on their first language. This phenomena is also supported by David Shariatmadari's account of his father's language changes in the BBC programme *Lost for Words* (2021).

In conjunction with the study of autobiographical memory, the natural history studies within this thesis support not only a dynamic character of language across the lifespan, but also patterns of individual variation which

necessitate a broad methodological spectrum (as opposed to just averaged group comparisons) in order to adequately examine the research questions under investigation.

Chapter 4 Learning Spanish Across the Lifespan: A Comparison of Classroom and Synchronous Online Learning

Having established the dynamic nature of language across the lifespan and the importance of individual differences, we now shift to practical studies of language (i.e., the second aspect of the title of the present work 'effects of language learning.' This first language learning chapter focuses on Spanish as it is the most common language studied in the UK which has not been traditionally taught in schools.

4.1 Introduction

4.1.1 Bilingualism and cognition

A contentious line of study in psychology and linguistics is the question of whether learning and using more than one language influences cognitive functions. Specially, the literature around this topic focuses on whether or not bilingual people show enhanced cognitive control compared to monolinguals. Historically, the evaluation of and consensus on the impact of multiple languages on cognitive functioning (particularly among children) has been on a pendulum, with some researchers finding it to have a positive impact and others a decidedly negative influence (for an in-depth review of the tumultuous literature around bilingualism and non-linguistic cognitive functions, see Mattschey, 2020). In the early 20th century, the predominate view held that bilingualism was damaging to a child's development (i.e., "bilingual disadvantage" theory), resulting in "mental confusion" (Saer, 1923) or even "mental retardation" (Darcy, 1963). In their seminal paper, Peal and Lambert (1962) brought this into question by concluding that "bilinguals appear to have

a more diversified set of mental abilities than monolinguals", fuelling the much debated "bilingual advantage" theory (p. 22). Much of the subsequent literature initially focused on "classic" bilinguals, that is bilinguals who acquired their second language in childhood and have a balanced command of both languages.

The findings on the topic are mixed, with some studies finding evidence for improved cognitive functioning in bilinguals (Bialystok et al., 2010; Bialystok & Viswanathan, 2009; Kovács & Mehler, 2009; Prior & Macwhinney, 2010) and others finding no difference between monolinguals and bilinguals (Paap et al., 2014, 2015, 2017; Paap & Greenberg, 2013). As noted by de Bruin et al. (2015) while few studies show evidence for a bilingual cognitive disadvantage, there is evidence of a publication bias in the literature towards studies fully supporting or contradicting the "bilingual advantage" theory. Accordingly, studies with polar results (i.e., finding there to be a difference between bilinguals and monolinguals, whether positive or negative), were more likely to be published across 169 conference proceedings from 1999 to 2012 (de Bruin et al., 2015). Indeed, a clear bias towards the publication of statistically significant results, in psychology as well as in other sciences, has long been flagged in the wider academic literature (Sterling, 1959; Sterling et al., 1995). A variety of variables have been suggested to influence the mixed findings in the literature on bilingualism and cognition, including age, socio-economic status, immigration status, cultural transferability of the cognitive tests used, and the operationalisation (via cognitive measures) and categorisation of cognitive functioning and bilingualism itself (Bak, 2016; van der Ploeg et al., 2020). In a recent meta-analysis of 170 studies, Ware and colleagues (2020) found evidence for a "bilingual advantage" which was moderated by age and the task under study. Samples with older adults compared to younger adults found

more evidence of improved cognitive functioning in bilinguals compared to monolinguals, specifically in the context of response times on incongruent trials. Further, bilinguals had increased speed and accuracy above monolinguals particularly in the Attention Network, Stroop, and Simon tasks (A. T. Ware et al., 2020).

Such a cognitive advantage has been suggested to be due to parallel activation of both languages in a bilingual mind/brain. During comprehension and production of language, bilinguals are thought to experience parallel activation of the languages they know (Beauvillain & Grainger, 1987; Colomé, 2001; A. Costa, 2005; de Groot et al., 2000; Kroll et al., 2008). Therefore, bilinguals are suggested to recruit domain general cognitive functions (e.g., attentional processes) to manage these competing linguistic systems and effectively select the appropriate language given the current linguistic environment (Kroll et al., 2014). This parallel activation theory has been supported by both eye-tracking and fMRI studies (Marian et al., 2003; Marian & Spivey, 2003). It is through constant monitoring of the linguistic environment, suppression/inhibition of the non-target language(s) and switching as the linguistic environment changes that these functions are thought to become more efficient, resulting in better cognitive control and performance, even in non-linguistic tasks (Bialystok, 1999, 2010, 2011; D. W. Green, 1998; D. W. Green & Abutalebi, 2013; Prior & Macwhinney, 2010).

Early models of bilingualism, such as the Inhibitory Control and the Bilingual Interactive Activation Models focused on inhibitory control as a fundamental cognitive function recruited for this management of cross-linguistic interference (Dijkstra et al., 1998; D. W. Green, 1998). Indeed, neuroimaging data supports bilingual language production as a dynamic inhibition process, recruiting both cortical and subcortical structures, in order

to resolve competition between languages and effectively select the target language (Abutalebi & Green, 2007). According to the Inhibitory Control Model, there are multiple levels of control including lower-level, automatic control for previously performed tasks and the higher-order, supervisory attentional system used in novel tasks (D. W. Green, 1998). Inhibition under this model is a reactive mechanism executed by the supervisory attentional system (Long et al., 2020). Reactive inhibition, therefore, becomes more efficient through continual competition, control, and switching between language schemas in the case of a bilingual (D. W. Green, 1998; Long et al., 2020).

In response to more recent findings in the bilingualism literature which the Inhibitory Control Model could not sufficiently explain (e.g., increased reaction time of bilinguals on non-linguistic tasks (A. Costa et al., 2008; Martin-Rhee & Bialystok, 2008) and in high though not low monitoring conditions (A. Costa et al., 2009), Green and Abutalebi (2013) put forth the Adaptive Control Hypothesis. According to the Adaptive Control Hypothesis, bilinguals may learn to optimally control or balance proactive and reactive methods of control dependent upon the monitored environment and the level of interference (i.e., interactional context) (D. W. Green & Abutalebi, 2013). Using a single language in distinct environments (e.g., work versus home), two languages within a conversation or between different speakers, or dense code-switching (i.e., using multiple languages within the same sentence or utterance) presents varied demands on the bilingual language control processes and is thus suggested to show different patterns of adaptive response. Of particular interest to the current study is the dual language interactional context, which reflects that of a language learning classroom wherein the target language (i.e., Spanish) and shared language of instruction (i.e., English) are both used within a conversation. In a dual language context, complex demands are made on the

cognitive control processes, resulting in a similarly complex adaptive response. "To reduce interactional cost, speakers must sustain attention to the current language goal and suppress interference but be in a position to switch languages on detection of an addressee with whom they converse in their other language. Reducing the interaction cost specifically requires that they manage the control dilemma that reducing interference also reduces sensitivity to relevant external cues" (D. W. Green & Abutalebi, 2013, p. 521). Therefore, according to this model, the areas of attentional control which experience increased demands, and thus are likely to adapt to become more efficient, are sustained attention, inhibitory attention, and attentional switching.

Furthermore, recent findings that acquiring a language in adulthood, even without achieving a native-like proficiency, may result in similar cognitive effects to that of bilingualism from childhood has led to a blossoming field of research in the cognitive effects of language learning (Bak, Nissan, et al., 2014; Bak, Vega-Mendoza, et al., 2014; Sullivan et al., 2014; Tao et al., 2011, 2015). Indeed, the Adaptive Control Hypothesis predicts that sustained attention, inhibition, and attentional switching would be increasingly recruited in a language learning environment (i.e., dual language interactional context), and therefore these areas of attentional control are likely to become more efficient. The cognitive measures used within the current study, and the subsequent language learning chapters as part of this thesis (Chapter 5 and Chapter 6), include a measure of these three attentional functions.

4.1.2 Language learning in older age

As reviewed in Chapter 1, there are a handful of studies which look specifically at the effect of language learning on cognition in the third-age against the backdrop of the current theory of cognitive reserve. Only one set

of studies in the review (Bak et al., 2016c; Long et al., 2020) explored classes with students across the lifespan (i.e., mixed ages in the classroom) as opposed to classes comprised solely of older learners. These studies measured the effect of intensive week-long residential language courses on the Isle of Skye in Scotland. Bak and colleagues (2016c) found that after just one week of intensive language learning (average of 14 hours of language learning), 33 participants aged 18-78 improved on an attentional switching task compared to passive controls. A subset of participants were tested 9-months after the course and those who continued to practice the target language (Scottish Gaelic) for at least 5-hours a week were found to have maintained the improvement in attentional switching. In a follow-up study with 105 learners across the lifespan, Long and colleagues (2020) found that in adults aged 21 to 85, while older participants scored lower on attentional switching measures, there was not a significant interaction between age and session supporting the findings of a cognitive impact of language learning across the lifespan. Notably, the Isle of Skye is widely considered one of the most beautiful places in the world, and in fact was rated by National Geographic Magazine as the 4th best island in the world (Lawrence-Jones, 2022; Tourtellot, 2007). This environmental context raises the question of if the improvement in attentional functions reported by Bak et al. (2016c) and Long et al. (2020) is due to a holiday or destination effect and substantial change in the participants' day-to-day environment, rather than the language learning experiences themselves.

4.1.3 Present study

An opportunity to address this question arose in the Summer of 2018 as the BBC (presenter and producer Michael Mosley) was looking to produce

a program on bilingualism (BBC Two, 2018; Mosley, 2018). The study was conducted in collaboration with Lingo Flamingo, a social enterprise based in Glasgow providing laid-back, tailored language learning classes for adults of all ages, including those living in care homes (*Lingo Flamingo - Our Story*, n.d.). Advised by Lingo Flamingo, Spanish was selected as the target language (see Section 4.2 for further rationale on the selection of the target language). The first part of the study was conducted in Glasgow as part of the BBC program. Additional funding was acquired to increase the sample size, though CoViD-19 meant that the in-person classes needed to be adapted to an online delivery. The current study extends the findings from intensive mixed-age language learning classes on the Isle of Skye to a normal daily life context.

Notably, the aims of this study were not to compare language learning to another activity or to a control group. Instead, the focus was on the effects of age in a non-residential context (following from Long et al., 2020) and the difference between in-person and online language learning with mixed-age classes. Specifically, the present study explored two primary research questions:

- Is there a cognitive effect of intensive language learning (a one-month course three times a week) in a non-residential context (i.e., students living at home and not staying at a language school) in participants across the lifespan?
    - Is there a difference in cognitive outcomes for students who took the classes online versus those who were in-person?
    - If so, what factors influence such effects (e.g., gender, age, years of education, hours of practice of the language, previous language background, etc.)?

- How did participants find the course to be in terms of feasibility and enjoyability?

4.2 Methods

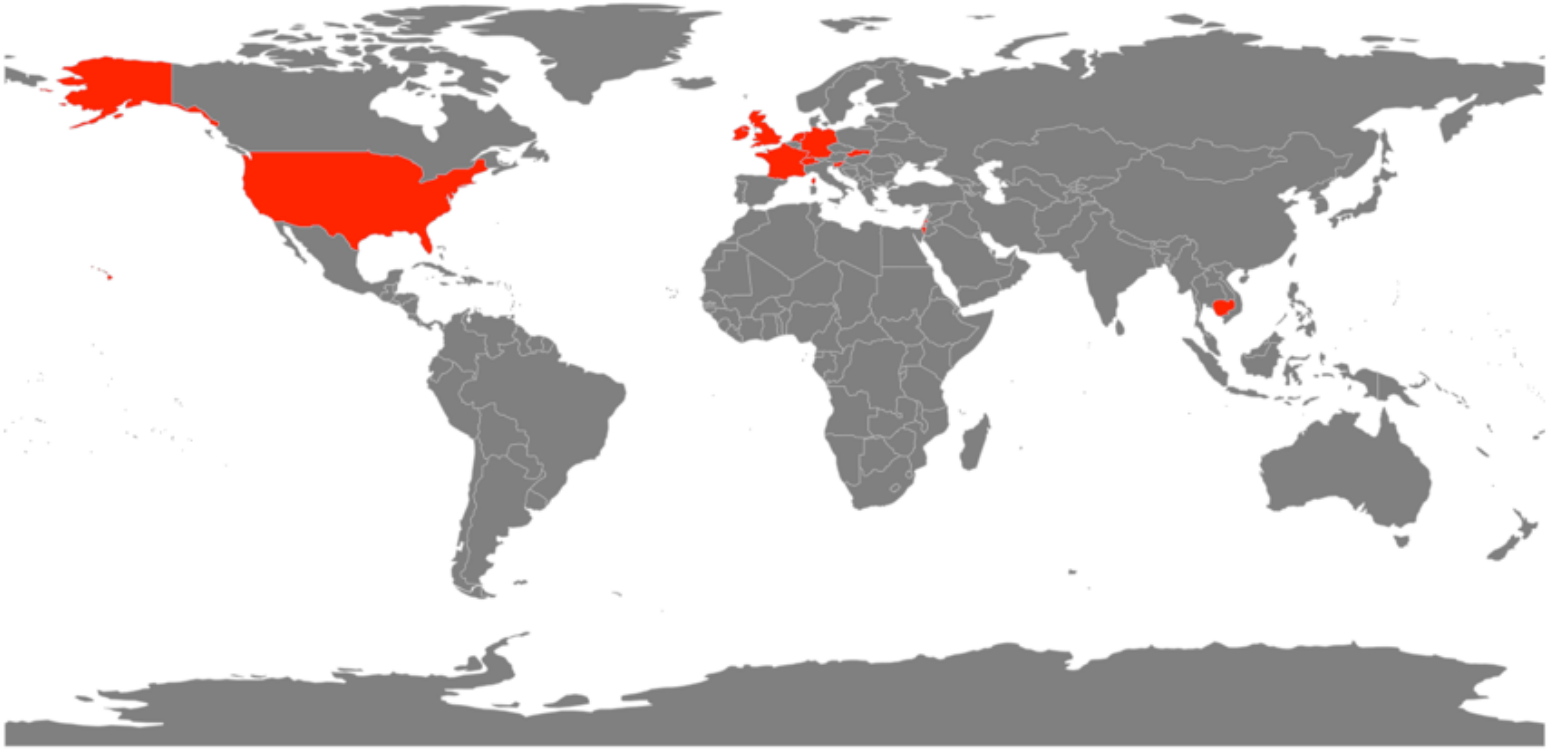
Two-hour beginner's Spanish classes took place three times a week for 4 weeks, for a total of 24 hours of class-based instruction (6 hours a week). As advised by Lingo Flamingo, Spanish was selected as the target language as it is a popular language often associated with holidays (in the UK context) and is not typically taught in schools (as is the case with French or German), therefore making recruitment of complete beginners across the lifespan more feasible. Indeed, according to the 2020 Duolingo Language Report, Spanish is the most popular language studied in the UK, followed by French and German (Watkins, 2020). Classes were offered to participants at no cost, though they were asked to participant in research session before and after the course. Inclusion criteria included having never studied Spanish before, being 18 or older, and available for 4 weeks to attend the classes. Eighty-six adults aged 18 – 77 (66 female) were recruited. Forty adults participated in courses in-person in 2018 while 46 adults participated in online classes, due to the CoViD-19 pandemic, in 2020. As the study relied on convenience sampling and data had already been collected in person, the sample size of the online courses was determined by the goal of gathering a comparable number of students. Further, the time frame for data collection was determined largely by the availability of funding, which expired just after the final round of in person classes That is, the stopping criteria was determined by circumstances outwith research control. To keep class sizes small, there were four rounds of classes both in-person and online (see Table 4.1). All classes were taught by the same teacher using the same materials. In-person classes took place in Glasgow, which limited participants

to those who would be able to commute three times a week. The online classes, however, were limited instead by time-zone rather than physical location, which resulted participants attending from over 13 countries (see Figure 4.1).

Table 4.1 Participant numbers by class

	In-person		**Online**	
	Started	**Completed**	**Started**	**Completed**
Class 1	12	9	12	7
Class 2	9	9	12	8
Class 3	9	7	14	12[1]
Class 4	10	5[1]	8	6
Total Course	40	30	46	33
Total Research	40	29	46	32

Figure 4.1 World map of participants in the online classes



4.2.1 In-person classes

Mixed-age, beginners Spanish classes took place in 2018 in-person as part of a BBC filming in Glasgow (BBC Two, 2018; Mosley, 2018). Four classes total were run, with 29 participants (*Mage* = 42, *SD* = 19.7) completing the course and accompanying research. As a total of 40 participants were recruited, the classes had a 75% retention rate with 30 participants completing the course (one participant completed the course but not the final research session).

4.2.2 Online classes

Four online Spanish classes took place in 2020 and were taught by the same teacher as those in-person, using the same textbook and learning materials. The online classes were the same as those in-person in terms of intensity, materials, content, and instructor. Class 1, Class 2, and Class 3 were all delivered via webroom.net, which was the teacher's preferred online platform. Due to technical issues on the teacher side as well as the platform requiring certain processing speeds which not all student computers supported, classes were moved to Zoom halfway through Class 3. Class 4 took place entirely on Zoom. A total of 33 students completed the courses (*Mage* = 40.4, *SD* = 15.2), and 32 completed the accompanying research sessions, of the 46 students recruited (72% retention rate). Demographics of participants who completed the course and research sessions are shown in Table 4.2.

Table 4.2 Demographics of participants who completed the classes and research sessions

Characteristic	**Overall**, N = 61[1]	**In-person**, N = 29[1]	**Online**, N = 32[1]
Class Number			
Class 1	—	9 (31%)	7 (22%)
Class 2	—	9 (31%)	8 (25%)
Class 3	—	7 (24%)	11 (34%)
Class 4	—	4 (14%)	6 (19%)
Participant Age	42 (18, 77)	44 (19, 77)	40 (18, 71)
Gender			
Female	46 (75%)	23 (79%)	23 (72%)
Male	15 (25%)	6 (21%)	9 (28%)
Handedness			
Right	56 (92%)	28 (97%)	28 (88%)
Left	4 (6.6%)	1 (3.4%)	3 (9.4%)
Both	1 (1.6%)	—	1 (3.1%)
Education Level			
Primary school	1 (1.7%)	—	1 (3.1%)
Secondary/High school	8 (13%)	7 (25%)	1 (3.1%)
Vocational degree/ Professional training	6 (10%)	5 (18%)	1 (3.1%)
University Degree	25 (42%)	10 (36%)	15 (47%)
Master's Degree	18 (30%)	5 (18%)	13 (41%)
Doctorate	2 (3.3%)	1 (3.6%)	1 (3.1%)

Characteristic	**Overall**, N = 61[1]	**In-person**, N = 29[1]	**Online**, N = 32[1]
(Missing)	1	1	—
No. Languages Before Class	3.51 (1, 9)	2.1 (1, 4)	4.78 (2, 9)

[1] n (%); Mean (Range)

4.2.3 Materials and procedure

The study is mixed methods, adopting an explanatory sequential design (see Section 2.5 for a discussion of the advantages of mixed methods). Following a repeated measures (pre- and post) design, participants attended research sessions consisting of questionnaires and an auditory cognitive test before the start of the first class and after the last class. The auditory cognitive test was the Test of Everyday Attention (Robertson et al., 1994), described in detail in Section 2.6.5.1, which is the primary outcome variable of the present study. Participants also completed the Language History Questionnaire (see Section 2.6.4.1), a free response question about their motivation for learning Spanish, as well as a series of language attitude questions, however this data was not analysed within the present study (see Appendix F and Table 4.3). The testing battery differed slightly between the in-person and online cohorts due to findings from research conducted in the interim resulting in an updated approach to covariates. As such, the online participants also completed Raven's Advanced Progressive Matrices, described in Section 2.6.5.2 (Raven et al., 1998), the Friendship Scale (Hawthorne, 2006), as well as a series of VAS style questions about the online learning experience. See Table 4.3 for a visualisation of the data available for each group and what was analysed within this study.

Table 4.3 Data collected within each group and analysed within the present study

	Group	
Data Collected	**In-person**	**Online**
TEA *	Yes	Yes
Raven's APM *	No	Yes
Friendship scale*	No	Yes
Language history *	Yes	Yes
Language attitudes	Yes	Yes
Motivation	Yes	Yes
Experience with online learning	No	Yes
Quiz *	Yes	Yes[1]

[1] data collected, though lost and therefore not analysed
* data analysed within the present study

In-person testing took place in groups in the classroom, while online testing was one-to-one due to technical limitations (see Section 2.7). During the last class, both groups were given a quiz to assess their learning from the course. The content of the quiz, much like the content of the course itself, was the same in-person and online. Unfortunately, the quiz data is only included for the in-person classes as the teacher misplaced the results from the online groups. After completing the course and second research testing session, participants were invited to attend a feedback back session. These feedback sessions were post-course focus groups to gather the students' opinions and experiences from the course, including questions around the learning experience, intensity, frequency, and duration of the course (see Appendix G).

The study was given ethical was approval by the University of Edinburgh PPLS Research Ethics Committee (ref: 115-1718/4 and 440-1819/8).

4.2.4 Analysis

In order to answer the two research questions of the present study, a mixed method analytic approach was adopted. Prior to the modelling stage, the quantitative outcome data (i.e., TEA scores) was analysed using descriptive statistics, and where the descriptive statistics suggested potential differences between groups or testing sessions, using inferential statistics in the form of *t*-tests or ANOVAs. Where parametric test assumptions are violated, non-parametric alternatives were used (e.g., Wilcoxon signed-rank test in lieu of a paired-sample *t*-test, Mann-Whitney U test in lieu of an independent sample *t*-test, etc.). As described in Section 3.2.4, Holm-Bonferroni corrected p-values are reported.

As the outcome measures of the primary quantitative research questions are the cognitive scores and the data is clustered by participants within classes, linear mixed effects modelling was selected as the method of data analysis (see Section 2.4.4). Models were fitted with participants nested within class; however, this produced either convergence issues or a singular fit, even in a null model suggesting the random effects structure was too complex to be supported by the data. As such, nesting within class was removed and the random intercept by participant grouping structure remained for all models reported. Consistent with previous research using the TEA, a trial random effect was not included as the models were intended to measure overall score on each TEA task rather than the change in scores (Long et al., 2020). In terms of fixed effects, categorical covariates/predictors were included using treatment coding while numeric covariates/predictors were mean-centred for interpretability (age of participants M = 41.95, years of education M = 16.82, average hours per week practicing Spanish M = 8.58, and composite language knowledge score M = 41.95). An interaction between testing session

(i.e., pre-course or post-course) and group (i.e., in-person or online) well as age of the participant and testing session (fixed effects) was investigated for each model within the model building strategy. Overall, the approach to model building adopted follows Barr et al. (2013)'s suggestion of beginning with the maximal random effects structure justified for the study design, or "keeping it maximal". All models were fit with Restricted Maximum Likelihood estimation and the bobyqa optimiser.

In order to draw inferences from the model and its parameters, case-based bootstrapping was used to compute 95% confidence intervals around the model parameter estimates. Bootstrapping is "a general approach to estimate the bias and the variance (and consequently the standard error) of an estimator under minimal assumptions" (Leeden et al., 2008, p. 405). There are a handful of approaches to apply bootstrapping to multi-level models, including parametric bootstrap, residual bootstrap, and case-based bootstrap. Under the parametric bootstrap, it is assumed that the explanatory variables are fixed as well as that the model specification and distribution(s) are correct (Leeden et al., 2008). With slightly weaker assumptions, residual bootstrapping assumes the explanatory variables to be fixed while only the model specification is assumed to be correct. The case-based bootstrap, alternatively, requires minimal assumptions. The hierarchical dependency in the data (i.e., the random effects structure) is assumed to be specified correctly, but no other assumptions are made of the data, the model, or the model fit. Further, case bootstrapping is most suitable with smaller sample sizes and was thus selected for the current analysis.

When employing case bootstrapping, a decision must be made about which level of the data to resample (e.g., observations, clusters, both, etc.). In the models reported as part of the current study, which includes multiple

observations from the same student, the participants (i.e., clusters) were resampled. While case bootstrapping is more robust, it can be less efficient than parametric or residual bootstrapping (Leeden et al., 2008). Therefore, the number of bootstrap resamples was set to 5,000 (as opposed to the usual 1,000 or 2,000). Further, percentile method confidence intervals are reported as it is not only common practice but also suitable for smaller sample sizes and cannot include invalid or impossible parameter values, unlike basic (AKA non-studentized pivotal method) or studentized pivotal methods (Carpenter & Bithell, 2000). Assumptions and case diagnostics were checked for all reported models.

Two post-hoc analyses were conducted following the modelling stage of analysis. The first set of analyses explored the statistical significance of the baseline differences between the online and in person groups in their self-rated proficiencies in languages prior to the course via an independent *t*-test. Following this, the rate of change in the ETR subtest of the TEA was compared between the two groups. As the descriptive statistics did not show a substantial difference, inferential statistics were not calculated. Second, as class attendance and time spent practicing Spanish during the course predicted ETD and ETR scores, the question of reverse casualty arose. Is it simply the case that participants who have difficulties concentrating get frustrated by the classes and do not attend? While this was not manipulated experimentally, the relationship between baseline attentional scores and resulting attendance was investigated via correlations in the current data.

Statistical analyses and data visualisations were conducted using R Version 4.1.3 (R Core Team, 2022), the lme4 (Bates et al., 2015), lmerresampler (Loy et al., 2022), rstatix (Kassambara, 2021), effectsize (Ben-Shachar et al., 2020), ggplot2 (Wickham, 2016), and patchwork (Pedersen, 2020) packages.

For the primary qualitative research question, the free-response questionnaire data and the feedback session data was transcribed orthographically/verbatim and analysed using reflexive thematic analysis (Braun & Clarke, 2006, 2021a) (see Section 2.3.1.5.3). Qualitative data was analysed and coded using NVivo 12 Pro (released January 2022) (*NVivo*, 2022). Prior to the initial coding stages, all participant names were anonymised (as described in Section 2.6.1). As the primary researcher throughout this project, I ensured prolonged engagement with the data and individual participants prior to analysis. I noted my initial impressions on potential analytic interests after each feedback session as well as after transcription. A critical realist epistemological stance was taken in this study, recognising the human influence on reality and truth which is situated and interpreted. As such, this epistemological stance allows for the analytic process to focus on the lived experience of the participants while considering the social and cultural influences on these experiences (Braun & Clarke, 2021a). Therefore, both a deductive (based on the interview schedule, particularly relating to question around the intensity of the course) and inductive approach to coding was adopted, using both semantic (explicit or overt) and latent (implicit or underlying) coding frameworks. Indeed, RTA allows for both inductive and deductive theme generation, which particularly suited the qualitative arm of the current research. This flexibility allowed for the analytic process to be shaped not only by pre-existing questions (e.g., around feasibility of the course) but also to be informed by the data and experiences of the participants themselves. As the role of age is a central question throughout the present research and thesis more generally, quotes reported in the results section include the participant's age as well as anonymised name.

4.3 Results

4.3.1 Dropouts

The retention rate between the classes delivered in-person (75%) and online (72%) was similar. Only Class 2 in-person finished with all of the students who started (Table *4.1*). The largest number of dropouts within a single class was 5 students, in Class 4 in-person and Class 1 online. Notably, the online classes did not see a significant increase in students who withdrew from the course. In the in-person classes, the majority of participants who withdrew or did not complete the course cited either scheduling difficulties or not wanting to travel three times a week for class as the reason for dropping out. Across the four online classes, 7 (54%) students withdrew due to technical difficulties and 3 (23%) never attended a class but completed the first research session. The demographics of the students who withdrew or did not start the classes are in Table 4.4 and compared to those who completed the course in Table 4.5. As seen in Table 4.5, the baseline TEA scores as well as the majority of demographic variables are similar between the participants who completed the course and those who dropped out.

Table 4.4 Demographics of students who dropped out

Characteristic	**Overall**, N = 23[1]	**In-person**, N = 10[1]	**Online**, N = 13[1]
Class Number			

Characteristic	**Overall**, N = 23[1]	**In-person**, N = 10[1]	**Online**, N = 13[1]
Class 1	—	3 (30%)	5 (38%)
Class 2	—	—	4 (31%)
Class 3	—	2 (20%)	2 (15%)
Class 4	—	5 (50%)	2 (15%)
Participant Age	38 (21, 65)	43 (22, 65)	34 (21, 58)
(Missing)	1	—	1
Gender			
Female	18 (78%)	7 (70%)	11 (85%)
Male	4 (17%)	3 (30%)	1 (7.7%)
Prefer not to answer	1 (4.3%)	—	1 (7.7%)
Handedness			
Right	20 (87%)	9 (90%)	11 (85%)
Left	3 (13%)	1 (10%)	2 (15%)
Education Level			
Primary school	—	—	—
Secondary/High school	7 (30%)	4 (40%)	3 (23%)
Vocational training/ Professional degree	2 (8.7%)	—	2 (15%)
University Degree	6 (26%)	3 (30%)	3 (23%)
Master's Degree	5 (22%)	1 (10%)	4 (31%)
Doctorate	1 (4.3%)	—	1 (7.7%)
(Missing)	2 (8.7%)	2 (20%)	—
No Languages before Course	2.35 (1, 5)	1.3 (1, 2)	3.5 (2, 5)

Characteristic	**Overall**, N = 23[1]	**In-person**, N = 10[1]	**Online**, N = 13[1]

[1] n (%); Mean (Range)

Table 4.5 Demographic and baseline scores of participants who completed the course versus dropped out

Characteristic	**Completed Course**, N = 61[1]	**Dropped Out**, N = 23[1]
Participant Age	42 (18, 77)	38 (21, 65)
(Missing)	—	1
Gender		
Female	46 (75%)	18 (78%)
Male	15 (25%)	4 (17%)
Prefer not to answer	—	1 (4.3%)
Handedness		
Right	56 (92%)	20 (87%)
Left	4 (6.6%)	3 (13%)
Both	1 (1.6%)	—
Number of Languages	3.51 (1, 9)	2.35 (1, 5)
Education Level		
Primary school	1 (1.7%)	—
Secondary/High school	8 (13%)	7 (30%)
Vocational degree/ Professional training	6 (10%)	2 (8.7%)
University Degree	25 (42%)	6 (26%)
Master's Degree	18 (30%)	5 (22%)

Characteristic	**Completed Course**, N = 61[1]	**Dropped Out**, N = 23[1]
Doctorate	2 (3.3%)	1 (4.3%)
(Missing)	1 (1.7%)	2 (8.7%)
Pre-course Elevator Task	98.59 (85.7, 100)	99.38 (85.7, 100)
Pre-course Elevator Task with Distraction	78 (10, 100)	74 (0, 100)
Pre-course Elevator Task with Reversal	58 (0, 100)	53 (0, 100)

[1] Mean (Range); n (%)

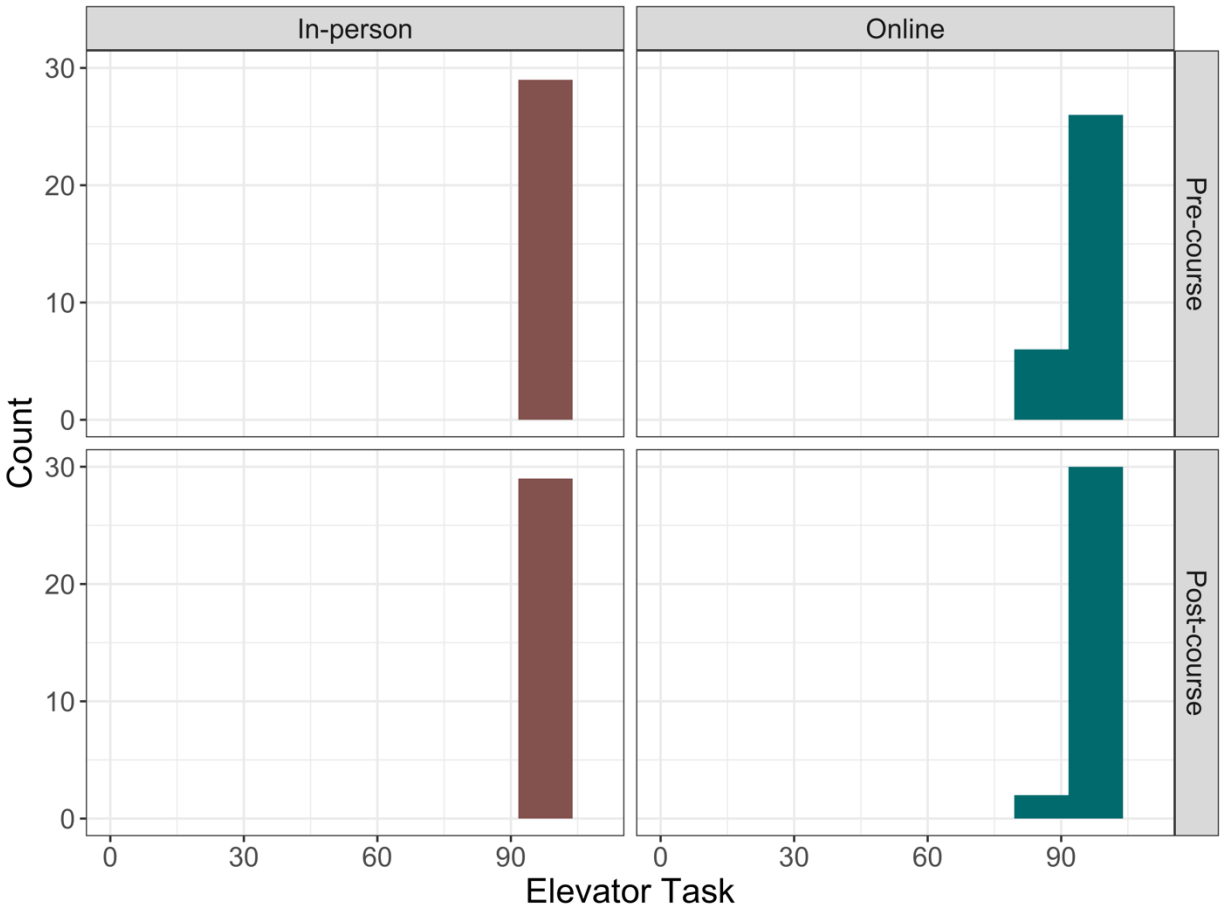
Sixty-one participants were included in the final sample for analysis, excluding those who dropped out of the course and the two who completed the course but not the final research session.

4.3.2 Cognitive outcomes

4.3.2.1 Elevator Counting

The Elevator Counting task, as expected given previous literature and that it is the least computationally difficult of the auditory subtests, had a ceiling effect, shown in Figure 4.2 (Bak et al., 2016c; Long et al., 2020; Vega-Mendoza et al., 2015). As such, it was not further analysed as part of this study.

Figure 4.2 Elevator Counting scores by group and session

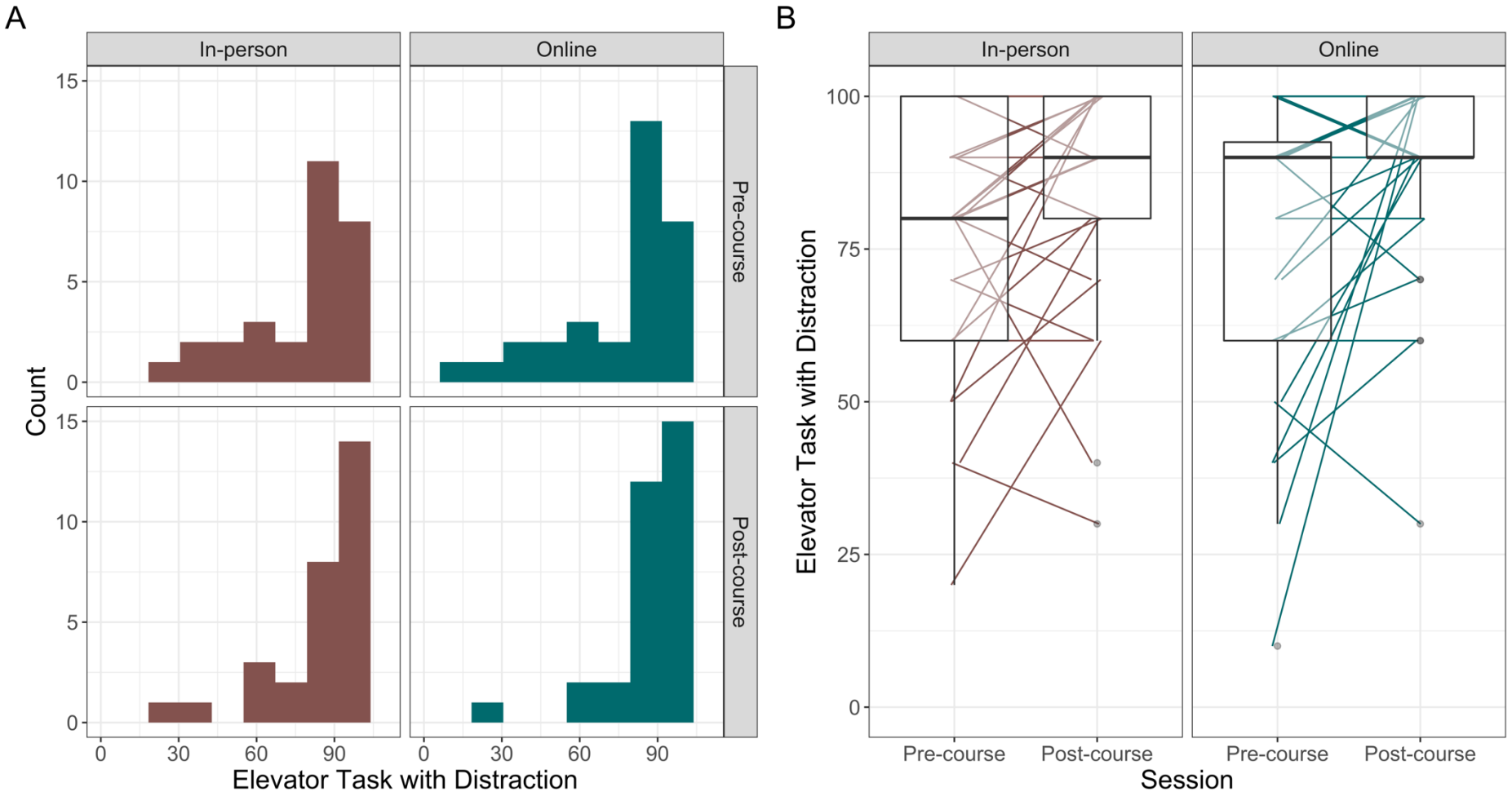


4.3.2.2 Elevator Counting with Distraction

The Elevator Counting with Distraction task showed more variability across sessions (Figure 4.3). At the pre-course session, the online group (M = 77.81, SD = 24.06, Mdn = 90) scored higher than the in-person group (M = 77.24, SD = 21.86, Mdn = 80), however this difference was not significant, $U(N_{\text{in-person}} = 29, N_{\text{online}} = 32) = 441.5$, $p = .746$. The overall change in scores before (M = 77.54, SD = 22.85, Mdn = 90) and after the course (M = 87.21, SD = 17.71, Mdn = 90), however, was significantly different, $W = 158.5$, $p < .001$, with a moderate effect size ($r = .44$). As such, a linear mixed model with random intercepts by participant was fitted, see Table 4.6. Controlling for group,

participant age, language knowledge before the class, participant gender, and participant education level, the testing session (β = 9.65, cluster sample bootstrap 95% CI: 4.21 – 15.3), hours on average spent practicing Spanish per week (β = 0.36, cluster sample bootstrap 95% CI: 0.01 – 0.84) and the number of classes missed (β = -4.38, cluster sample bootstrap 95% CI: -8.69 – -0.20) significantly predicted ETD scores. As described in Section 4.2.3, the in-person and online groups have slightly different data, due to the online group including an updated testing battery and missing the quiz data. To explore these variables, the larger dataset was subset by group. The in-person ETD model includes the quiz data (see Table H1) while the on-line ETD model includes Ravens APM and the Friendship scale (see Table H2). The resulting models are reported in Appendix H.

Figure 4.3 Elevator Counting with Distraction scores



Note: Figure A shows histograms of ETD scores by group (in-person or online) and testing session (pre-course and post-course). Figure B shows 2 boxplots of ETD scores by testing session with the lines representing individual participants, one for the in-person group and one for the online group.

Table 4.6 ETD model

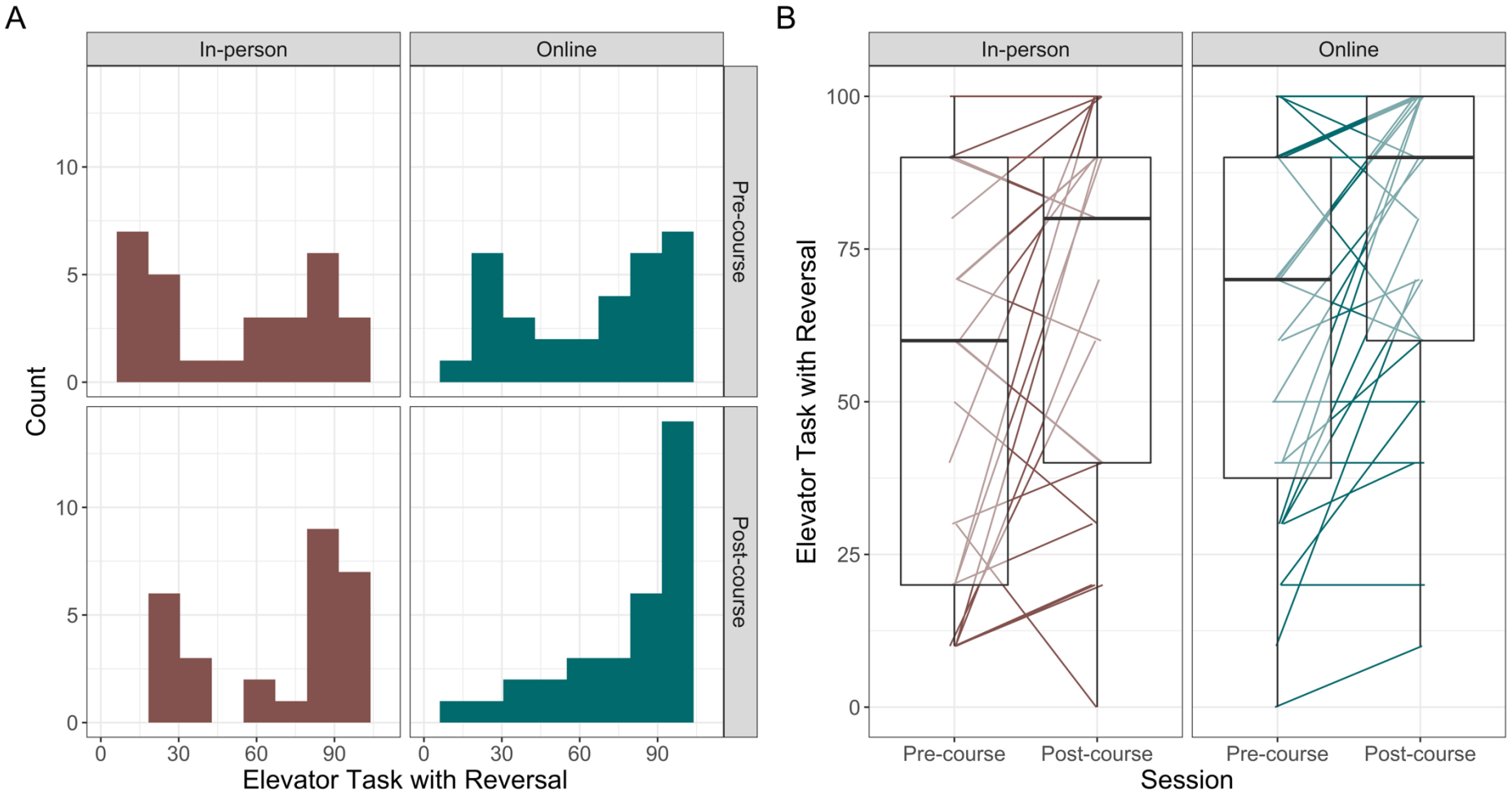
Predictors	**Beta**	**95% CI**[1]
(Intercept)	78.0	**69.9, 86.2**
Session		
Pre-course	—	—
Post-course	9.65	**4.21, 15.3**
Group		
In-person	—	—
Online	6.21	-4.11, 16.6
Age of Participant	-0.22	-0.46, 0.069
Composite Language Background Score	-0.05	-0.25, 0.19
Classes Missed (out of 12)	-4.38	**-8.69, -0.20**
Hours Practicing Spanish	0.36	**0.01, 0.84**
Gender		
Female	—	—
Male	1.38	-8.34, 10.3
Years of Education	-0.02	-1.92, 1.68
Observations	114	
ICC	.42	
Marginal R^2 / Conditional R^2	.14 /.50	

[1] CI = Confidence Interval, case-based bootstrap with 5,000 samples

4.3.2.3 Elevator Counting with Reversal

The Elevator Counting with Reversal task scores also showed variability (i.e., no ceiling effect) as seen in Figure 4.4. At the pre-course session, the online group (M = 63.44, SD = 31.58, Mdn = 70) scored higher than the in-person group (M = 51.72, SD = 34.13, Mdn = 60), however this difference was not significant, $U(N_{\text{in-person}} = 29, N_{\text{online}} = 32) = 360.5$, $p = .134$. The overall change in scores before (M = 57.87, SD = 33.07, Mdn = 60) and after the course (M = 73.11, SD = 29.75, Mdn = 90), though, was significantly different, $W = 207$, $p < .001$, with a large effect size ($r = .50$). As such, a linear mixed model with random intercepts by participant fitted using the bobyqa optimizer, see Table 4.7. Controlling for group, language knowledge before the class, participant gender, and participant education level, the testing session (β = 16.0, cluster sample bootstrap 95% CI: 9.85 – 22.6), participant age (β = -1.03, cluster sample bootstrap 95% CI: -1.31 – -0.74), classes missed (β = -6.56, cluster sample bootstrap 95% CI: -13.1 – -0.58), and average hours per week spent practicing Spanish (β = 0.86, cluster sample bootstrap 95% CI: 0.37 – 1.40) were significant predictors of ETR scores. As with the ETD models, the ETR models were subset by group to further explore the variables available within the two groups, reported in Appendix I. The in-person ETR model includes the quiz data (see Table I1) while the on-line ETR model includes Ravens APM and the Friendship scale (see Table I2).

Figure 4.4 Elevator Counting with Reversal Scores



Note: Figure A shows histograms of ETR scores by group (in-person or online) and testing session (pre-course and post-course). Figure B shows 2 boxplots of ETR scores by testing session with the lines representing individual participants, one for the in-person group and one for the online group.

Table 4.7 ETR Model

Predictors	**Beta**	**95% CI[1]**
Intercept	63.1	**49.7, 74.4**
Session		
Pre-course	—	—
Post-course	16.0	**8.95, 22.6**
Group		
In-person	—	—
Online	10.8	-1.79, 21.3
Age of Participant	-1.03	**-1.31, -0.74**
Composite Language Background Score	-0.10	-0.43, 0.17
Classes Missed (out of 12)	-6.56	**-13.1, -0.52**
Hours Practicing Spanish	0.87	**0.37, 1.40**
Gender		
Female	—	—
Male	-2.12	-15.0, 11.3
Years of Education	2.05	-0.99, 4.26
Observations	114	
ICC	.42	
Marginal R^2 / Conditional R^2	.44 / .67	

[1] CI = Confidence Interval, case-based bootstrap with 5,000 samples

4.3.3 Post-hoc analyses

The first set of post-hoc analyses explored the baseline differences between the online and in-person cohorts of students. The online cohort (*M* = 63.44, *SD* = 31.58, *Mdn* = 70) did perform higher than the in-person group (*M* = 51.52, *SD* = 34.13, *Mdn* = 60) at baseline in the measure of attentional switching (ETR), although this difference was not statistically significant. This difference could be explained by the different language backgrounds of participants in the two groups. Indeed, based on self-rated proficiency of known languages prior to the course, where a score of 20 indicated a rating of fluent in the 4 categories (i.e., command, understanding, reading, writing), the in-person (*M* = 26, *SD* = 7.92, *Mdn* = 24) and online groups (*M* = 56.41, *SD* = 26.2, *Mdn* = 53) are significantly different ($t(37) = -6.26$, $p < .001$). However, the rate of change in ETR scores before and after the course are similar across groups: in-person (*M* = 15.17, *SD* = 30.43, *Mdn* = 10) and online (*M* = 15.31, *SD* = 23.69, *Mdn* = 24).

Second, the correlation between baseline TEA scores and class attendance and time spent practicing Spanish were explored. Baseline scores on the ETD were not significantly correlated with class attendance ($r(59) = -.124$, $p = .341$) or the average hours spent practicing the language outwith the class ($r(56) = .159$, $p = .233$). Similarly, baseline ETR scores were neither correlated with class attendance ($r(59) = -.142$, $p = .27$) nor with time spent practicing the language ($r(56) = .204$, $p = .125$).

4.3.4 Student experience

During the feedback sessions after the course, the students were asked to describe the course in 1 or 2 words, shown in Figure 4.5.

Figure 4.5 Word cloud of the course described in 1 or 2 words

As shown in the word cloud in Figure 4.5, a variety of topics are covered in the course description, which was further supported and reflected in the feedback sessions and post-course questionnaires. Particularly relevant to the research questions within the current study are the themes: "Learning is fun: new language, new friends"; "Time well invested"; and "Spanish today, motor bike riding tomorrow".

4.3.4.1 Learning in fun: new language, new friends

As seen in Figure 4.5, the most common description of the Spanish class experience, across both groups of students, was 'fun'. For example, Julie (22) describes the course as "*very fun. Very, very fun. And I like the group as well".* The post-course feedback sessions further illimunated that the learning experience was found to be fun and enjoyable by students due to the challenge and resulting sense of achievement as well as the cultivated social

relationships. Daniel (27), who took the course online during lockdown, described the course as, *"ideal because, I don't know about the others, but I had nothing else to do so I was really looking forward to it. It was nice to pick up things really fast"*. Outside of feelings of achievement being fun, the social relationships from the course were continuously highlighted as enjoyable. Not only was the social aspect of the course pinpointed as a key source of the fun learning experience, but it was also described as a motivating factor for successful learning: *"I've done other classes before, and I didn't expect to form such a bond with the group in such a short time. I think I picked up the language so much faster because of the social side of things and bonding with the group"* (Delilah, 27). Elise (36) further describes that *"interactiveness as well helped a lot in this class, it helped us retain information easier, at least for me."* Layla (33) reflected that while she does not generally find learning to be fun, this was not the case in the Spanish classes:

> *"I would say, 'fun learning' because usually when you learn it is a bit dull. It is, 'oh, I don't want to do this. I don't want to do that.' You have to put effort and time. At school usually you have it there and it is a bit like 'ugh'. It is like you don't want to do that, it is boring, and it is not fun. But it was really fun this time…it was absolutely lovely. And I just keep telling people 'just don't be afraid! Online is fine, you can still have social interaction and everything.' Yeah, it was amazing".*

Indeed, a fundamental element of the fun learning experience was the social interaction and engagement amongst the peer group as well as with the teacher, who was continuous described as being incredibly knowledgeable as well as bringing a great sense of humour to the classroom. In fact, many students expressed feelings of genuine sadness that the classes were over in the post-course feedback session. Isla (46) goes so far as to say that any future language learning classes are likely to fall short in comparison to her experience learning Spanish, "*I think too like I will maybe go to a class in the*

future, and I will be like 'oh, they are not as nice as the people that I did my Spanish class with.' We have really got on well as a group" (Isla, 46).

4.3.4.2 Time well invested

Related to the theme of the learning experience being fun and enjoyable is the theme that the courses were a good investment of time, or *"well worth doing"* (Penelope, 63). Indeed, Elisabeth (23) reflected that she *"would recommend [doing the course]"* to others. Overall, the consensus among participants in both groups was that the course was intense, but worth it the time investment. However, this feedback and opinion, of course, comes from the participants who completed the course. For the online group, many students reported that the intensity of the course was only feasible due to lockdown measures or working from home. This is further supported by the most common reason for dropping out of the course in the in-person group being the intensity (as resulting scheduling conflicts and travel time), whereas technical issues were the most prevalence reason in the online group.

It was universally acknowledged that the courses were intense in terms of frequency 3 times a week, but many students felt the length being only 4-weeks made it more palatable. As Stanley (41) puts it, *"if you told me at the beginning that it was a five-week or a six-week course, I would have been maybe a little bit... I think, I think four felt manageable"*. The intensity of the course, in fact, was highlighted as a key source of achievement and progress in the class. *"I forget things a lot and if it hadn't been so frequently, I think my progress would have been really poor. And being able to do it very frequently, I think has made quite an impact on me"* (Matilda, 67). Jim (71) similarly felt the intensity was fundamental to his learning success and motivation, though was happy to have a break from classes three times a week: *"think it certainly for*

me, it was essential to have the four weeks to accumulate enough that you felt you could build on that. To do any less, I don't think you'd have been as committed. But I must say at this stage, I'd be glad to take a break." Likewise, Jasper (39) suggested that the course intensity was both motivating and necessary to *"learn a language effectively. You know, if you're not in the target context, sitting here in [a non-Spanish speaking country] I don't interact with Spanish speakers. So, the fact is that I have seen the progress that can be made by really making a priority of it. Same thing if you go the gym, or whatever you do in life, you kind of need to say, 'This is the time' every day or every week."*

Furthermore, many students highlighted that the intensity of the classes particularly complimented the style of the teacher who typically spent the first half hour of the class revising, which "*casually reinforc[ed] what we've done previously"* (Scarlett, 57). Other students expressed both surprise and a sense of pride in how much content was covered in just one month and the associated achieved learning outcomes: *"I feel I can say more in Spanish after 4 weeks than any other language that I have ever learnt."* (Florence, 40). Millie (46) similarly reflected that *"I know more Spanish after 4 weeks than other languages that I have been studying for months or even a year."* Jesse (36) shared on his surprise at how much he and his classmates had learned while taking the quiz after the last class: *"we learned a lot more than I thought, more than I thought we would. I, even when we're doing that exam, I realized I knew a lot more Spanish than I thought I would after this amount of learning if that makes sense. I think we all feel like that, which is really good."*

The online group, given the context of the courses taking place during global CoViD-19 lockdowns, particularly highlighted that the classes helped to provide a sense of purpose and structure to the week, which otherwise blurred together due to daily life being so intensely altered by lockdown measures.

Further, the social impact of the classes was acknowledged as a motivating factor for attending during lockdown (e.g., to prevent "letting down" the teacher and peers). As described by Jasper (39), the classes offered a sense of control amongst the pervasive feeling of most things being outwith anyone's control due to the pandemic:

> *"To have a bit of control, to say right, 2-4pm I have to be there, and I have to have done my homework in advance. It does kind of ground you a little bit more and and give you a little bit of control because you have appointments that you have to keep, and you'll be upset with yourself, and you'll be letting your teacher and your classmates down."*

Indeed, as Evelyn (38) labels it, the classes functioned as *"my escape thing"* from the stresses of the pandemic, with Jesse (36) describing that *"it was nice to have some focus."*

As evidenced by the quotes included here, the theme of the courses being time well invested is motivated by numerous factors, including the language and learning achieved, but also the social interaction and engagement amongst the peer group. The social impact of the classes was powerfully felt by some participant outwith classroom hours, *"I was looking forward to each class and then afterwards feeling better than I was before, and so that was really nice"* (Delilah, 27). Florence (40) similarly reflected that she *"would finish work really stressed out at like 5pm, but I would feel so much better after the class at 7pm. It really helped".*

In the post-course questionnaire, all students expressed an interest in continuing to learn Spanish. Particularly in the online groups, all four cohorts of participants self-organised means of remaining in touch after the classes had completed, be it Facebook groups or WhatsApp chats. In fact, at the time of writing (two years after the end of the study) three of the four online cohorts of students have continued to learn Spanish together with the same teacher.

4.3.4.3 Spanish today, motor bike riding tomorrow

Particularly salient for the older participants on the course was the transformative nature of the language learning experience and feelings of success and improved self-confidence after completing the 4-weeks of Spanish classes. Isla (46) explains how not only her opinions overall towards language learning have changed as a result of the course, but also what she believes she is able to achieve:

> *"I think of the negative attitudes to learning languages that I had when I was at school and when I started this; I cannot believe that I can acutally do it. So, you know, [the course has been] confidence building. I was scared before we started. On the first day I was like a little nervous going into the class, you know, you have this screen of people all looking at you. And I felt really nervous about it. And then it has just been… after the first lesson I think the nerves are gone and you are doing the first homework and everything. It was brilliant. I really started looking forward to them. It's been so lovely. I feel so lucky to have been involved. It's been life changing, honestly. It was a really good way to start, and I think it [Spanish] is something that I will do for the rest of my life now."*

The impression of the courses as being life changing was further echoed by Penelope (63) who went on the share that she also felt intimidated entering into the first class as she *"was the oldest by 20 years at least. But then I thought, it actually doesn't matter, I am doing this for curiosity and then I really started enjoying it. The group was lovely".* In fact, Penelope further elaborated that this feeling of intimidation was quickly replaced by increased feelings of self-confidence:

> *"I went to coffee with my friends…this morning and they asked, 'why did you do that Spanish thing?' And I thought, well, if I stop wanting to do and try new things, that's a sign of old age. I enjoyed it, and I could do it. It proved to myself that I can. Who knows, maybe I will learn motor bike riding next".*

The suggested transferability of the confidence gained in the Spanish classes to other activities and her general outlook on life is particularly notable in Penelope's reflection. Interestingly, Avery (64) elaborated that her experience of finding confidence and a sense of achievement through the courses was largely due to the continued effort required by the intensity of the course, *"I found I had to push myself quite hard to try and take on board what [the teacher] was saying, on an ongoing basis over the four weeks, it wasn't just one particular class. And in being challenged, I feel I've really achieved something. I really feel I have benefited hugely from [taking the course]."* Indeed, the transformative nature of language learning was echoed throughout the different classes, both online and in-person, and amongst younger participants as well. As Florence (40) describes, she *"didn't realize [before the Spanish course] languages could be so powerful and, honestly, life changing."*

4.4 Discussion

The current study set out to explore the impact of training in the form of intensive (6 hours for 4-weeks, 24 total hours), non-residential language learning classes both cognitively and socially in mixed age language learning classes. Overall, across measures of inhibition and attentional switching, students scored higher at the post-course session both in-person and online. While in the literature most often attentional switching specifically is influenced by language learning training within a short period of time, in the present study inhibitory attention was also found to show a significant improvement after the course in both in-person and online classes. Further, students who practiced the language through homework or other means outwith the

classroom scored higher on both of these measures, with those who missed classes scoring lower. This finding suggests that indeed the time spent training in the language learning classroom and practicing the language predicts scores on attentional tests (specifically inhibition and attentional switching), as supported by previous literature which found students who practiced the language for at least 5 hours a week after the course completed maintained their improvement in cognitive scores (Bak et al., 2016c).

Attendance and time spent practicing the language outwith class significantly predicting attentional scores raises a question of reverse causality – is it simply the case that participants who have difficulties concentrating get frustrated by the classes and do not attend? While this was not manipulated experimentally, the relationship between baseline attentional scores and resulting attendance was not correlated in the current data.

Furthermore, adults who participated in language classes online did not perform differently than those who completed the classes in person. This suggests that the medium of language learning delivery does not significantly impact potential cognitive outcomes of language training. While the online cohort did have higher baseline scores on the measure of attention switching prior to the course, this difference can be well explained by the varied demographic profile and language backgrounds between the groups. The in-person group took place in Glasgow, with largely monolingual English-speaking participants who had learned an additional language in school, whereas, as seen in Figure 4.1 and Table 4.2, the online group consisted of a more diverse group of multilinguals. Post-hoc analyses confirmed that the online group was significantly more multilingual than the in-person group according to their self-rated knowledge and use of languages. The pre-course difference between groups being due to baseline differences in multilingualism

is further supported by the fact that both groups performed similarly in overall change in ETR scores after the course. The higher baseline score in attentional measures among active multilinguals is supported by previous literature (Bak, Vega-Mendoza, et al., 2014; Vega-Mendoza et al., 2015).

Also supporting previous literature (Bak et al., 2016c; Long et al., 2020) is the finding older adults in the sample scored lower overall on attentional switching scores, though did improve after the course. Across both the in-person and online groups, participants ranged in age from 18 to 77, with a mean age of 40. Therefore, the present findings should be interpreted with caution as it relates to older age in particular. While the majority of participants were not older adults as typically defined in the third-age language learning literature (i.e., 60 plus), the sample does include a range of adults across the lifespan. Notably, there was not an interaction between age and session for either of the attentional measures. While cognitive performance generally varies with age, there is no evidence in the present study supporting age-dependent impact of language learning on attention. Instead, these findings support the notion of an adaptable brain across the lifespan. This complements the findings reported above that baseline attentional scores are not related to resulting class attendance or time engaging with the language outwith the class. Collectively, these findings support the idea that there can be a cognitive benefit, for adults undertaking a new language learning experience or training. Notably, the qualitative findings further support that language learning experiences can benefit adults across the lifespan.

Students felt their perceived success in learning was due to both the intensity and social aspects of the classes. The course intensity provided motivation and what was deemed a good start to the language learning journey across both groups but was particularly appreciated by the online

group for providing structure to an otherwise structureless lockdown. Further, such an intense course can be particularly effective with adept teachers who integrate revision into each lesson. For the older participants in particular, the language learning classes presented a transformational experience which was reported to be was both empowering and enjoyable. Taken together, the quantitative and qualitative data suggest that not only can adults across the lifespan experience cognitive benefits from language learning 6 hours a week, but also that this experience is both enjoyable and rewarding regardless of age and delivery (in-person or online). So much so, in fact, that all participants in the present study expressed interest in continuing to learn Spanish, particularly with the same peer group.

While the relationship between success in language learning and attentional measures was not able to be fully explored in the present study due to loss of data, findings from the in-person group cautiously suggest the reported cognitive effect does not depend on how well students learn the language (see Table H1 and Table I1). This provisional finding suggests that the intensity of language learning experiences plays a more important role in cognitive outcomes than achievement. Future research should explore this relationship further to determine if it is a genuine effect or artefact in the current study.

The present study contributes to the literature on cognitive effects of language training as a novel investigation of mixed-age adult classes in-person compared to online for 6 hours a week across 4 weeks. Further, the novel inclusion of quiz scores as a measure of achievement was planned, though as mentioned due to missing data was not able to be fully explored. Extending findings from intensive residential language courses on the Isle of Skye, the present study found an improvement in attentional functions (namely

inhibition and attentional switching) across the lifespan, without participants learning in a beautiful holiday destination. These effects were not significantly different when learning took place online or in-person in Glasgow. The intensity of training in terms of time learning and practicing the language both within and outwith the classroom significantly predicted cognitive outcomes. Provisional findings further suggest that intensity plays a more important role than achievement, which is an important line of investigation for future research to further illuminate. Lastly, and arguably most importantly in terms of the likelihood of participants continuing to learn and use the language (and therefore maintain any cognitive improvement), participants across the lifespan found the courses both in-person and online to be highly enjoyable and worthwhile both linguistically and socially.

Chapter 5 Learning Chinese Across the Lifespan: Learning Writing and Speaking as Two Separate Languages

Language learning can entail both auditory (spoken language) and visual (written language) input, which gives rise to the question of the role of modality in measures of cognitive functions. Chinese languages[1] *offer a unique opportunity to study the role of written versus spoken language learning and visual versus auditory cognitive measures. The study reported in this chapter is based on active data collection from January 2020 until December 2020, including three rounds of language classes. The first two rounds of classes were discussed in the context of the unexpected switch to online delivery due to the CoViD-19 pandemic in:*

> Zhao, L. X., Blankinship, B., Duan, Z., Huang, H., Sun, J., & Bak, T. H. (2020). Comparing face-to-face and online teaching of written and spoken Chinese to adult learners: An Edinburgh-Sheffield case study. *International Journal of Chinese Language Teaching*, *1*(1), 83-98. doi: 10.46451/ijclt.2020.06.05

As the rationale for the study and the design remain the same, Section 5.1.1, 5.1.2, 5.1.3, and part of 5.2.1 are taken from this paper, all of which are based mainly on my contribution. However, the results reported in this chapter include more data than was available at the time of writing the paper as well as address different research questions, and thus are unique to this chapter.

[1] Due to the substantial difference between different variants of spoken Chinese, though all share the same written system, I think it is more accurate to speak of languages in the plural form.

5.1 Introduction

5.1.1 Teaching written versus spoken Chinese

The Chinese writing system, unique in the world in having preserved elements from different stages of script development, from pictographic, through ideographic to phonological, belongs to the most iconic and fascinating aspects of Chinese language and culture. However, in terms of teaching Chinese as a foreign language (CFL), characters are often seen as a burden and challenge to learners. Accordingly, the question whether the characters should be introduced to the learner at the same time as the spoken words, or later, remains a topic of intense debate.

Accordingly, contrasting views exist as to whether there should be a time difference between the start of the language programme and the introduction of characters: delayed character introduction (DCI) versus immediate character introduction (ICI), following Ye's (2013) English translation. The DCI approach has been proposed based on the learning experience of native Chinese speakers (Dew, 2005; Jorden & Walton, 1987; Packard, 1990; Swihart, 2004; Unger et al., 1993). Dew (2005) observed that native Chinese speakers usually learn Chinese characters after having acquired speaking and listening skills for five to six years. Swihart (2004) maintained that "most students need time to learn to connect the pronunciation of Chinese characters with their shapes, and only then can they read" (p. xii).

The DCI approach is based on the idea that speakers of other languages should follow a similar trajectory as native Chinese speakers in the learning of characters. This assumption has been increasingly challenged with the argument that adult second language learners should be treated differently from childhood native speakers. As Wan (2018) pointed out, Chinese children

have mastered around 2500-3000 words in speaking and listening before they start to learn characters. They have also been exposed visually to characters on a daily basis, due to their ubiquitous presence in every-day life in China.

Apart from these theoretical arguments, the DCI approach has also been challenged empirically. In the 1950s, two DCI methods were trialled based on the learning experience of Chinese children and the experience of increasing the literacy level of the adult Chinese population, respectively. Neither method achieved satisfactory results (Li & Ren, 1986). In the first method, students were taught pinyin for five to six months. Students did not start learning characters until they had mastered several hundred words in pinyin and therefore ended up with the dual tasks of learning new characters and the characters they had learned in pinyin. Thus, the bottleneck of learning characters was moved to a later point in time, but not broken (Wan, 2018). In the second method, students learned 700-800 words and basic grammar points before they learned corresponding characters intensively in the following two weeks. It was found that the students did not achieve a high accuracy in recognising the characters, and, in the meantime, they also forgot many of the words and grammar points they had learned.

Based on these considerations, the ICI approach, i.e., the introduction of Chinese characters from the first lesson, has been proposed as an alternative (I. Liu, 1983; Ma, 2018; Wan, 2018; J. Zhao, 2008). This approach has gained popularity and, currently, the majority of textbooks used in CFL courses introduce characters from the first stages of instruction (Allen, 2008).

A radically different approach has been taken by the so-called Heisig-Richardson Method. It is based on the popular book series by James Heisig "Remembering Japanese Kanji", adapted to Chinese by Richardson (1998). Focusing entirely on the written language without any reference to the

pronunciation, it makes liberal use of mnemonics, presenting stories designed to facilitate remembering characters and even encouraging learners to invent their own stories. Accordingly, the method has been criticised for its arbitrariness (So, 2008) as well as for its failure to explain ideophonetic characters (Jin, 2010).

The diversity of methods of introducing characters in CFL teaching contrasts with the dearth of empirical evidence comparing them. A recent study by Osborne et al. (2020) compared four different methods of teaching Chinese characters in Irish schools, finding only slight differences in pupils' performance. As the authors emphasise in their conclusions, further studies are needed to explore this issue systematically.

5.1.2 Teach and learning written Chinese online

Apart from the question of the best time to introduce characters, there is also a debate as to the best method of introducing characters, particularly online. The use of SCMS (Synchronous Computer Mediated Communication) in CFL teaching was pioneered by Wang and her colleagues (Wang, 2004; Wang & Chen 2007, 2009, 2012), who argue that a crucial component in online language learning is synchronous interaction.

In (2013), Stickler and Shi attempted to identify the ways in which Chinese teachers' intentions matched with student perceptions and expectations during online multimodal tutorials. Using a multimodal analysis of synchronous online spoken interactions and stimulated recall, they found that a mismatch between teacher intentions and student perceptions during online learning can not only lead to negative outcomes such as anxiety and communication failure, but even to totally abandoning the online language learning experience. In another study Stickler and Shi (2015) conducted an

online, eye-tracking study with beginner to low-intermediate learners of Chinese involving a reading comprehension task and interactive speaking task. Also employing stimulated recall interviews, they found that learners at higher levels largely used characters for comprehension while those at average or poor levels relied more heavily on pinyin. The intermediate learners in this study were found to make use of both pinyin and characters in the comprehension task.

Indeed, authors such as Zhan and Chen (2014) and Sun et al. (2014) have argued in favour of technology integration in learning Chinese characters, particularly for CFL teaching. Zhan and Chen (2014) developed the so-called REEE model (radical awareness, enforcement of sound-meaning connections of characters, enforcement of sound-meaning-form connections of characters, and evaluation) whereas Sun et al. (2014) proposed a character teaching system that integrated character structure theory with 2D contour morphing technology. Both argued that integrating technology in the teaching of characters can provide motivation for continued learning and an additional element of interest and engagement in the learning process.

5.1.3 Chinese characters as a potential aid in learning Chinese

As pointed out in preceding sections, when it comes to learning Chinese characters, there is a fundamental difference between native Chinese speakers and those learning Chinese as a foreign language. A Chinese speaking person enters the world of written Chinese with a well-established knowledge of the spoken language. Thus, new characters can be mapped onto already existing sound meaning correspondences. For a non-native Chinese speaking person, in contrast, learning written and spoken Chinese could easily appear as

learning two separate languages, competing and interfering rather than facilitating each other.

The Chinese writing system shows important differences to other writing systems in use today. Firstly, it is possible to learn the characters relatively independently from their spoken counterparts. This applies not only to pictograms and ideograms, but also to phono-semantic compounds. Even phono-semantic compounds can be learned without knowledge of their sound, with the semantic radical providing a cue to their meaning. Indeed, as demonstrated by Williams (2013), learners of Chinese as a second language do not seem to benefit from the phonological information contained in the characters. Indeed, even for native speakers the phonological clues are not always helpful as they refer to older pronunciation, Old and Middle Chinese, and the sounds may have changed substantially since then.

Secondly, from a cognitive point of view, the compositional nature of Chinese characters provides a network of visual and semantic associations which does not have any parallel in the Chinese spoken language. Taking the character 明 ming “bright” for example, comprising the radicals for 日 ri “sun” and 月 yue “moon”. This conceptually plausible and easy to remember connection between the written form of “bright”, “sun” and “moon” does not correspond to any comparable relationship between the entirely unrelated sounds of the spoken words: “ming”, “ri” and “yue”. Moreover, the radicals derived from the characters for “sun” and “moon” appear in many other characters, making it easier for the learner to recognise them.

Some radicals can be associated with whole clusters of characters, often linked to each other semantically. For instance, the radical for 金 jin “gold” appears in names of different metals (e.g., 银 yin “silver”), and also in objects

made out of metal (e.g., 铃 ling “bell”). Thus, confronted with a new and unknown character containing the 钅 “gold” radical, the Chinese learner will be able to guess that the word is likely to have at least some connection to metal. This applies not only to objects and nouns denoting them, but also to verbs, e.g., in the character for 喝 he “to drink”, which includes the radical for 口 “mouth”. In contrast, it is not possible to find similar relationships between spoken words, as neither onset, rhyme, nor tone can be associated with specific meanings or semantic categories.

5.1.4 Present study

Very often, the complexity of the Chinese characters is portrayed as a burden to a learner, making the acquisition of the language much more difficult than it would have been if only the spoken language (and the phonological pinyin transcription of it) were used. In this study, we take an opposite point of view. We set out to determine whether the introduction of Chinese characters could make learning Chinese easier rather than more difficult. Instead of following the DCI or ICI approaches, we are examining whether Chinese characters can be introduced before any spoken language. If learning spoken and written Chinese can be conceptualised as learning two different languages (or to be more precise, two separate lexical systems) and the compositionality and structure of Chinese characters can offer for a beginner an easier point of entry into the language than its spoken version, could it be that learning the characters alone might in fact be easier than learning the spoken language alone?

Furthermore, capitalising on this unique feature of Chinese, the present study is able to explore if cognitive effects of language learning are specific to

the modality of input and training (e.g., visual or auditory). While language can be visual, it is primarily, at least in how it is first learnt, an auditory phenomenon. To date, an overwhelming majority of studies on bilingualism and cognition have used visual tests such as the Simon Task, Stroop Task, and ANT (specifically the conflict network scores). Visual measures generally tend to dominate neuropsychological assessment, which has in turn influenced the type of measures adopted in the bilingualism literature. In a recent meta-analysis of the bilingual advantage in the literature, only visual measures of attention were included (A. T. Ware et al., 2020). This is due to the inclusion criteria set by the authors wherein the task needed to be used in 7 or more publications, reflecting the dominance of visual measures of attention in the wider literature (A. T. Ware et al., 2020). Within the studies of language learning in the third age reviewed in Chapter 1, visual measures of attention were used in all but two studies (Bak et al., 2016 and Long et al., 2020 who used the auditory TEA).

Accordingly, we have designed a new four-week course of Chinese for beginners, introducing the characters without any reference to their pronunciation. For comparison, we designed, in parallel, a course of spoken Chinese only, without any reference to written Chinese, neither pinyin nor characters. We then compared both courses in a crossover design, so that all participants have received both writing and speaking classes but in a different order (for details see Section 5.2.1).

The present study, therefore, has two primary research questions:

- Does the order of classes (i.e., speaking followed by writing or writing followed by speaking) influence the rates of change in cognitive measures of the training modality, or is there a transfer across modalities? In other words, does learning the written

character system result in a change in visual cognitive measures and not auditory, and vice versa?

- How do attitudes towards the two modalities (i.e., speaking Chinese and writing characters) change over time?

As few studies have simultaneously explored auditory and visual measures of attention, the present study presented a unique opportunity to undertake an exploratory analysis of this relationship. To my knowledge, Ooi et al. (2018) is only study to include the TEA and ANT in the same testing battery, though they were not directly compared. Thus, the following exploratory research question was also investigated: how do the visual and auditory measures of attention correlate between groups across testing sessions? Correlations across both groups in both auditory and visual measures of attention across all time points of measurement would support the modality of input being transferrable to modality of testing. A lack of significant correlations within groups and across time points of measurement would suggest that modality specific input results in a modality specific effect on attention.

5.2 Methods

5.2.1 Research design

Following a randomised cross-over trial design with a two-week wash out period, two-hour beginner's adult Mandarin classes took place two times a week in four-week blocks (see Figure 5.1). A brand-new Chinese beginners' course for adults was developed specifically for this study. It consisted of two separate parts: a four-weeks course of written Chinese (using simplified characters) without any reference to pronunciation and a parallel course of

spoken Chinese, designed around the same topics, but without any reference to the written language, neither characters nor pinyin. Two professionally trained and experienced CFL teachers employed by the Confucius Institute at the University of Sheffield designed and delivered the courses. As the main objective of the study design was to completely separate the teaching of spoken and written Chinese, students participated in both classes, for a total of 32 hours of language learning (16 hours per block).

The randomised cross-over trial design, a method well established in therapeutic and clinical trials, is particularly suitable for comparing different treatment options (for further details see Senn (2002) or Sibbald and Roberts (1998)). Randomised cross-over trials are a repeated measures design wherein each experimental unit (i.e., participant) is randomly assigned to study arms comprised of a sequence treatments (Senn, 2002; Sibbald & Roberts, 1998; Wellek & Blettner, 2012). The simplest, and most common model, as adopted in the present study, is an AB/BA design. Participants allocated to the AB study arm undergo treatment A followed by treatment B, whereas those in the BA study arm first undergo treatment B then treatment A. Therefore, as a function of this design, every student in the present study was due to participate in both parts of the study, half of them starting with the written course followed by the spoken one, the other half the other way round.

The students were divided randomly (using a stratified randomization procedure with the strata: age, gender, and education level, using the randomizeR package in R (Uschner et al., 2018) into two groups (see Appendix J). In this way, each participant functions as their own control as both courses can be evaluated within the individual, allowing for comparisons at eh individual rather than only group level (Dwan et al., 2019; Senn, 2002; Wellek & Blettner, 2012). For example, the Writing-Speaking (WS) group received a

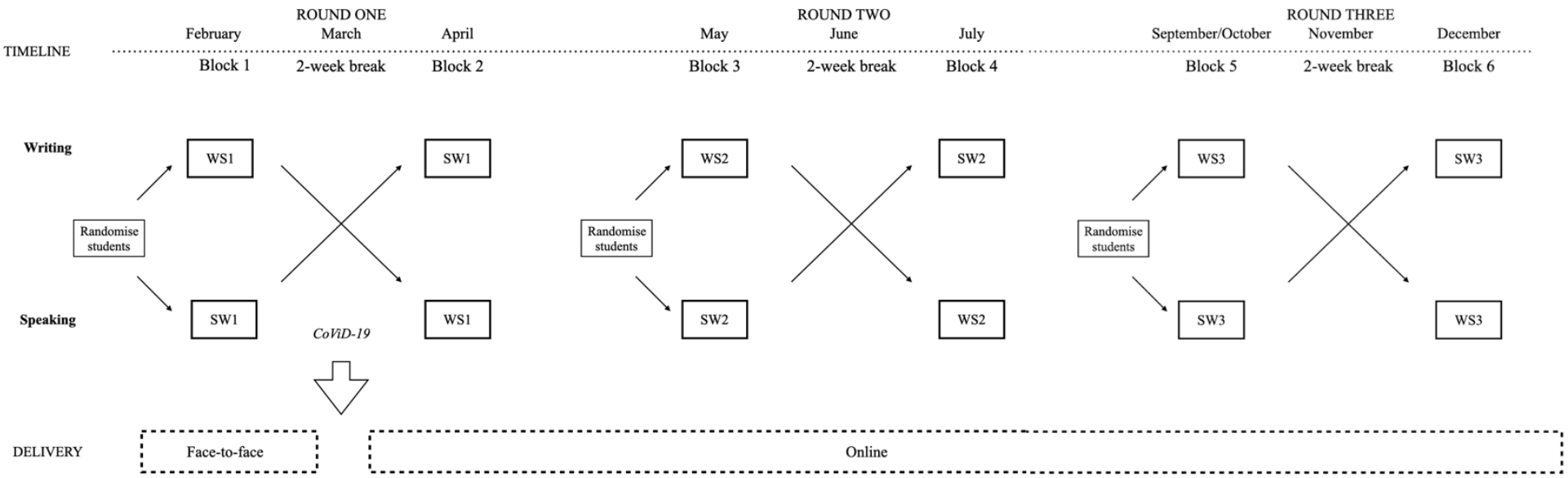
four-week intensive (two two-hour sessions per week) Chinese course on characters only (using simplified characters) without being taught how to pronounce the learned sentences. After a two-week break, the group received another four-week intensive course for the same content but in spoken Mandarin Chinese only (without using pinyin). The Speaking-Writing (SW) group had exactly the opposite curriculum: an intensive four-week course of spoken Chinese only, followed, after a two-weeks break, by four weeks of teaching written Chinese.

The two weeks interval functioned as a wash-out period, which is commonly used in cross-over trials to delineate the effects of each intervention or training and minimise carry-over effects from one treatment to another (Sibbald & Roberts, 1998). The two-week period was selected to strike a balance between retention of students and including a sufficient period of time with no courses to disentangle the effects of the first course. Based on personal communication with the Director of the Sheffield Confucius Institute it was deemed that any longer than 2 weeks between classes would present a serious concern around the retention of students. Given the cross-over design, participants must complete both classes in order to have a complete data point. Including a washout period equal to the length of classes (i.e., 4 weeks) would have been ideal, however, it was not feasible. Further, a 4-week washout period would have made the study run for a total of 12 weeks. As the study relied on convenience sampling, minimising deterrents in the recruitment and retention of participants was a priority.

To control for the specific influence of individual teachers, the two teachers swapped in the middle of each block, so that each group got two weeks with one, then the other two weeks with the other teacher. After the last class in block 2 (i.e., the last class overall), students were given a quiz to assess

their learning across both courses. As described in Zhao and colleagues (2020), CoViD-19 lockdowns unexpectedly took place during the wash-out period of the first round of classes, which took place in-person in Sheffield, resulting in all of the following classes taking place online via Zoom. The teaching materials and specific details around course delivery both in-person and online are discussed in depth in Zhao et al. (2020).

Figure 5.1 Research design



5.2.2 Participants

A total of 100 participants aged 18-60 (51 female) from 25 countries (see Figure 5.2) were recruited across three groups: Writing-Speaking, Speaking-Writing, and Traditional. Inclusion criteria included being 18 or older not having any prior knowledge of Chinese languages and in the case of the in-person courses, having the ability to travel to Sheffield twice a week to attend classes. Participants were recruited through the usual advertisement channels of the Confucius Institute at the University of Sheffield and through social media (Twitter, Instagram, and Facebook). As with all language learning studies presented in this thesis, this study relied on convenience sampling and participants to volunteer to take the language courses. Since the classes were bespoke to the current study both in terms of content and presentation, as many participants as possible were recruited within the scope of the Sheffield Confucius Institute resources and timeline of available funding.

In the Traditional classes students were taught in the traditional pedagogical approach of introducing both the character system and spoken language at the same time as well as using pinyin. As the Traditional group was not the focus of the research questions under investigation, the 19 students in these classes were removed from the present analysis. As originally randomised, see Table J1, there were 40 students in the Speaking-Writing class (M_{age} = 30.5, SD = 9.09) and 41 students in the Writing-Speaking class (M_{age} = 30.5, SD = 8.59). The demographic data of the 55 participants who completed the courses is summarised in the Table 5.1**.**

Figure 5.2 World map of participants in the online classes

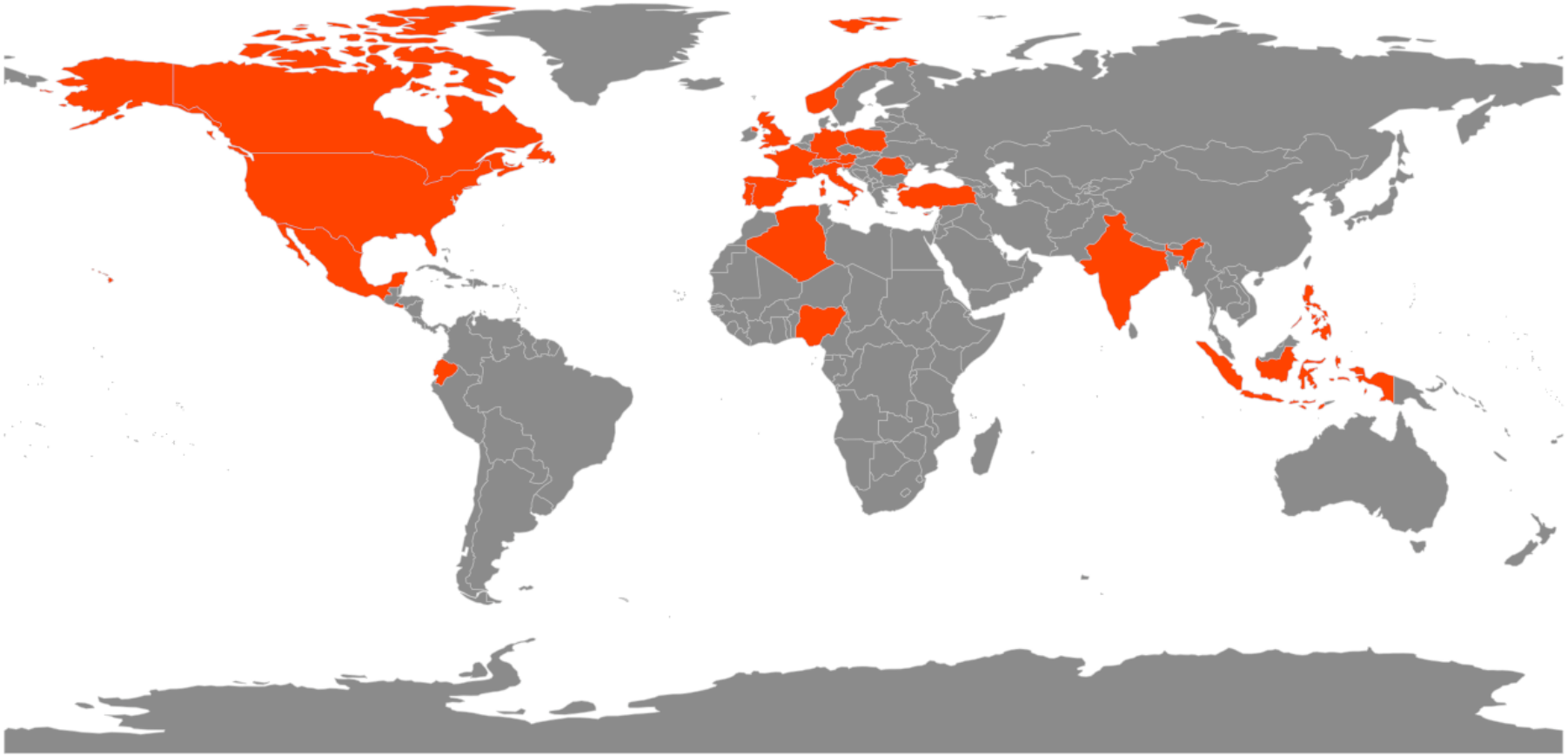


Table 5.1 Demographics of participants who completed the course

Characteristic	Overall, N = 55[1]	Class Order	
		SW, N = 27[1]	**WS**, N = 28[1]
Round			
Round 1	17 (31%)	8 (30%)	9 (32%)
Round 2	19 (35%)	9 (33%)	10 (36%)
Round 3	19 (35%)	10 (37%)	9 (32%)
Participant Age	31 (18, 60)	32 (20, 60)	30 (18, 48)
Handedness			
Left	1 (1.8%)	1 (3.7%)	—
Right	54 (98%)	26 (96%)	28 (100%)
Gender			
Female	28 (51%)	12 (44%)	16 (57%)
Male	26 (47%)	15 (56%)	11 (39%)
Prefer to self-describe	1 (1.8%)	—	1 (3.6%)
Education Level			
Primary school	1 (1.8%)	—	1 (3.6%)
Secondary/high school	2 (3.6%)	2 (7.4%)	—
Vocational training/ professional degree	1 (1.8%)	1 (3.7%)	—
Undergraduate Student	11 (20%)	5 (19%)	6 (21%)
University Degree	14 (25%)	10 (37%)	4 (14%)
Postgraduate Student	7 (13%)	4 (15%)	3 (11%)
Master's Degree	13 (24%)	4 (15%)	9 (32%)
PhD Student	2 (3.6%)	—	2 (7.1%)

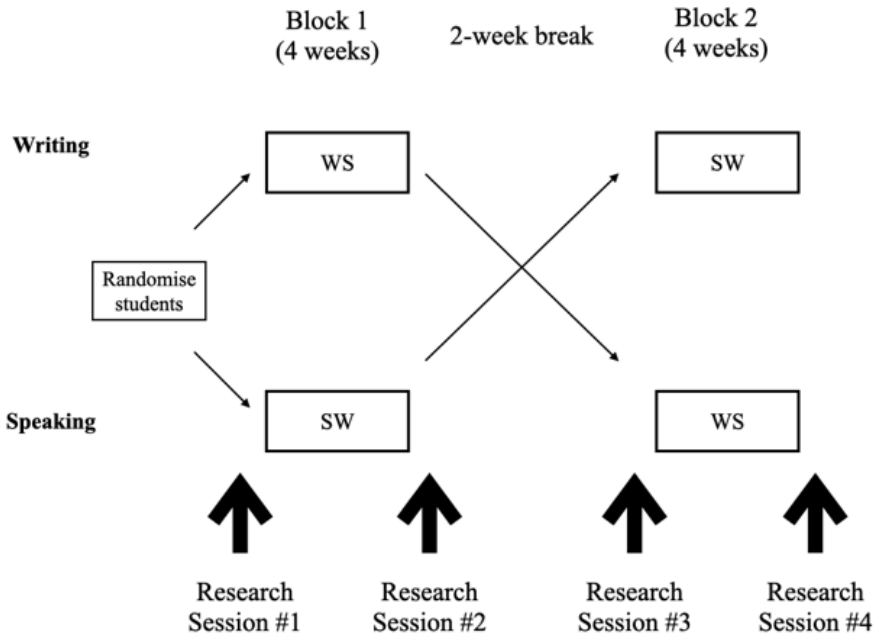
Characteristic	Overall, N = 55[1]	Class Order: SW, N = 27[1]	Class Order: WS, N = 28[1]
Doctorate	4 (7.3%)	1 (3.7%)	3 (11%)
No. Languages Before Class	4.04 (1, 8)	4.15 (1, 8)	3.92 (1, 8)

[1] n (%); Mean (Range)

5.2.3 Materials

Given the design of the study, data was collected at a total of 4 time points (i.e., before and after each class), shown in Figure 5.3. Each individual research sessions included questionnaires to gather demographic information, language attitudes (of specific interest in the present study attitudes towards language difficulty), and opinions on the courses as well as two cognitive tests. Of specific interest to the present study are the language attitude questions asked about the perceived difficulty of learning Chinese, which were rated on 6-point, fully labelled, ascending, horizontal Likert-style scales. The second research session also included Raven's APM as a measure of non-verbal general intelligence (see Section 2.6.5.2).

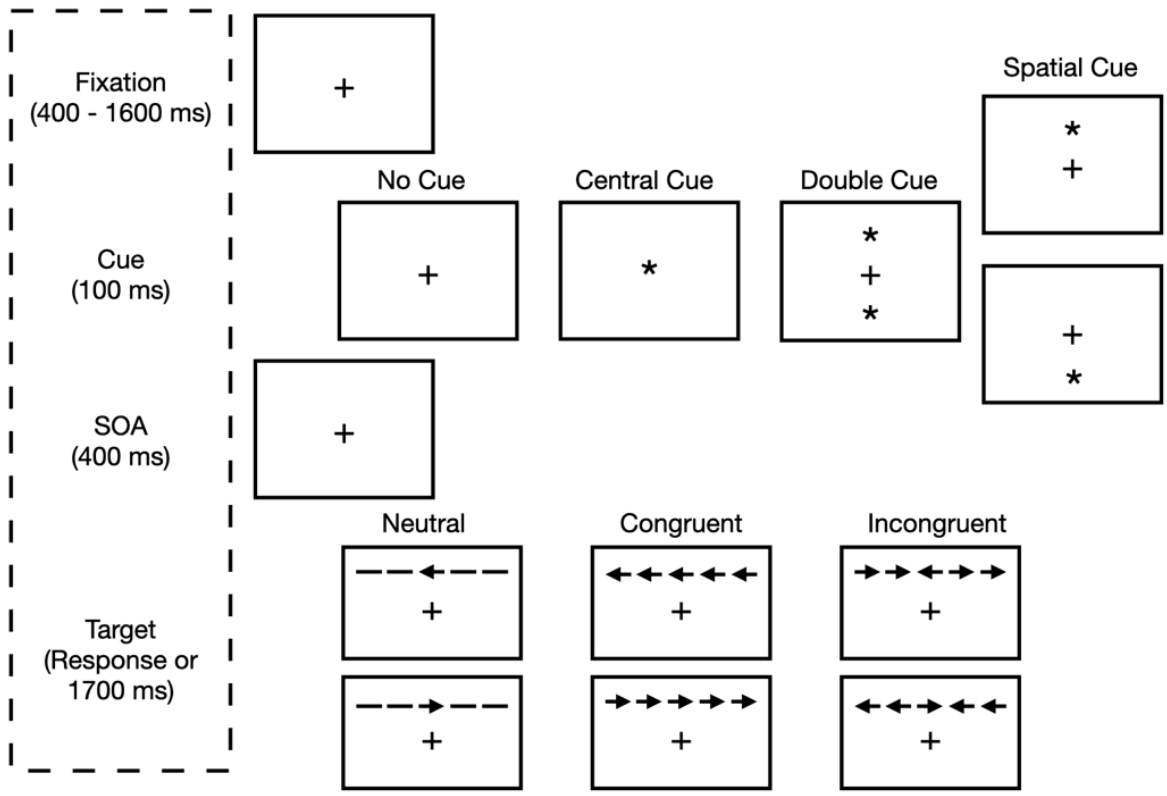
Figure 5.3 Visualisation of time points of data collection



Two measures of attentional functions were used in this study: the TEA (Robertson et al., 1994) and Attention Network Test (ANT) (Fan et al., 2002). As the two language classes include a different modality of language input (e.g., the speaking class being auditory and the written class being visual), so too did the cognitive measures. The TEA (as described in Section 2.6.5.1) is an auditory measure of attention whereas the ANT is a visual measure of attention. As with the TEA, the ANT was based on Posner and Peterson's (1990) neuro-anatomical multi-system attentional model. Therefore, both measures include three subtests. The ANT is a computer-based modified flanker task which simultaneously measures the efficiency of three attentional networks: alerting, orienting, and executive control (referred to as Conflict by Fan et al. (2002)), see Figure 5.4 for a diagram of the task. The experiment is comprised of 2 within trial factors: cue type (no cue, center cue, double cue, spatial cue) and

flanker type (neutral, congruent, incongruent). Participants are asked to indicate to the direction of the central pointing arrow, with their reaction time (RT; in ms) and accuracy being recorded for each trial. The full test consists of a 24-trial practice block with feedback and three experimental blocks without feedback. Each experimental block is comprised of 96 trials presented in random order (i.e., 2 repetitions of 4 cue conditions x 2 target locations x 2 target directions x 3 flanker conditions). In the in-person research sessions, the ANT was presented using E-Prime 2.0 (*E-Prime*, 2012) software whereas in the online research sessions the task was coded using the jsPsych JavaScript library (J. R. de Leeuw, 2015).

Figure 5.4 ANT experimental procedure



Note. Stimulus-onset asynchrony (SOA); the sequence of events and timing of a single trial is displayed in column with dotted lines and all possible stimuli associated with each event is displayed on the right. The four cue (warning) types as well as the three flanker (target) conditions are equally probable throughout the task. Targets appear above or below the fixation point with equal probability.

The three network scores of the ANT are calculated using orthogonal subtractions of performance (RTs) in selected conditions, as per Fan and colleagues (2002):

(i) Alerting: $RT_{no\ cue} - RT_{double\ cue}$

(ii) Orienting: $RT_{center\ cue} - RT_{spatial\ cue}$

(iii) Conflict: summed across cue types, $RT_{incongruent} - RT_{congruent}$

Since its publication in 2002, the ANT has been adopted by numerous studies, particularly within the neuropsychological literature (MacLeod et al., 2010). Within the literature on the cognitive effects of bilingualism specifically, Ware and colleagues (2020) meta-analysis of 170 studies found the conflict network of the ANT was highly sensitive and more consistently indicate "superior executive functioning in bilinguals" compared to Stroop, Simon, and Flanker tasks (p. 10). Indeed, the average effect size of ANT executive control network was significantly larger than Stroop, Simon, and Flanker tasks. Notably, MacLeod and colleagues (2010)'s psychometric analysis of 15 studies using the ANT concluded that the executive control network measure was the most reliable, followed by the orienting and alerting network measures. Using Magnetic Resonance Imaging, Xiao et al. (2016) found the ANT network measures to be significantly correlated with both global and regional brain network efficiency. In particular, the executive control network as measured by the ANT was found to be significantly related to global brain efficiency. In fact, across most studies of bilingualism and language learning using the ANT as an outcome measure of interest, the conflict scores (i.e., executive control network) in particular are used. As such, the executive control network is the primary measure of interest from the ANT within the present study, though the other two networks will also be analysed.

In terms of application in pre/post testing to evaluate the effects of an intervention or training on attentional functions, the ANT has been used as an outcome measure in studies of mindfulness training (Jha et al., 2007) and short-term meditation (Tang et al., 2007). Relating to the reliability of repeated measurement, Ishigami and Klein (2010, 2011) have conducted a series of studies using both young adults and older adults and found the ANT to be overall robust across 10 sessions of repeated testing.

The study was given ethical was approval by the University of Edinburgh PPLS Research Ethics Committee (ref: 185-1920/2).

5.2.4 Analysis

For the TEA, the same analytic approach was taken as described in Chapter 4 (see Section 4.2.4), including descriptive statistics followed by inferential statistics where appropriate with Holm-Bonferroni correct *p*-values being reported. To capture the repeated testing and hierarchical clustering of participants within classes within the data, linear mixed effects modelling was selected as the method of analysis. Models were fitted with a random intercept and random slope by session with participants nested within class; however, this produced either convergence issues or a singular fit, even in a null model suggesting the random effects structure was too complex to be supported by the data. In an iterative process, the random slope by session was removed, which did not resolve the convergence issues. Next, nesting within class was removed. As such, the random intercept by participant grouping structure remained for all models of TEA data reported below. In terms of fixed effects, for all models (both TEA and ANT), categorical covariates/predictors were included using treatment coding while numeric covariates/predictors were mean-centered for interpretability. As the primary question in the present research is the order effects of the classes, group by session interactions were explored in all models. All models were fit with Restricted Maximum Likelihood estimation and the bobyqa optimiser. As described in Section 4.2.4, inference criteria of these models use included case-based bootstrapping to compute 95% confidence intervals around the model parameter estimates.

For the ANT, the analytic strategy adopted mirrors that commonly found in the literature: first conducting a general analysis of RTs by cue type

and flanker condition, then exploring network effects individually (A. Costa et al., 2008, 2009; Ishigami & Klein, 2010, 2011; MacLeod et al., 2010). Rather than using ANOVAs which are pervasive in the ANT literature, linear mixed models were adopted to allow for a more effective modelling of the repeated measures of the data (i.e., random intercept by participant) as well as a random slope of session. All RTs < 200 ms and > 1500 ms for correct responses in each condition were removed (2.10% of the data). Network effects were analysed with linear mixed models with a random intercept by participant. Where appropriate, pairwise comparisons via least-squares means (AKA estimated marginal means) were computed to further explore main effects. Notably, as the conflict effect is calculated by subtracting RTs on congruent conditions from the RTs on incongruent conditions, lower scores reflect better conflict resolution. On the other hand, for the alerting and orienting effects, lower scores reflect poorer performance.

In order to explore the relationship between the ANT and TEA scores between groups, a correlation analysis was conducted. While the analysis was exploratory, FDR corrections for multiple testing were applied as described in Section 3.2.4. With correlation in particular, the more coefficients calculated the more likely it is to obtain a significant value due to random noise or an association which is accidental or spurious (Curtin & Schulz, 1998; Haig, 2003; Knudson & Lindsey, 2014). Therefore, instead of controlling for the more conservative FWER via Holm-Bonferroni corrections, FDR corrections were employed. Given the correlation between attentional measures was an exploratory analysis, the priority was to identify as many significant features as possible whilst incurring a relatively low proportion of type I errors (i.e., false positives). Indeed, for more exploratory analyses FDR is preferable over FWER as it has comparatively higher statistical power, meaning small and medium

effect sizes are more likely to be detected, which can then be analysed in future research (Benjamini et al., 2006; Benjamini & Hochberg, 1995).

Participant language attitudes were explored descriptively using data visualisation and percentages. Assumptions and case diagnostics were checked for all reported models. Statistical analyses and data visualisations were conducted using R Version 4.1.3 (R Core Team, 2022), the lme4 (Bates et al., 2015), lmerresampler (Loy et al., 2022), emmeans (Russel, 2022), rstatix (Kassambara, 2021), effectsize (Ben-Shachar et al., 2020), ggplot2 (Wickham, 2016), and patchwork (Pedersen, 2020) packages.

5.3 Results

5.3.1 Dropouts

The retention rate between the Speaking-Writing and Writing-Speaking classes was the same (68%). As seen in Table 5.2, the majority (85%) of dropouts took place in the first round of classes when the CoViD-19 lockdowns were first implemented. Of the remaining 4 participants who dropped out from the online classes, 3 never attended the first block of classes and 1 dropped out during the washout period due an increase in demands at work making the schedule of the second block of classes incompatible. Notably, being allocated the Writing class first did not result in an increase in dropouts, nor did participants systematically dropout after the Speaking class avoiding the Writing class, or vice versa. Furthermore, as seen in Table 5.3, the profile of participants who dropped out of the courses did not differ in any variable of interest from those who completed the courses, including cognitive measures.

Table 5.2 Demographics of participants who dropped out

Characteristic	Overall, N = 26[1]	Class Order	
		SW, N = 13[1]	**WS**, N = 13[1]
Round			
Round 1	22 (85%)	12 (92%)	10 (77%)
Round 2	2 (7.7%)	1 (7.7%)	1 (7.7%)
Round 3	2 (7.7%)	—	2 (15%)
Period of Dropout			
Before Class 1	5 (19%)	1 (7.7%)	4 (31%)
Block 1	14 (54%)	10 (77%)	4 (31%)
Washout	7 (26%)	2 (15%)	5 (38%)
Participant Age	29 (20, 51)	28 (20, 40)	31 (20, 51)
(Missing)	3	1	2
Gender			
Female	14 (56%)	9 (69%)	5 (42%)
Male	11 (44%)	4 (31%)	7 (58%)
(Missing)	1	—	1
Education Level			
Primary school	—	—	—
Secondary/high school	2 (8.7%)	1 (8.3%)	1 (9.1%)
Vocational training/ professional degree	3 (13%)	—	3 (27%)
Undergraduate Student	6 (26%)	4 (33%)	2 (18%)
University Degree	3 (13%)	2 (17%)	1 (9.1%)
Postgraduate Student	3 (13%)	1 (8.3%)	2 (18%)

Characteristic	Overall, N = 26[1]	Class Order	
		SW, N = 13[1]	WS, N = 13[1]
Master's Degree	4 (17%)	3 (25%)	1 (9.1%)
PhD Student	—	—	—
Doctorate	2 (8.7%)	1 (8.3%)	1 (9.1%)
(Missing)	3	1	2
No. Languages Before Class	2.96 (1, 5)	3.33 (1, 5)	2.55 (1, 5)

[1] n (%); Mean (Range)

Table 5.3 Demographics and baseline scores of participants who completed the course versus dropped out

		Completed the Study	
Characteristic	**Overall**, N = 81[1]	**No**, N = 26[1]	**Yes**, N = 55[1]
Group			
Speaking-Writing	40 (49%)	13 (50%)	27 (49%)
Writing-Speaking	41 (51%)	13 (50%)	28 (51%)
Round			
Round 1	39 (48%)	22 (85%)	17 (31%)
Round 2	21 (26%)	2 (7.7%)	19 (35%)
Round 3	21 (26%)	2 (7.7%)	19 (35%)
Participant Age	30 (18, 60)	29 (20, 51)	31 (18, 60)
(Missing)	3	3	—
Gender			
Female	42 (52%)	14 (56%)	28 (51%)
Male	37 (46%)	11 (44%)	26 (47%)
Prefer to self-describe	1 (1.3%)	—	1 (1.8%)
(Missing)	1	1	—
Pre-course 1 Elevator Task	98.39 (85.87, 100)	98.86 (85.87, 100)	98.18 (85.87, 100)
(Missing)	1	1	—
Pre-course 1 Elevator Task with Distraction	80 (10, 100)	76 (10, 100)	81 (10, 100)
(Missing)	1	1	—
Pre-course 1 Elevator Task with Reversal	66 (0, 100)	64 (0, 100)	72 (0, 100)

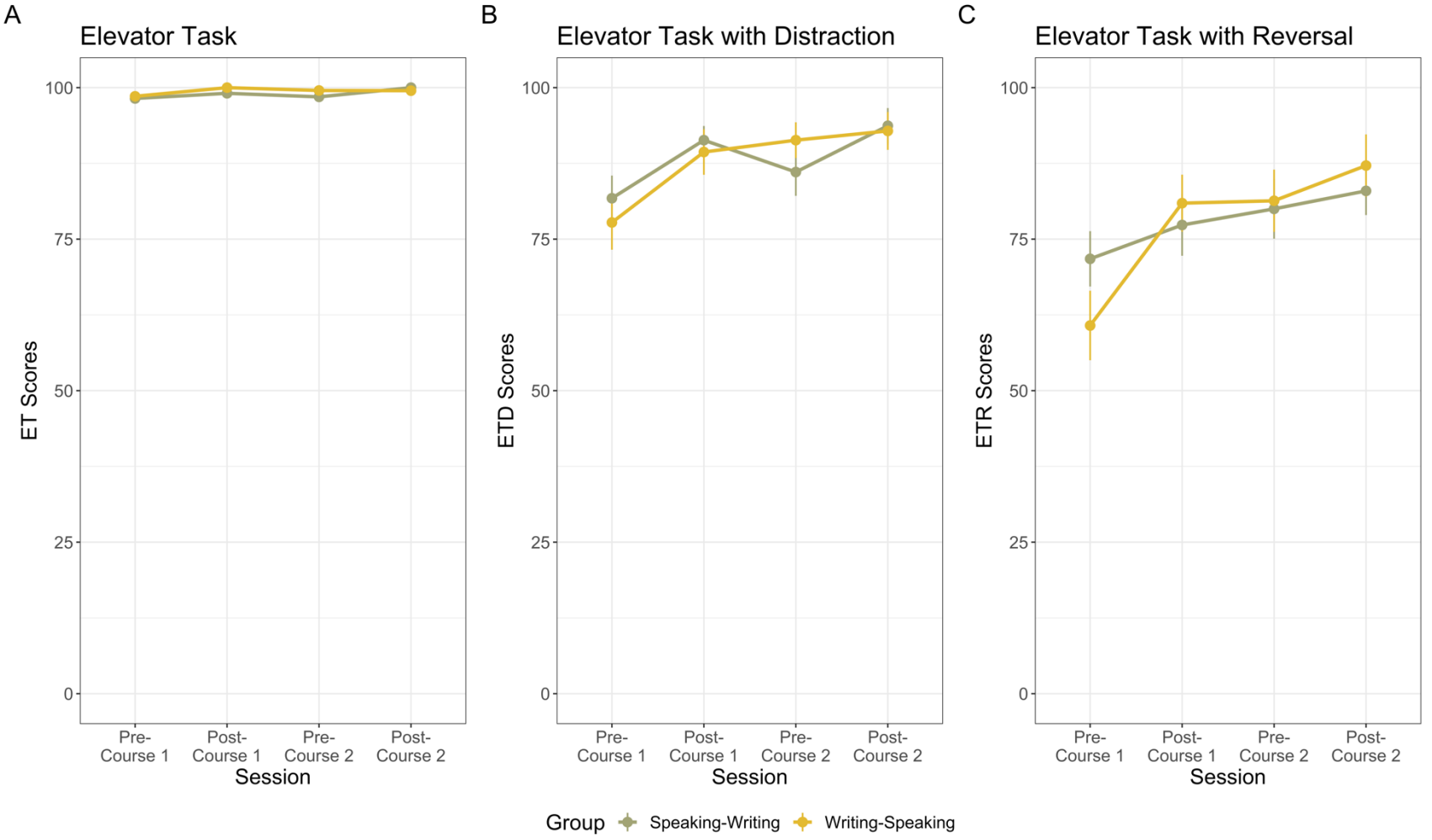
Characteristic	Overall, N = 81[1]	Completed the Study	
		No, N = 26[1]	Yes, N = 55[1]
(Missing)	1	1	—
Raven's APM Score	77 (15, 100)	68 (15, 90)	79 (16, 100)
(Missing)	19	19	—

[1] n (%); Mean (Range)

5.3.2 Auditory measures of attention

The overall pattern of scores across the four testing sessions in the three auditory measures of attention are shown in Figure 5.5 and discussed in further detail in the subsequent sections.

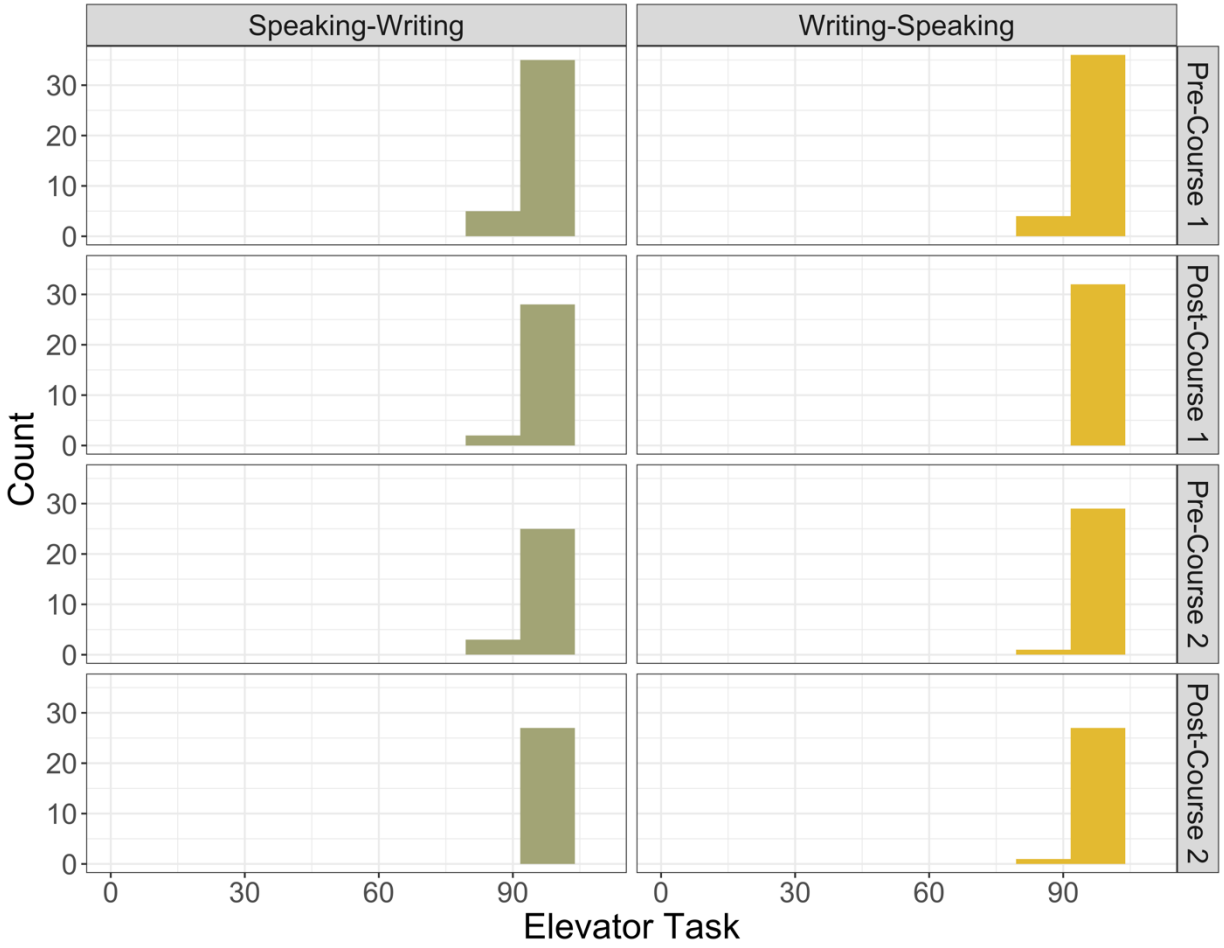
Figure 5.5 Test of Everyday Attention scores by group and session



5.3.2.1 Elevator Counting

The Elevator Counting task, as expected given previous literature and that it is the least computationally difficult of the auditory subtests, had a ceiling effect, shown in Figure 5.6 (Bak et al., 2016c; Long et al., 2020; Vega-Mendoza et al., 2015). As such, it was not further analysed as part of this study.

Figure 5.6 Elevator Counting scores by group and session



5.3.2.2 Elevator Counting with Distraction

The Elevator Counting with Distraction task showed more variability across sessions (Figure 5.7). At baseline, both groups scored similarly: Speaking-Writing group (*M* = 81.75, *SD* = 23.63, *Mdn* = 90), Writing-Speaking group (*M* = 77.75, *SD* = 28.42, *Mdn* = 90). Overall, as seen in Figure 5.5, there was an increase in scores before (*M* = 79.75, *SD* = 26.05, *Mdn* = 90) and after (*M* = 90.32, *SD* = 17.55, *Mdn* = 100) the first course, *W* = 309.5, *p* = .03, *r* = .42, with a moderate effect size. After the washout period, ETD scores remained essentially the same (*M* = 88.79, *SD* = 18.55, *Mdn* = 95), slightly increasing again after the second course (*M* = 93.27, *SD* = 15.76, *Mdn* = 100). Including both rounds of classes, ETD scores significantly improved from baseline measures, *W* = 147, *p* < .001, *r* = .52, with a large effect size. As such, a linear mixed model with random intercepts by participant was fit, reported in Table 5.4. Controlling for group, participant gender, participant age, quiz score, activities during the washout period, learning Chinese during the washout period, and hours spent practicing Chinese in during each class, testing session, language knowledge before the class (β = 0.18, cluster sample bootstrap 95% CI: 0.09 – 0.27), Raven's APM score (β = 0.23, cluster sample bootstrap 95% CI: 0.03 – 0.46) significantly predicted ETD scores. Compared to baseline, ETD scores were significantly higher after the first course (β = 10.00, cluster sample bootstrap 95% CI: 2.99 – 16.72) and after the second course (β = 12.61, cluster sample bootstrap 95% CI: 5.66 – 19.36). The group by session interaction was not significant.

Figure 5.7 Elevator Counting with Distraction scores by group and session

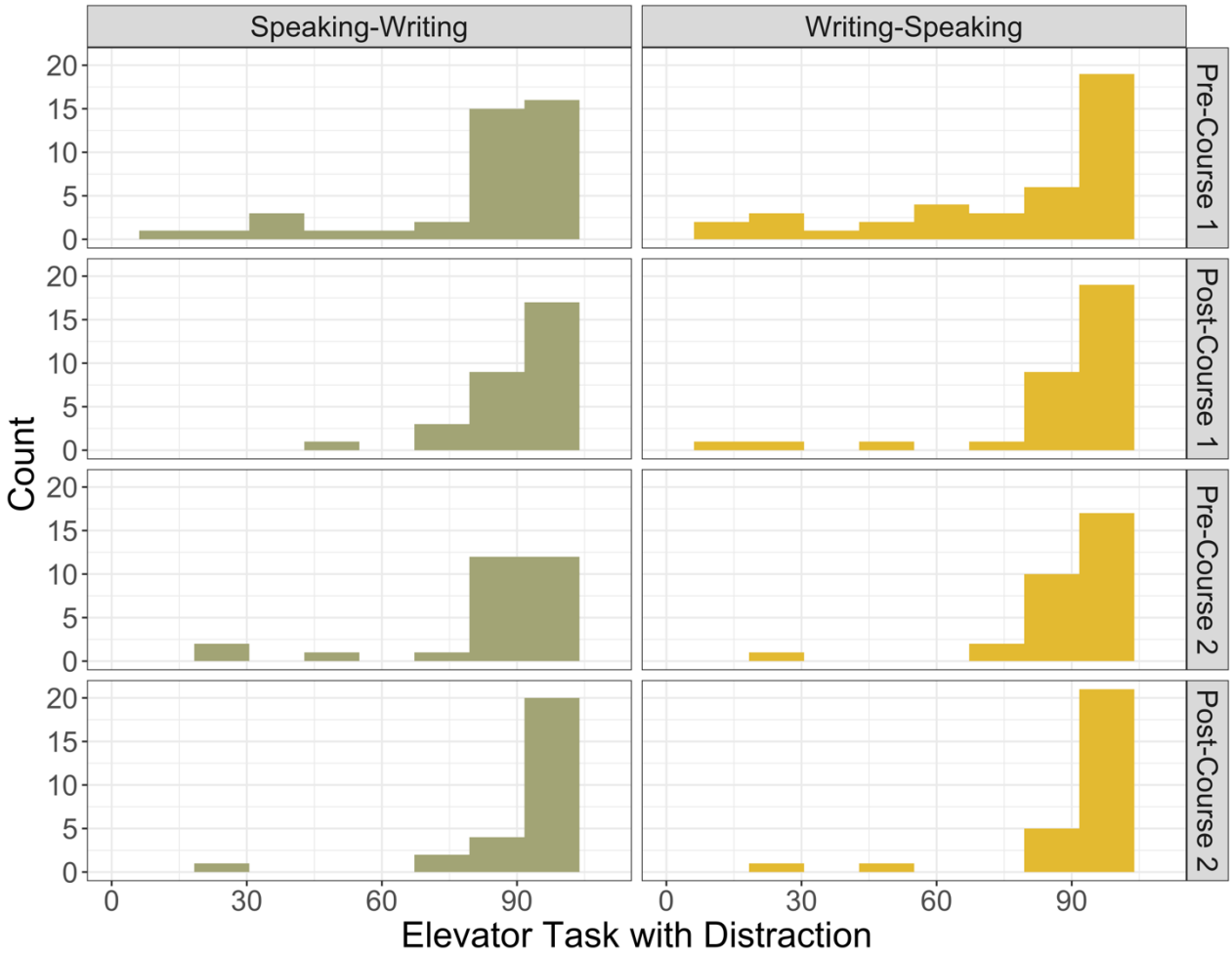


Table 5.4 ETD Model

Predictors	Beta	95% CI[1]
(Intercept)	52.93	**28.15, 77.13**
Session		
Pre-Course 1	—	—
Post-Course 1	10.00	**2.99, 16.72**
Pre-Course 2	4.35	-3.68, 12.37
Post-Course 2	12.61	**5.68, 19.36**
Group		
Speaking-Writing	—	—
Writing-Speaking	-3.68	-12.99, 5.67
Age of Participant	0.09	-0.16, 0.36
Composite Language Background Score	0.18	**0.09, 0.27**
Years of Education	-0.19	-0.90, 0.50
Gender		
Female	—	—
Male	-1.76	-6.64, 3.34
Quiz Score	0.01	-0.09, 0.12
Raven's APM Score	0.23	**0.03, 0.46**
Block 1 Hours Practicing Chinese	-0.18	-0.60, 0.20
Washout Period Activity (Hours per Week)	0.11	-0.12, 0.34
Block 2 Hours Practicing Chinese	0.32	-0.02, 0.66

Predictors	Beta	95% CI[1]
Learning before Class 2		
No	—	—
Yes	5.66	-0.01, 9.27
Group x Session		
Writing-Speaking x Post-Course 1	-5.22	-17.44, 7.07
Writing-Speaking x Pre-Course 2	3.04	-8.87, 15.22
Writing-Speaking x Post-Course 2	-3.04	-14.15, 8.94
Observations	184	
ICC	.38	
Marginal R^2 / Conditional R^2	.15 / .47	

[1] CI = Confidence Interval, case-based bootstrap with 5,000 samples

5.3.2.3 Elevator Counting with Reversal

The Elevator Counting with Reversal task also showed variability of scores across sessions (Figure 5.8). Prior the first course, the Speaking-Writing group (*M* = 71.75, *SD* = 28.90, *Mdn* = 80) scored higher than the Writing-Speaking group (*M* = 60.75, *SD* = 36.26, *Mdn* = 70), however this difference was not significant, $U(N_{SW} = 40, N_{WS} = 41) = 930.5$, $p = .206$. Overall, as seen in Figure 5.5, there was an increase in scores before (*M* = 66.25, *SD* = 33.02, *Mdn* = 80) and after (*M* = 79.19, *SD* = 27, *Mdn* = 90) the first course. However, after corrections for multiple testing this difference was not statistically

significant, $W = 309.5$, $p = .058$. After the washout period, ETR scores remained essentially the same ($M = 80.69$, $SD = 26.94$, $Mdn = 90$), significantly increasing again after the second course ($M = 85.09$, $SD = 26.94$, $Mdn = 100$) compared to the before to the second course, $W = 159$, $p = .044$, $r = .31$, with a moderate effect size. Including both rounds of classes, ETR scores significantly improved from baseline measures, $W = 147$, $p < .001$, $r = .57$, with a large effect size.

Therefore, a linear mixed model with random intercepts by participant was fit. The full fixed effects structure with a three-way interaction of group by session by participant age was multicollinear with extremely high VIF (variance inflation factor) score around 1000 (typically, a VIF ≥ 10 is deemed problematic). As such, the final fixed effect structure included an interaction between group and session, reported in Table 5.5. Controlling for group, participant gender, participant age, quiz score, activities during the washout period, learning Chinese during the washout period, and hours spent practicing Chinese in both classes, testing session, language knowledge before the class (β = 0.09, cluster sample bootstrap 95% CI: 0.01 – 0.18) and Raven's APM score (β = 0.74, cluster sample bootstrap 95% CI: 0.62 – 0.86) significantly predicted ETR scores. Compared to baseline, ETR scores were significantly higher after the second course (β = 11.74, cluster sample bootstrap 95% CI: 1.72 – 21.89), as can be seen descriptively in Figure 5.5 and Figure 5.8. The group by session interaction was not significant.

Figure 5.8 Elevator Counting with Reversal scores by group and session

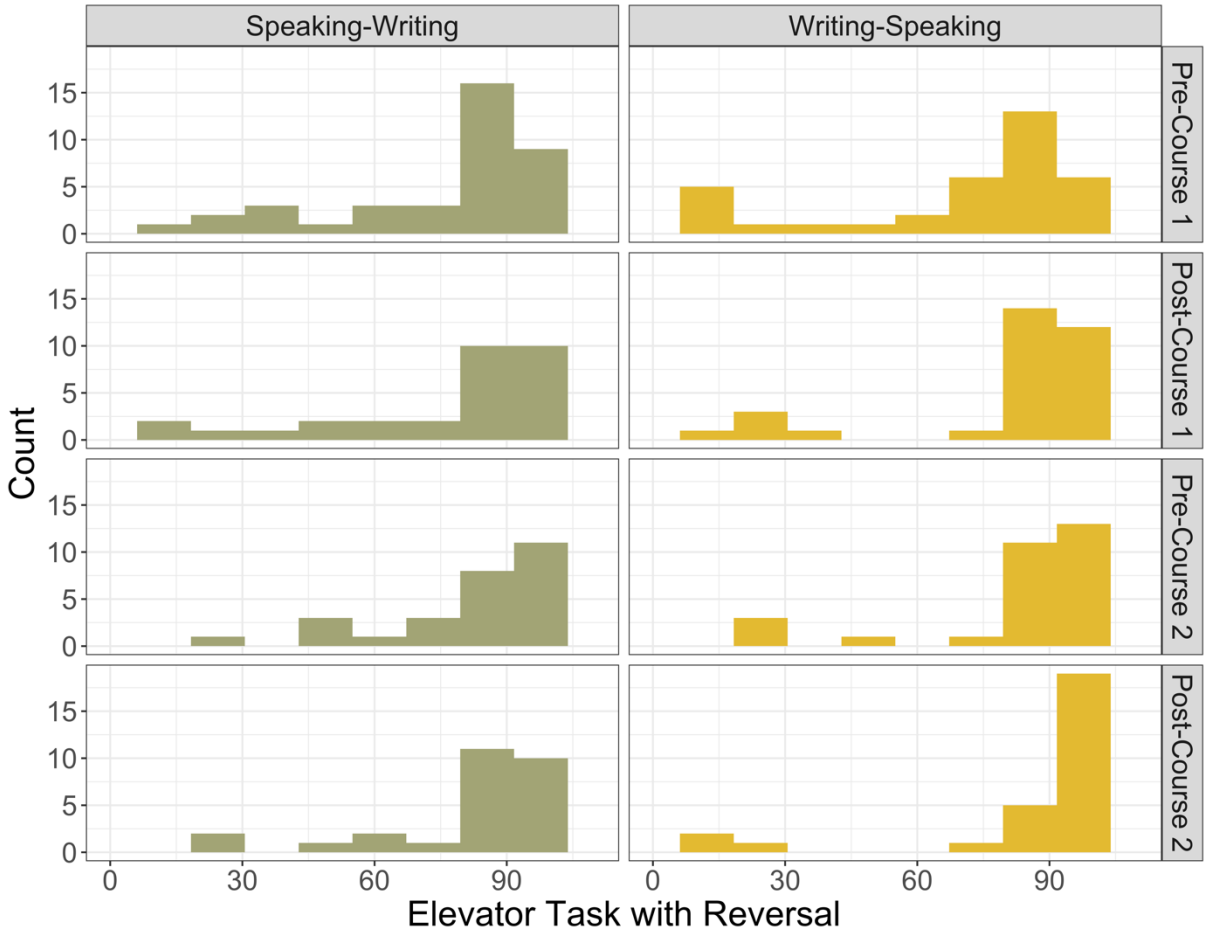


Table 5.5 ETR model

Predictors	**Beta**	**95% CI[1]**
(Intercept)	71.47	**49.77, 85.17**
Session		
Pre-course 1	—	—
Post-course 1	5.65	-5.20, 17.18
Pre-course 2	9.57	-1.29, 19.90
Post-course 2	11.74	**1.72, 21.89**
Group		
Speaking-Writing	—	—
Writing-Speaking	-12.83	-22.77, 2.96
Age of Participant	-0.19	-0.58, 0.22
Composite Language Background Score	0.09	**0.01, 0.18**
Years of Education	0.66	-0.20, 1.50
Gender		
Female	—	—
Male	-0.01	-4.39, 4.43
Quiz Score	0.08	-0.03, 0.19
Raven's APM Score	0.74	**0.62, 0.86**
Block 1 Hours Practicing Chinese	0.47	-0.24, 0.71
Washout Period Activity (Hours per Week)	0.21	-0.01, 0.41
Block 2 Hours Practicing Chinese	0.14	-0.27, 0.52

Predictors	Beta	95% CI[1]
Learning before Class 2		
No	—	—
Yes	4.94	-0.36, 9.95
Group x Session		
Writing-Speaking x Post-Course 1	3.91	-9.16, 16.63
Writing-Speaking x Pre-Course 2	-0.00003	-12.6, 12.06
Writing-Speaking x Post-Course 2	5.22	-6.73, 17.12
Observations	184	
ICC	.63	
Marginal R^2 / Conditional R^2	.28 / .74	

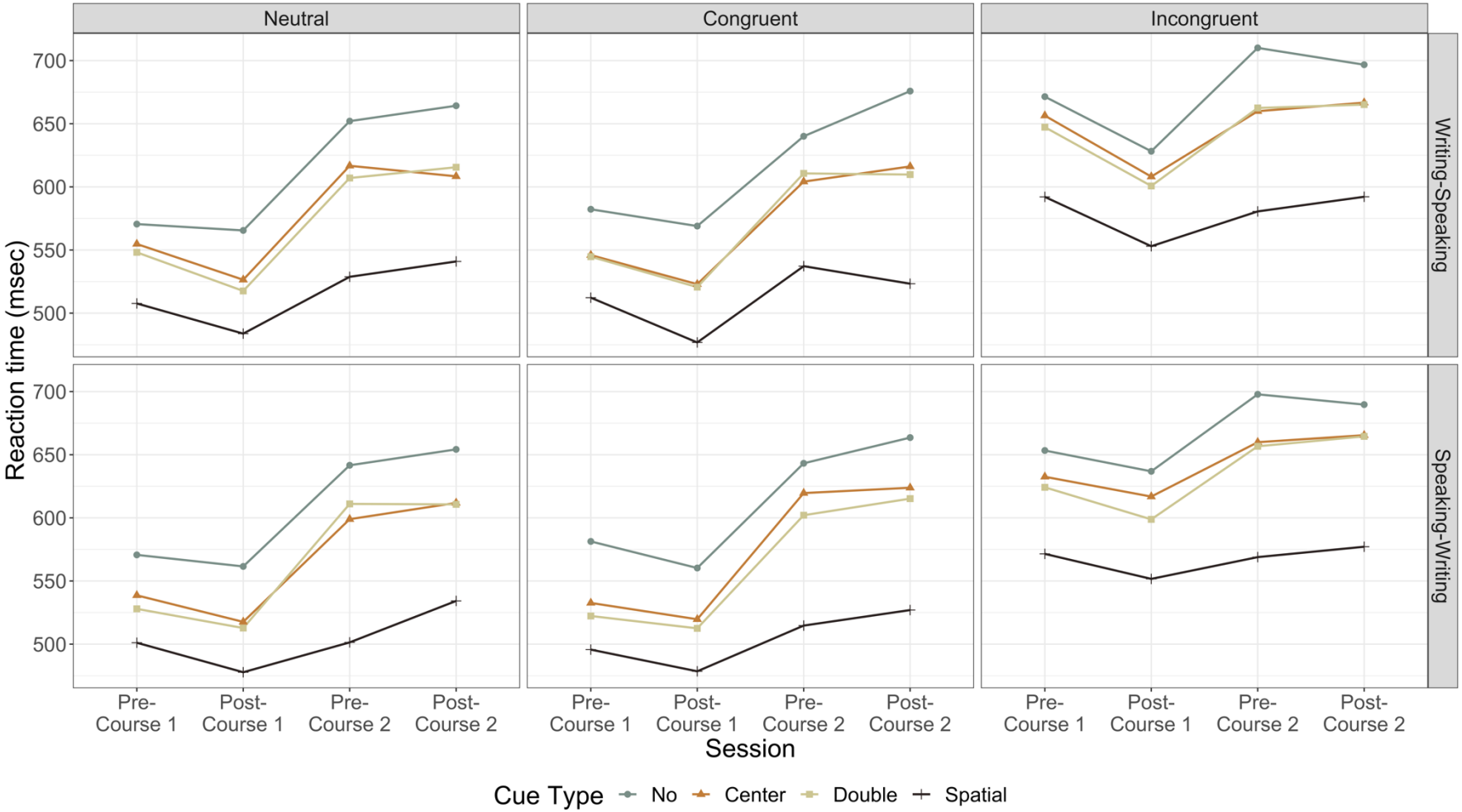
[1] CI = Confidence Interval, case-based bootstrap with 5,000 samples

5.3.3 Visual measures of attention

Figure 5.9 summarises the RTs by group from correct trials in Sessions 1 through 4. The efficiency of the three measures networks across sessions by group are shown in Figure 5.10. The full model of the general ANT analysis exploring a 4-way interaction between cue type (no cue, center, double, spatial), flanker condition (neutral, congruent, incongruent), session (pre-course 1, post-course 1, pre-course 2, post-course 2), and group (Speaking-Writing, Writing-Speaking) with random slopes by session and random intercepts by participant is reported in Appendix K (see Table K1).

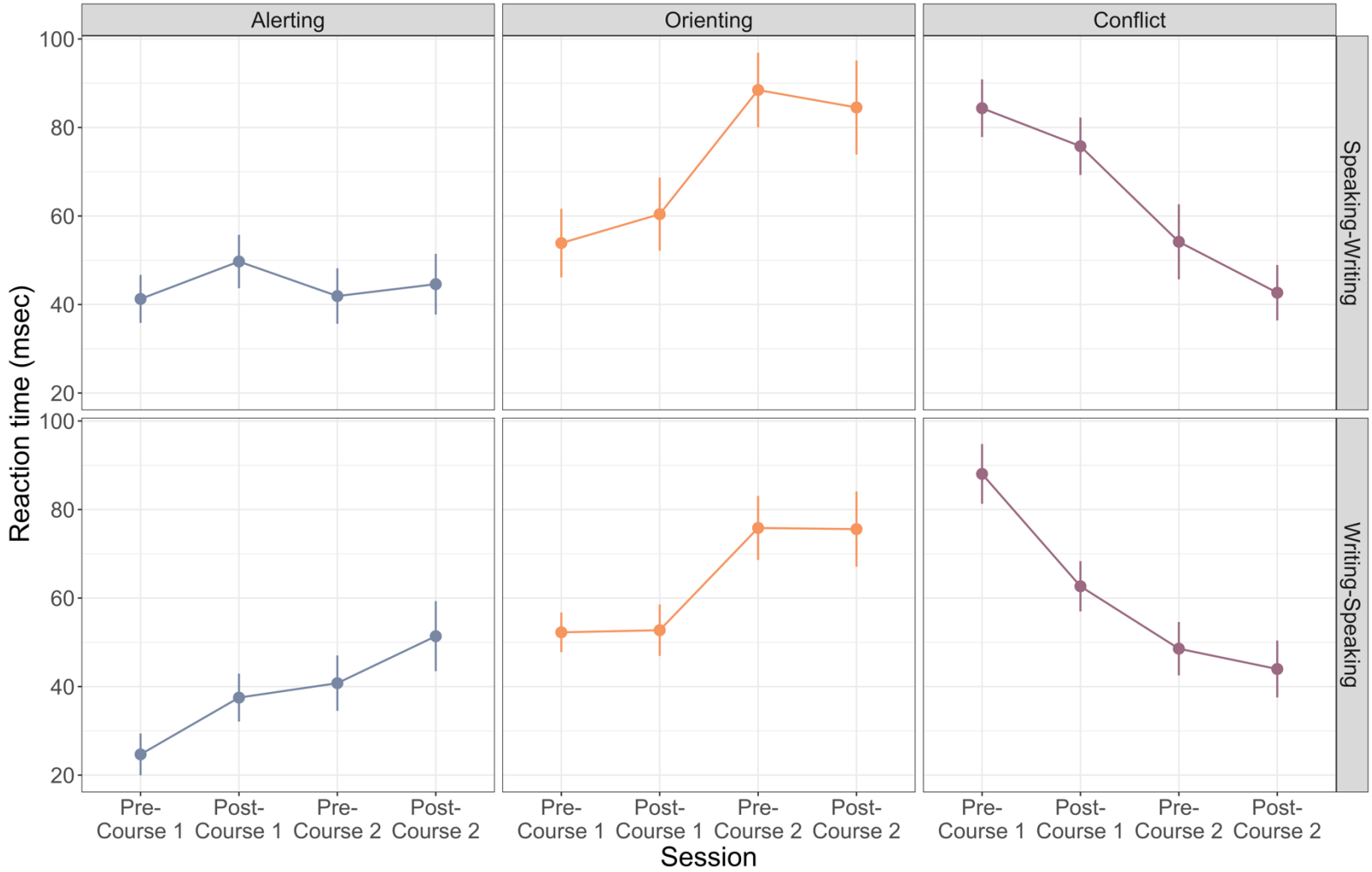
There was a significant main effect of (a) cue type, (b) flanker condition, and (c) session as well as interactions between (d) cue type and flanker condition, (e) cue type and session, (f) flanker condition and session, and (g) flanker condition, session, and group (see Appendix K for visualisations of significant interactions). The interactions, which moderate interpretations of the main effects, are of particular relevance to the research questions under investigation. The cue type by flanker condition interaction reveals that the congruency effect was increased when participants were alerted by non-spatial cues (e.g., center cue and double cue), β = -18, cluster sample bootstrap 95% CI: -36 – -0.30. The cue type and session interaction reflect a decrease in RT compared to baseline across testing sessions across all cue types compared to no cue. Post-hoc pairwise comparisons further clarified that in the washout period, increased speed in response to the spatial cues are driving the Pre-Course 2 to Post-Course 1 interaction. The flanker condition by session interaction showed that compared the baseline, incongruent conditions were responded to more quickly after course 1 (β = -40, cluster sample bootstrap 95% CI: -40 – -59), before course 2 (β = -49, cluster sample bootstrap 95% CI: -70 – -28), and after course 2 (β = -66, cluster sample bootstrap 95% CI: -87 – -45). In terms of the 3-way interaction, the interaction of flanker condition and session differs by group such that Post-Course 1, the Speaking-Writing group was slower to respond to incongruent trials than the Writing-Speaking group (β = 32, cluster sample bootstrap 95% CI: 5-59). No other 3-way interactions nor the 4-way interaction was significant in this model. The lack of an interaction between session and group suggests that the pattern of changes in RTs generally across sessions was not predicted by the order in which the classes were taken.

Figure 5.9 Mean RT from correct trials by cue type and flanker condition across testing sessions



Note: Lower values represent better (faster) performance or reaction times

Figure 5.10 Attention network scores by group and session



Note. Network scores calculated according to Fan et al. (2002). Error bars show the standard error of the mean. Conflict network scores are interpreted as lower values indicating better performance.

5.3.3.1 Alerting network

The alerting network scores, as seen in Figure 5.10, increase slightly across the testing sessions, though largely in the Writing-Speaking group. As seen in Table 5.6 the main effect of session is not significant in a linear mixed model with random intercept by participant. Visual inspection of Figure 5.10 shows a marked difference in baseline alerting scores before the Writing-Speaking and Speaking-Writing groups. However, controlling for covariates in the model there is not a main effect of group. Indeed, controlling for group, session, participant age, participant gender, years of education, quiz score, Raven's APM score, activities during the washout period, learning Chinese during the washout period, and hours spent practicing Chinese in each class, only language knowledge before the course was a significant predictor of alerting network scores (β = -0.29, cluster sample bootstrap 95% CI: -0.46 – -0.12). However, there is a significant interaction between group and session such that post-course 2 compared to baseline, the Writing-Speaking group alerting network scores are higher than the Speaking-Writing group (β = 23.12, cluster sample bootstrap 95% CI: 0.55 – 45.67) (see in Figure 5.10 and Appendix L).

Table 5.6 Alerting network model

Predictors	Beta	95% CI[1]
(Intercept)	56.72	**15.58, 98.29**
Group		
Speaking-Writing	—	—
Writing-Speaking	-10.18	-26.77, 5.04
Session		
Pre-course 1	—	—
Post-course 1	12.91	-3.14, 29.06
Pre-course 2	6.44	-9.01, 21.63
Post-course 2	5.88	-10.84, 21.93
Age of Participant	-0.13	-0.60, 0.55
Composite Language Background Score	-0.29	**-0.46, -0.12**
Years of Education	-0.06	-0.20, 1.50
Gender		
Female	—	—
Male	6.78	-15.27, 26.69
Raven's APM Score	-0.19	-0.51, 0.09
Quiz Score	0.22	-0.06, 0.46
Block 1 Hours Practicing Chinese	-0.19	-0.78, 0.38
Washout Period Activity (Hours per Week)	0.14	-0.35, 0.65
Block 2 Hours Practicing Chinese	0.01	-0.69, 0.79

Predictors	Beta	95% CI[1]
Learning before Class 2		
No	—	—
Yes	-7.93	-19.12, 2.06
Group x Session		
Writing-Speaking x Post-Course 1	-3.90	-23.47, 16.81
Writing-Speaking x Pre-Course 2	10.21	-11.24, 32.01
Writing-Speaking x Post-Course 2	23.12	**0.55, 45.67**
Observations	179	
ICC	.17	
Marginal R^2 / Conditional R^2	.16 / .30	

[1] CI = Confidence Interval, case-based bootstrap with 5,000 samples

5.3.3.2 Orienting Network

The orienting network scores also appear to increase over time as seen in Figure 5.10. A linear mixed model with random intercepts by participant was fitted, Table 5.7. Controlling for group, participant age, participant gender, years of education, language knowledge before the course, quiz score, Raven's APM score, activities during the washout period, learning Chinese during the washout period, and hours spent practicing Chinese in each class, only session was a significant predictor of orienting scores. Compared to baseline (Pre-Course 1), orienting network scores were significantly higher before the second course (β = 27.15, cluster sample bootstrap 95% CI: 10.03 – 45.02) and after

the second course (β = 26.49, cluster sample bootstrap 95% CI: 7.84 – 45.83). The group by session interaction was not significant.

Table 5.7 Orienting network model

Predictors	**Beta**	**95% CI[1]**
(Intercept)	116.94	**77.54, 154.75**
Group		
Speaking-Writing	—	—
Writing-Speaking	-16.29	-31.37, 1.07
Session		
Pre-course 1	—	—
Post-course 1	5.21	-13.77, 24.63
Pre-course 2	27.15	**10.03, 45.02**
Post-course 2	26.49	**7.84, 45.83**
Age of Participant	-1.27	-1.78, 0.70
Composite Language Background Score	-0.0004	-0.16, 0.15
Years of Education	-0.86	-2.55, 0.90
Gender		
Female	—	—
Male	-10.40	-18.81, 2.62
Raven's APM Score	0.16	-0.12, 0.43
Quiz Score	-0.25	-0.46, 0.05
Block 1 Hours Practicing Chinese	0.42	-0.17, 0.98

Predictors	Beta	95% CI[1]
Washout Period Activity (Hours per Week)	0.09	-0.38, 0.54
Block 2 Hours Practicing Chinese	0.33	-0.51, 1.17
Learning before Class 2		
No	—	—
Yes	-0.31	-10.52, 8.85
Group x Session		
Writing-Speaking x Post-Course 1	-3.62	-26.26, 17.92
Writing-Speaking x Pre-Course 2	-16.09	-37.19, 5.05
Writing-Speaking x Post-Course 2	-11.86	-35.87, 11.57
Observations	179	
ICC	.53	
Marginal R^2 / Conditional R^2	.21 / .63	

[1] CI = Confidence Interval, case-based bootstrap with 5,000 samples

5.3.3.3 Executive network

The conflict effect scores decrease over time, suggesting increased efficiency of the executive network across testing sessions. A linear mixed model with random intercepts by participants was fitted, see Table 5.8. Controlling for group, participant gender, years of education, Raven's APM score, quiz score, activities during the washout period, learning Chinese during

the washout period, and hours spent practicing Chinese in each class, session, participant age (β = 0.93, cluster sample bootstrap 95% CI: 0.41 – 1.29), and language knowledge before the course (β = -0.27, cluster sample bootstrap 95% CI: -0.41 – -0.13) were a significant predictors of conflict effect scores. Compared to baseline (Pre-Course 1), conflict network scores were more efficient before the second course (β = -31.31, cluster sample bootstrap 95% CI: -45.04 – -16.17) and after the second course (β = -38.95, cluster sample bootstrap 95% CI: -52.39 – -25.64). The group by session interaction was not significant.

Table 5.8 Executive control network model

Predictors	**Beta**	**95% CI[1]**
(Intercept)	79.79	**62.90, 96.67**
Group		
Speaking-Writing	—	—
Writing-Speaking	2.46	-11.94, 17.78
Session		
Pre-course 1	—	—
Post-course 1	-6.75	-20.97, 7.07
Pre-course 2	-31.31	**-45.04, -16.17**
Post-course 2	-38.95	**-52.39, -25.64**
Age of Participant	0.93	**0.41, 1.29**
Composite Language Background Score	-0.27	**-0.41, -0.13**
Years of Education	0.75	-0.65, 2.21
Gender		

Predictors	**Beta**	**95% CI[1]**
Female	—	—
Male	-7.65	-13.87, 0.25
Raven's APM Score	-0.01	-0.21, 0.19
Quiz Score	0.33	-0.12, 0.58
Block 1 Hours Practicing Chinese	0.64	-0.26, 1.08
Washout Period Activity (Hours per Week)	0.51	-0.18, 0.86
Block 2 Hours Practicing Chinese	-0.09	-0.83, 0.52
Learning before Class 2		
No	—	—
Yes	-12.44	-19.20, 4.27
Group x Session		
Writing-Speaking x Post-Course 1	-12.85	-31.73, 6.37
Writing-Speaking x Pre-Course 2	8.79	-11.63, 28.28
Writing-Speaking x Post-Course 2	10.57	-8.96, 30.22
Observations	179	
ICC	.42	
Marginal R^2 / Conditional R^2	.28 / .58	

[1] CI = Confidence Interval, case-based bootstrap with 5,000 samples

5.3.4 Exploratory relationship between visual and auditory measures of attention

Correlation matrices by session for the Speaking-Writing group are shown in Figure 5.11 and for the Writing-Speaking group in Figure 5.12. Of note, the standard deviation of the ET scores Post-Course 2 in the Speaking-Writing group and Post-Course 1 in the Writing-Speaking group is 0 (see Figure 5.6), therefore a correlation coefficient cannot be calculated. For these two time points, however, the ET variable has been included in the correlation matrix plot for consistency.

The two groups show different correlations across the visual and auditory measures. Further, the magnitudes and direction of correlations differ by group, particularly in testing session 4. While the two cognitive measures in the Writing-Speaking group did not indicate any significant correlation across the four testing sessions, this was not the case for the Speaking-Writing group scores. Prior to the writing course (pre-course 2 testing session), ETR and conflict network were negatively correlated, $r(28) = -.63$, $q < .001$ in the Speaking-Writing group. After the writing course (post-course 2 testing session), ETR and conflict network scores were again negatively correlated, $r(25) = -.59$, $q = .02$. As explained in Section 5.2.3 lower scores in the ANT conflict network scores are a marker of better performance, or more efficient conflict resolution, hence the negative correlation. Further, in the Speaking-Writing group after the writing course, ETD was negatively correlated with altering scores, $r(25) = -.54$, $q = .03$, and positively correlated with orienting scores $r(25) = .50$, $q = .03$.

Figure 5.11 Speaking-Writing group auditory and visual measures of attention correlations by session

Note: In Figure D, correlations were not able to be computed due to ET scores having a standard deviation of 0. A grey rounded square represents a statistically significant correlation according to the FDR corrected *q*-value.

Figure 5.12 Writing-Speaking group auditory and visual measures of attention correlations by session

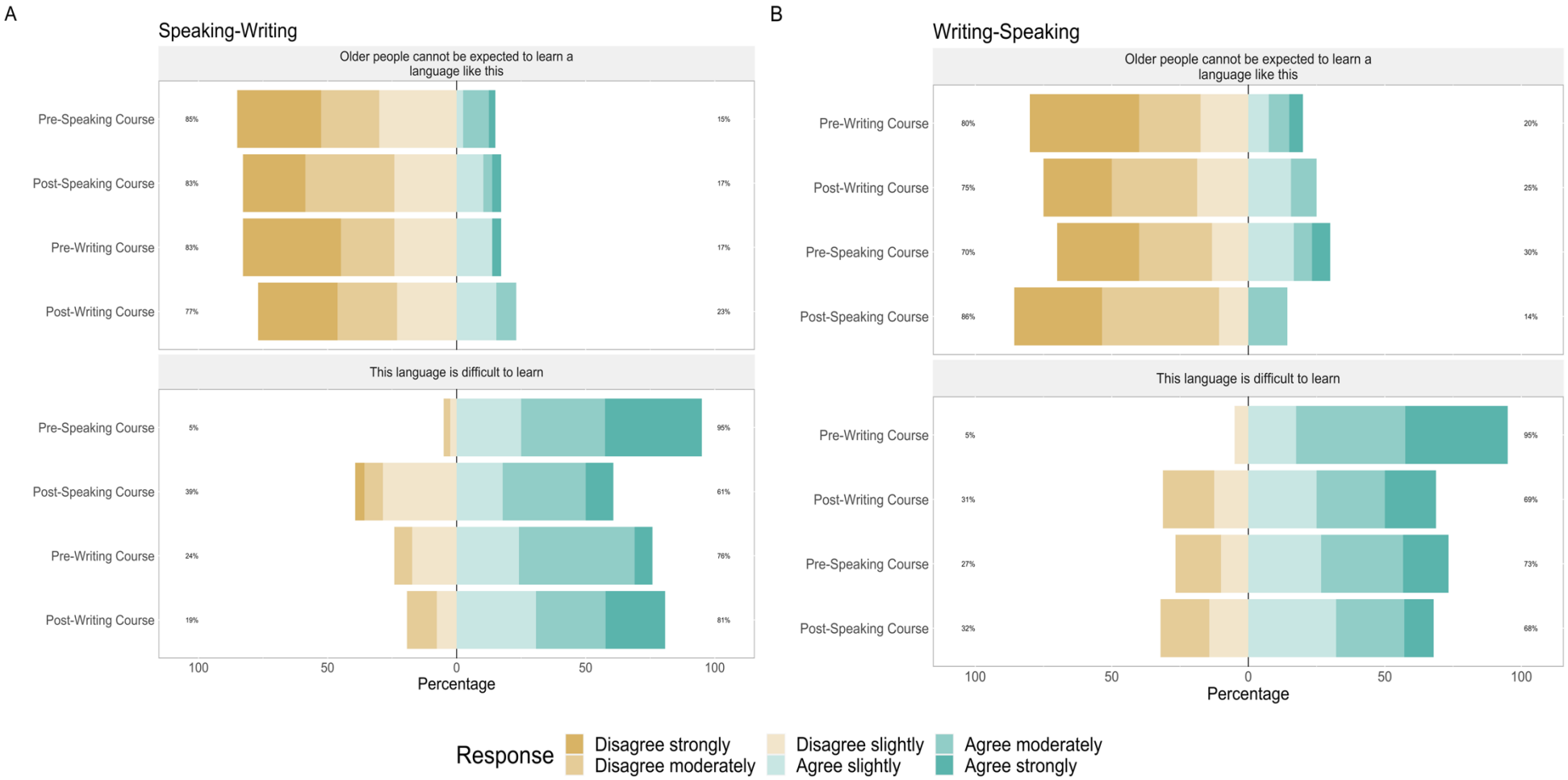
Note: In Figure B, correlations were not able to be computed due to ET scores having a standard deviation of 0. A grey rounded square represents a statistically significant correlation according to the FDR corrected *q*-value.

5.3.5 Language attitudes: Perceived difficulty

The data presented below relates to attitudes, therefore the presentation of results departs from the strict quantitative separation of results and discussion.

As shown in Figure 5.13, both groups highly agreed (95% agreement) with the statement "*Chinese is a difficult language to learn*". After the speaking course (course 1), the Speaking-Writing group showed a 34% decrease in agreement with this statement, which interestingly returned to higher agreement that the language was difficult to learn prior to the writing course. The Writing-Speaking group decreased 27% in agreement following the writing course, which remained relatively stable through the washout period and speaking course. By the end of both courses, the Speaking-Writing group perceived the language overall as more difficult to learn (81% agreement) than the Writing-Speaking group (68% agreement). This difference could explain the divergent pattern of responses to the first question, "*Older people cannot be expected to learn a language like this*". While overall both groups largely disagreement with this statement, by the end of the courses the Speaking-Writing group showed a 23% agreement that Chinese is not a suitable language for older adults to learn (compared to 14% agreement in the Writing-Speaking group), perhaps due to generally finding the language more difficult than the Writing-Speaking group.

Figure 5.13 Language attitude questions by group and session



Looking further into which aspects of the language were perceived as the most difficult element (grammar, pronunciation, vocabulary, or writing/characters), an interesting pattern develops which also differs by group (Figure 5.14). Prior to starting the courses, overall, both groups showed similarly high levels of agreement that Chinese grammar presented a challenge (Speaking-Writing 72% agreement, Writing-Speaking 68% agreement). However, after the first course, whether the speaking or writing class, participants showed a shift to lower levels of agreement (Speaking-Writing 51% decrease in agreement, Writing-Speaking 40% decrease in agreement). Across the washout period and writing class, the Speaking-Writing class remained relatively stable in this level of agreement of grammar being a difficulty in Chinese. The Writing-Speaking class, however, continued to show further overall decrease in agreement, with 82% of this group disagreeing that grammar was a primary difficult in learning Chinese by the end of the Speaking class.

Related to writing and characters, the Writing-Speaking group showed shift towards less agreement of writing being the main difficulty in learning Chinese (26% decrease in agreement) following the writing course. On the other hand, the Speaking-Writing group increased 3% in agreement. This is particularly interesting when considering that following the writing course, 100% of participants in the Writing-Speaking group thought pronunciation was the most difficult aspect of learning Chinese. Overall, pronunciation and writing were viewed as the main difficulties in learning Chinese both before the classes and after both courses, with a trend suggesting the order of classes influenced the strength of these opinions.

Figure 5.14 Perceived difficulty of learning Chinese by group across testing sessions

5.4 Discussion

This study aimed to explore if modality of language learning input, operationalised by uniquely designed 2-hour classes of Mandarin Chinese characters (visual) and spoken language (auditory) twice a week for 4 weeks influenced (a) cognitive effects measured in the trained modality and (b) student perceptions of difficulty of the target language.

In terms of the cognitive results, different trends emerged across the auditory and visual measures of attention. For the ETD scores, an auditory measure of inhibitory attention, participants overall improved after the first course and after the second course. In the auditory measure of attentional switching (ETR), scores after the first course were not significantly different from the baseline but did significantly improve after the second course.

Additionally, across both measures, participants with higher self-rated proficiency in other languages before the course and who performed better on the measure of general non-verbal intelligence (Raven's APM) scored higher. These two predictors were independent (r = .10), meaning participants whose scores were higher on Raven's APM were not more likely to be more multilingual, and vice versa.

Further, across both groups scores on the TEA remained stable after the first course and before the second course (i.e., during the 2-week wash-out period) and increased again after more language learning took place in the second course. There are two interesting potential implications of this finding. First, broadly speaking this confirms the lack of testing effects within the TEA auditory subtests. The finding that participants did not continue to improve during the washout period is in line with the purpose of the TEA to monitor natural history and rehabilitation of patients with brain injury (Robertson et al., 1994, 1996). Second, this pattern suggests that the cognitive effects of

language learning from a 4-week class as measured by the TEA are stable (i.e., not immediately lost) across 2 weeks without further language use.

The three network scores within the visual measure of attention (ANT) follow a somewhat different pattern than the auditory measure and do indicate evidence of practice effects. The primary network of interest in the present study was the executive control network as measured the conflict scores. The efficiency of the conflict network significantly increased before and after the second course compared to baseline in both groups. Visual inspection of Figure 5.10 suggests the Writing-Speaking group improved more than the Speaking-Writing group after course 1 (the writing class), though this was not statistically significant. The alerting and orienting network scores, which the literature suggests are less reliable, were less robust. The alerting network scores were only significantly higher after the second course compared to the baseline in the Writing-Speaking group compared to the Speaking-Writing group. Overall, the alerting network scores did not show any robust changes across the four testing time points. Lastly, the orienting network scores improved most during the wash-out period, indicating practice effects (Figure 5.10). Evidence of practice effects in the conflict network scores was also found as both groups continued to improve in performance during the washout period, though the difference in scores after the washout period (Pre-course 2) compared to before the washout period/after the first course (Post-Course 1) was not significant.

The general analysis of the ANT stimuli supports these network findings. Replicating previous literature, improved executive network efficiency was driven by faster conflict resolution (RTs) to incongruent flanker conditions and the orienting network efficiency due to faster RTs to spatial cues across testing sessions (A. Costa et al., 2008, 2009; Ishigami & Klein, 2010, 2011). Further, self-

rated command of languages before the course negatively predicted alerting network scores while it positively predicted executive network efficiency. That is, more multilingual participants before the course initially had slower alerting network scores, though faster conflict network scores. This partially replicates as well as partially contradicts Costa et al.'s (2008) finding that bilingual's have more efficient alerting and executive control networks as measured by the ANT.

Additionally, across all attentional measures, auditory and visual, age was a significant predictor only in the visual executive network measure. ANT conflict scores were slower as age increased, which post-hoc analyses reveal that older participants had slower reaction times when responding to incongruent trials in particular. In their study of comparing attention network efficiency as measured by the ANT of young adults, healthy ageing adults and adults with Alzheimer's Disease, Fernandez-Duque and Black (2006) found conflict resolution to be impaired in AD, but not in students (the young adult group) or healthy ageing adults. The current results do not suggest impairment with age, as the older participants also improved in executive network scores after the language classes. Rather, at baseline older participants showed slower conflict resolution efficiency compared to younger participants.

Moving to the question of modality specificity, overall, the order of classes as explored by the 'group' variable in the model building strategy was not a significant predictor of the attentional measures. Neither a main effect of group nor an interaction of group by session was found in the auditory measures of attention. For the visual measures, the general analysis of ANT stimuli suggested a three-way interaction whereby the flanker condition and session interaction are moderated by group (see Figure 5.9 and Appendix K). Specifically, the Writing-Speaking group was faster to respond to incongruent

trials after the writing course (Post-Course 1). The Speaking-Writing group also showed increased conflict resolution efficiency after the first course (Speaking course), though at a smaller interval than the Writing-Speaking group. After the Writing course (Post-Course 2), the Speaking-Writing group also showed faster response times to incongruent trials compared to before the second course, but not at the same rate as the Writing-Speaking group. This interaction suggests that there may be a mild modality effect in favour of the writing class resulting in greater efficiency of visual conflict resolution as measured by the ANT. Notably, however with transfer as there was also improved performance after the speaking course, though not the same degree. This is an interesting trend meriting future research to determine if indeed taking the Writing class (i.e., visual language input) first offers an initially greater improvement in response times to incongruent trails in a visual measure of attention over and above exclusively auditory input in the form of the Speaking class. This could, however, be due to individual differences between the groups, although group assignment was randomised, and therefore should be explored in future research to see if this pattern replicates.

Although we did find different cognitive effects of attention, they did not seem to be exclusively tied or bound to modality. In the visual measure (ANT), there is a visual trend (though statistically not significant) seen in Figure 5.10 wherein the performance on the executive control network scores improves at a greater rate following the Writing course. Taken together with the significant 3-way interaction discussed above, this suggests a mild modality effect in favour of visual input and measurement with transfer. There is overall evidence of transfer effects such that after the written course students improved in the auditory measure (TEA) and after the spoken course on the visual measure (ANT).

While there were measurable effects on attention after the language courses, the strength of these effects were not as strong as in the previous study (Chapter 4). This may be due to (a) baseline differences in the groups even with randomisation or (b) varied intensity and frequency per week. While the Spanish course included classes for 6 hours a week (3 times a week for 4 weeks), each block of the Chinese course involved only 4 hours of class a week (2 times a week for 4 weeks). It could, therefore, be the case that 4 hours a week alone (across 4 weeks) does not produce robust cognitive effects. While indeed some participants did improve both the auditory and visual measures of attention after the first class, across both cognitive measures the effects were strongest after the second class. These findings imply that while 4 hours across 4 weeks may not be a strong enough dosage of language learning, cognitive effects can more reliably be seen when maintaining that frequency but increasing the duration to 8 weeks, resulting in 32 hours total.

Additionally relating to group effects, the alerting network group by session interaction reflects an interesting pattern in the data whereby the Writing-Speaking group network scores increase across the four testing sessions though the Speaking-Writing group follows a different pattern. After the first course alerting network scores decrease (i.e., reflecting less efficiency), increasing again after the washout period and stabilising after the second course. Ishigami and Klein (2011) found that the reliability of the alerting network scores was significant only after more than 7 testing sessions. Indeed, MacLeod and colleagues (2010) concluded that of the three network scores, the alerting network was the least reliable: "Researchers and clinicians interested in using the ANT to examine individual differences and network-specific effects will be disappointed by the relatively low reliability of the network scores for orienting and alerting" (p. 647). Therefore, as the current

study included only 4 time points and the reliability of the alerting network scores themselves are questionable, this interaction should be interpreted cautiously.

Group or class order effects were found, however, in language attitudes. Writing and pronunciation were viewed generally as the greatest challenge to learning Chinese, which was regarded overall as a difficult language to learn. Notably, the Pre-course 1 questionnaire was completed prior to randomisation, so participants provided their language attitudes without knowing which class they would be taking first. Across both groups at baseline, writing was regarded as the most difficult aspect of learning Chinese (92% and 98% agreement). Interestingly, after the completing first class and prior to the second class, whichever class came second, participants rated as a more difficult aspect of the language. This is particularly notable as participants were randomised into these groups and were not aware of their group allocation at the baseline measurement.

It would be difficult to attribute any of these changes to a potential examiner effect. Throughout the collection of the language attitudes data both during Round 1 in person and Rounds 2 and 3 online (see Figure 5.1), the researcher was not present. During in person data collection, I stepped out of the room while participants completed the questionnaire and in the online delivery participants completed the questionnaire independently online.

This study is novel in its approach to teaching Chinese as a foreign language as well as due to the, albeit unplanned, inclusion of both in-person and online classes. The effectiveness from the perspectives of the students and teachers of online versus in person learning is discussed in Zhao et al (2020). To my knowledge, the present study is the first to directly compare performance on the TEA as an auditory measure of attention and the ANT as a

visual measure. Ishigami and colleagues (2016) compared older adults' (aged 55-80) performance on the ANT-I (Attention Network Test-Interaction which includes auditory and visual cues within the same test) and a variety of standardised neuropsychological tests. Similarly, van der Leeuw and colleagues (2017) compared older adults' (aged 80 or older) performance on the 3 auditory subtests of the TEA with an assortment of standardised neuropsychological tests. While Ooi and colleagues (2018) used both the ANT and TEA in their study looking at attentional control among bilinguals with different switching frequencies and contexts, the ANT and TEA were not analysed in relation to one another.

In the present study, statistically significant associations between the auditory and visual measures of attention were found only in the Speaking-Writing group and only in testing sessions before course 2 and after course 2. After the second course, there was significant positive correlation between ETD and orienting scores, though a negative relationship between ETD and alerting scores in the Speaking-Writing group. As both the alerting and orienting network scores are interpreted the same, there is not a clear reason why the direction of the relationship with ETD scores would be different. Nor do the aspects of attention these measures are intended to theoretically measure suggest a varied direction in the potential relationship. Considering that session was a significant predictor in the ETD and orienting models (see Table 5.4 and Table 5.7) though not in the alerting model (see Table 5.6), it may be that case that the positive relationship between ETD scores and orienting network scores reflects the fact that both improved across testing sessions, particularly in session 4. The negative correlation with alerting network scores may reflect the relative stability of the observed scores while the ETD scores improved. However, these relationships are only observed in one testing

session in one group, and thus should not be given substantial weight and be interpreted with caution.

A negative correlation between ETR scores and executive network scores was the only relationship to be observed in more than one testing session – before the writing course (Pre-Course 2) and after the writing course (Post-Course 2) only in the Speaking-Writing group. As higher ETR scores reflect improve attentional switching and lower executive network scores reflect more efficient conflict resolution, the negative correlation suggests that as ETR scores improve so too do conflict network scores. Both the ETR and conflict network subtests are the most cognitively demanding measures of attention in their respective testing battery and therefore show the greatest variability in scores. It is therefore not surprising that these two measures were found to correlate in the Speaking-Writing group. However, there were no significant correlations between auditory and visual measures of attention in the Writing-Speaking group. Further, the magnitude of correlations was different between the two groups across testing sessions, particularly notable in fourth testing session after course 2. In the Speaking-Writing group there are 7 total correlation coefficients in the matrix with an *r* of .4 or higher (i.e., medium to large effect sizes), while in the Writing-Speaking group there are only 2 coefficients .4 or higher. Particularly striking is different magnitude and direction between correlation coefficients between groups among ETR and conflict scores. While the negative correlation between ETR and conflict scores is large (r = .-59), the relationship within the Writing-Speaking group is nearly negligible (r = .05) (see Figure 5.11 and Figure 5.12). The group differences in correlations between the auditory and visual measures of attention (in terms of magnitude as well as the direction of correlations), in combination the lack of a main effect of group in the individual models suggest that the correlations

may be an artefact in the data or should at minimum be interpreted with caution. An analytic extension of the analysis presented in this chapter would be a crossed-lagged panel model.

However, there are notable limitations of present study. While great lengths were taken to randomise the groups based on potentially relevant demographic variables (i.e., age, gender, education level), there were still differences in baseline scores of the TEA and ANT, particularly in the ETR and alerting networks scores, between groups. Study design and resources permitting, future studies where these measures are a primary outcome variable should include cognitive test scores as a stratum in the stratified randomisation procedure. Additionally, dropouts and participant attrition across the study, while unavoidable, undermine statistical power. It is worth noting in terms of feasibility of teaching Chinese characters and spoken language as separate systems, though, that the majority of dropouts as reported by participants were due to CoViD-19. It can be postulated that if this study has taken place earlier, the in-person groups would likely have had lower rate of dropouts, but there is no way of confirming this hypothesis. However, cross-over trials often encounter difficulties with retention due to the length of time in which the study must take place in order to include 2 (or more) treatment periods and a washout period (Senn, 2002). Future studies with larger samples to complete all four research sessions may be better powered to explore if the findings presented here are robust and disentangle possible artefacts in the data.

While the strictest possible separation between the spoken and written language modality was an essential feature of the experimental design of this study, we recognise that in the typical language classroom, from both practical and pedagogical perspectives, the written and spoken languages are unable

to be effectively dissociated and experimentally controlled. The results of this study are not intended to advance that spoken and written Chinese should only be taught separately.

In conclusion, learning an entirely visual language system in the form of Chinese characters did not have an exclusive effect on visually measured attentional functions. Similarly, receiving exclusively auditory input in the form of the spoken Mandarin Chinese did not have an exclusive effect on auditorily measured attentional functions. A trend suggesting a slightly modality effect which was more visual than auditory was found. Notably, at the end of the study participants did improve in both visual and auditory measures of attention. This suggests that the cognitive effects of language learning, even when learning is modality specific, are generalisable across modalities.

Chapter 6 Learning Scots Across the Lifespan: How Language Status Influences Learning

In the previous two chapters, language learning studies were presented using the two most widely spoken languages in the world apart from English: Spanish and Chinese. In this chapter, we now turn our attention to the other end of the spectrum: a minority language which is still struggling to be recognised as a language. This chapter examines the issue of how the social status of a language might influence the learning process both in terms of student opinions towards the target language and the cognitive effects of learning the language, using Scots which is debated in some circles as being a dialect of English rather than a language.

6.1 Introduction

6.1.1 Linguistics prestige, function, and status

Prestige, function, and status are three terms which can converge or dissociate in the description of a language or linguistic system. Mackey (1989) distinguishes prestige, function and status as

> "the difference between past, present, and future. The prestige of a language depends on its record, or what people think its record to have been. The function of a language is what people acutally do with it. The status of a language depends on what people can do with it, its potential. Status therefore is the sum total of what you can do with a language – legally, culturally, economically, politically, and, of course, demographically. This is not necessarily the same as what you do with the language, although the two notions are obviously related, and indeed interdependent. They can also be connected with the prestige of the language" (p. 4).

These concepts are not always clearly delineated as they are interconnected. For example, by changing the functions of a language its status can too be modified which in the long-term may have an impact on prestige. Mackey (1989) goes on to illustrate the differences between these terms with the following examples: "Classical Latin has had a lot of prestige but it has few functions. Swahili has a lot of functions but little prestige. Irish Gaelic has status, official status, but few exclusive functions" (p. 4).

Linguistic prestige (and stigma) can also be thought of as the social value or socially perceived worth of a language. In the larger field of sociolinguistics, one of the key means of distinguishing concepts of prestige or overall social acceptance of a language/speech variety is the idea of covert and overt prestige, popularised by Labov (1966, 1997) and Trudgill (1972). Overt prestige is typically associated with the standard variety of a language which is generally acknowledged to be the socially "correct" or "proper" form (Trudgill, 1972; Van Herk, 2018). Alternatively, covert prestige is "the linguistic equivalent of street credibility" (Van Herk, 2018, p. 56). That is, typically non-standard language varieties regarded as high in prestige by members of a speech community, typically relating to group identity.

Evident by the classification of overt and covert prestige, the larger notion of prestige cannot be removed from social, political, and economic forces. That is, power and prestige are directly associated (Bonfiglio, 2002; Van Herk, 2018). Indeed, social and linguistic prestige are typically correlated such that a language which low social prestige will also have lower linguistic prestige, and vice versa. As Bonfiglio (2002) describes "there is nothing in the particular language itself that determines its worth: it is the connection of the language in question to the phenomena of power that determines the value of that language and that contributes to the standardization process. The

language in and *of* power then becomes an ideal" (p. 23). The standardization process links to a language's function and ultimately its status linguistically, culturally, and legally. For example, the standard variety of a language is defined as one which carries the highest status and (overt) prestige in its speech community (Van Herk, 2018). This often accompanies the language variety being codified, or written down, and subsequently used for official purposes such as law, politics, and the in the media. Therefore, in the case of standard varieties there is high status and high prestige which results in increased function.

Within the concepts of status as operationalised by Mackey (1989), there are multiple types. A language can have varying levels of linguistic, demographic, cultural, and legal status, each potentially developing from extra-linguistic sources such as social and political forces. Linguistic status is the ability to use to the language due to standardisation and/or orthographic conventions while demographic status is the number of speakers or people who understand the language (i.e., size). Further, cultural status is the variety of cultural products such as books, films, papers, etc. and cultural activities in which the language is used by the speech community. Lastly, legal status is the official or legal recognition given to the language by governing bodies.

Another aspect of status as future looking is language vitality or endangerment. Indeed, Ethnologue defines language status by the size of the speech community (cultural status), official recognition (legal status), and language vitality ('Ethnologue: Languages of the World.', 2022). Speaker population is divided into 3 levels: large with more than 1,000,000 users, mid-sized with 10,000 to 1,000,000 users, and small with fewer than 10,000 users. Language vitality is categorised according to the Expanded Graded Intergenerational Disruption Scale (EGIDS), which was developed to provide a

comprehensive, multidimensional assessment of the state of the world's languages in terms of development versus endangerment (Lewis & Simons, 2010). The EGIDS evaluates the endangerment and vitality of the world's languages on a 13-level scale, expanding upon Fishman's (1991) 8-level scale (Lewis & Simons, 2010). As the EGIDS measures the level of disruption of intergenerational transmission of the language in question, the higher the rating on the scale, the weaker and more endangered the language (Fishman, 1991; Lewis & Simons, 2010). Ethnologue characterises the vitality of a language by grouping the EGDIS levels into 4 broader categories:

1. "Institutional (EGIDS 0-4): the language has been developed to the point that it is used and sustained by institutions beyond the home and community.
2. Stable (EGIDS 5-6a): the language is not being sustained by formal institutions, but it is still the norm in the home and community that all children learn and use the language.
3. Endangered (EGIDS 6b-9): it is no longer the norm that children learn and use this language.
4. Extinct (EGIDS 10): the language has fallen completely out of use and no one retains a sense of ethnic identity associated with the language" ('Ethnologue: Languages of the World.', 2022).

According to this operationalisation of status in terms of language vitality as well as Mackey (1989)'s definitions, the two languages in the preceding language learning chapters are both high in status. Both Spanish and Mandarin Chinese have a large speaker population (more than 1,000,000 users) and are categorised as institutional languages in terms of vitality ('Ethnologue: Languages of the World.', 2022). Indeed, Spanish and Mandarin Chinese also have high function and high levels of prestige. However, in the

present study the focus shifts to a language with a different status and prestige, namely Scots. According to the Ethnologue, Scots has a large speaker population, though is stable rather than institutional in terms of language vitality ('Ethnologue: Languages of the World.', 2022). Scots is an example of a language with low prestige (both overt and covert), increasing legal status, substantial demographic status (number of speakers), and often leisure or home/community functions.

6.1.2 A brief history of Scots

Scots, a Germanic language, is one of the four national languages of Scotland which include Scots, Scottish Gaelic, British Sign Language, and English. Both Scots, historically spoken in the Lowlands of Scotland, and Scottish Gaelic, a Celtic language historically dominant in the Highlands and Islands of Scotland, are recognised as minority languages (*Languages - Gov.Scot*, n.d.). In fact, by national percentage of speakers, Scots is one of the most widely spoken minority native languages in Europe (Central Intelligence Agency, n.d.). Further, both Scots and Scottish Gaelic are recognised as vulnerable languages by UNESCO (i.e., "vulnerable" and "definitely endangered", respectively) (L. Evans, 2011; UNESCO, 2010). However, unlike Scottish Gaelic, Scots is considered by some as "a slovenly debased dialect… corrupt English… coarse slang… a language that has never existed" (Kay, 1986, p. 13) or "a joke, a weirdly spelled and -accented local variety of English" (Matheson & Matheson, 2000; Nosowitz, 2018).

Indeed, English and Scots share a common linguistic ancestor: Old English (Stuart-Smith, 2004). The two languages largely diverged from Early Middle English, when the Norman Conquest in the mid-11th century introduced heavy French influence on English. Consequently, much of the lexicon and

phonology of Old English remained in Scots, but disappeared from English (Kay, 1986). Indeed, according to Kay (1986), "Scots and English were dialects arising from a common root which developed independently due to political, rather than linguistic factors" (p. 59). As with English, Scots has several dialects which are mutually comprehensible. However, unlike English, there is currently not a unified orthography or recognised modern standard, leading to many debating the delineation and definition of Scots as a language (J. Costa, 2015).

Historically, Scots was a widely spoken and utilised language, recognised as related though separate from English (Johnston Jr, 2007). Indeed, in the sixteenth century Scots was used by the royal administration both in official documents and proceedings as well as in various works of literature (Millar, 2006). Scots has enjoyed a rich and diverse literary history, particularly within poetry. As Johnson Jr (2007) outlines, Scots has "a good-sized and sometimes brilliant corpus of literature from the early fourteenth to the early seventeenth centuries, exhibiting a range of genres, styles and registers comparable to any Western European national language" (p. 106). However, by the 18th century the relationship between Scots and English had become increasingly diglossic, with Scots progressively regarded as a dialect of the socially and politically dominant Standard Scottish English (Millar, 2006). With the unification of Scotland and England in 1603 under James I (previously King James VI of Scotland), English became the language of power and prestige, and soon after the language of religion as mass-printed Bibles were available only in English. As Kay (1986) describes, "from then on, God spoke English" (p. 59). In fact, John Knox, father figure of the Reformation in Scotland, "is the first known Scot in history to attempt to conform his writing to English models." (Kay, 1986, p. 60).

By the middle of the 19th century, Scotland was highly literate (in English) due to the reformation emphasis on reading the Bible, and such figures as John Knox, and the blossoming Scottish Enlightenment. Within this widely literate population, a contextual continuum is noted as developing between "dense" Scots and Scottish Standard English (McCLure, 1979). Speakers moved along this continuum largely depending on the context in which they found themselves: at home or socially with friends and family where Scots was largely used, or more professionally where standard English was appropriate. By the end of the 20th century, as Millar (2006) explains, "Scots had, to a very large extent, become a socially conditioned dialect, prone to exactly the same centripetal forces as other English dialects" (p. 64). At the same time, however, Scots had found a niche as a language of flourishing poetry. For example, Robert Burns, the national poet of Scotland, uses both Scots and English throughout his work. Modernist poetry, such as the works of Hugh MacDiarmid, often equated the use of Scots with nationhood (McCLure, 2000).

6.1.3 Scots's language status, prestige, and function today

Both historical and current socio-political factors influence the perception of Scots today, resulting in heavily divided opinions on the importance, relevance, and legitimacy of the language. The debate around Scots as a language or dialect is as much political, social, and ideological than it is linguistic. In terms of mutual intelligibility, English and Scots are as similar as Danish and Norwegian. Mutual intelligibility, however, is not a universal metric to determine language status. For example, Mandarin and Cantonese are largely unintelligible, though both are sometimes considered to be dialects of Chinese (Siegel, 2010). Indeed, the demarcation of a language or dialect is

often shrouded by politics, as popularised (originally expressed in Yiddish) by the linguist Max Weinreich, "a language is a dialect with an army and a navy".

The narrative of Scots as "bad English" or "slang" has been interpreted as a political tool of linguistic imperialism which turned into a pervasive narrative removing national identity and replacing it with social class demarcations (Matheson & Matheson, 2000). Indeed, Scots is predominantly considered to be used by the working class, particularly in urban areas where "it is generally seen as the language of vulgarity and lack of education and sophistication" (J. Costa, 2015; Johnston Jr, 2007; Matheson & Matheson, 2000, p. 216; Stuart-Smith, 2004). Within the UK context specifically, an individual's accent and manner of speaking is generally regarded as a direct proxy indicating socio-economic status and social class (Edwards, 1976). Further, often Scots speakers do not consider themselves to be speakers of Scots in an attempt to distance themselves from what is often perceived as a lower class dialect, reflecting the history of the language not being permitted in educational, administrative, or broadcasting contexts (Matheson & Matheson, 2000). In fact, many Scottish people report feelings of embarrassment around Scots, so-called "Scottish cringe" which can be defined as "a persistent sense of inferiority, self-pity – and embarrassment at overt expressions of Scottishness" (Brooks, 2020; Gardiner, 2016). Following from its expression of nationalism in modernist poetry in the early 20th century with key figures such as Hugh MacDiarmid, the modern political association with Scots is typically one of pro-Scottish independence. In fact, the Scottish National Party has promoted Scots as a fundamental element to national identity, including in their pre-election manifesto that the 2011 census should include a question about Scots (*Languages - Gov.Scot*, n.d.).

As an illustration of the complex status and prestige of Scots amongst its modern speakers, a 2010 Scottish government-commissioned survey reported that 85% of the sample spoke Scots (with 45% reporting to speak Scots a lot/fairly often), however, 65% of this same sample did not consider Scots to be a language in its own right, "more just a way of speaking" (TNS-BMRB, 2010, p. 2). Further, 67% of respondents reported that they likely used Scots without being aware of it (i.e., "I probably use Scots, but am not really aware of it"). In the 2011 census[2], wherein a question on Scots was included for the first time, 1.5 million people reported to be able to speak Scots and 1.9 million reported the ability to speak, read, write, or understand Scots (*Languages - Gov.Scot*, n.d.; Scots Language Center, 2011).

Perpetuating the narrative of Scots as "bad English", or more specifically simply phonetic-style translations of the Scottish English accent, is the recent Wikipedia scandal (Brooks & Hern, 2020). Forty-nine percent of all articles on Scots language on Wikipedia (at least 20,000 articles) were revealed to be written and edited by a non-Scots speaking teenager in the United States. Starting at the age of 12, this individual's fabricated articles containing English grammatical structures substituted with Scots sounding words was not discovered until 7 years later when a reddit user brought to light the poor quality of Scots Wikipedia. As described by the whistle-blower, "This is going to sound incredibly hyperbolic and hysterical, but I think this person has possibly done more damage to the Scots language than anyone else in history. They engaged in cultural vandalism on a hitherto unprecedented scale." (Brooks & Hern, 2020).

[2] The most recent Scottish census was scheduled for 2021, though was delayed by a year until 2022 and as such the results are not yet available.

When Scottish writer James Kelman's 1994 novel *How Late it Was How Late* won the Booker Prize, the opinion of Scots as not only the same as English, but further wrong and inferior was made clear on a public, international stage (Gardiner, 2016). One judge threatened to resign over the decision calling it a "disgrace" and a columnist for *The Times* referred to the work as "literary vandalism" and the author as "an illiterate savage" (Gardiner, 2016; Wood, 2014). In response to this criticism, Kelman accepted his award by stating, "my culture and my language have the right to exist" (Kelman, 1994).

Recently minority languages in Scotland are gaining increasing prominence due to both government language policies and language learning resources becoming increasingly available. In 2015 the Scottish government officially launched their Scots language policy, outlining steps to increase "support for Scots in education, encouragement of a greater profile for Scots in the media, and the establishment of a network of Scots co-ordinators" (Hyslop & Allan, 2015, p. 2). In 2020 Scottish Gaelic was added to the popular language learning app Duolingo with more than 127,000 registered users in the first two months (Brooks, 2020). Further, in 2019 The Open University in Scotland launched a successful free online Scots course in collaboration with Education Scotland, "teaching Scots through the context in which it's spoken, highlighting the role of Scots in Scottish culture and society, past and present" (BBC, 2019; The Open University, 2019).

6.1.4 Linguistic distance and cognition

From a cognitive perspective, the delineation of Scots as a language or dialect has been explored in a series of studies by Kirk and colleagues, largely with speakers of Scottish Standard English and Dundonian Scots (Declerck et al., 2021; Kirk et al., 2014, 2018, 2022). Much of this work has supported Scots

to be represented cognitively as separate from English as speakers of the two languages show switching costs typical of (unbalanced) bilinguals (Kirk et al., 2018, 2022).

More generally, the question of if the linguistic distance between two languages influences how they are processed cognitively have been explored in the literature among users of Standard and Swiss German (Vorwerg et al., 2019), Standard Modern and Cypriot Greek (K. Antoniou et al., 2016) and, in the case of different written standards, Nynorsk and Bokmål Norwegian (Lundquist & Vangsnes, 2018). For example, Vorwerg and colleagues (2019) found that Swiss-German speakers had longer naming latencies when switching from Standard German to Swiss German. Such a phenomenon (i.e., switch cost asymmetry) is well documented in bilingual switching. Further Lundquist and Vangsnes (2018) found that "stable" dialect speakers (meaning the dialect was spoken at home as well as socially) were able to not only accurately predict impending linguistic material in an eye tracking task, but also edited these predictions based on the dialect mode (e.g., Oslo dialect or Sogn dialect of Western Norway). The authors argued that such a pattern of behaviour suggests that this group should be treated as bilingual. Additionally, Antoniou and colleagues (2016) found evidence in support of a multilingual and bidialectal cognitive control advantage among children (M_{age} = 7.8) moderated by proficiency.

However, just as the so-called "bilingual advantage" literature includes studies with a variety of nuanced findings, not always in support of a clear-cut cognitive advantage due to speaking two languages, so too is the bidialectal literature. For example, using the Simon task with bilingual, bidialectal, and monolingual older adults (M_{age} = 70.8), Kirk et al. (2014) failed to replicate

Bialystok and colleagues' (2004) influential finding of improved inhibitory control in older adult bilinguals.

As such, regardless of the categorisation of Scots as a language in its own right or a variety of English, the literature suggests that diglossic and bilingual language control functions via similar mechanisms, involving inhibitory control processes. Therefore, the question arises of if there is a cognitive effect of short, intensive language training in Scots.

6.1.5 Present study

The present study set out to investigate how the status and prestige of a language, using Scots, influences learning and student opinions of the target language. Specifically, two research questions were investigated:

- How do student opinion about the language and its status and prestige change after taking a course on Scots?
- In the context of the preceding chapters, is there a measurable cognitive effect after the language learning classes?

6.2 Methods

Free, two-hour Scots classes took place online via Zoom three times a week for three weeks for a total of 18 hours of language learning from June to September 2020. Dr Michael Dempster, a native Scots speaker as well as Scots Scriever from 2018 to 2022 (an exclusive residency at the National Library of Scotland and Creative Scotland to produce original creative work in Scots) and Director of the Scots Language Centre, developed, designed, and delivered the Scots courses. Dr Dempster acted not only as the teacher, but a collaborating member of the research team on this project.

6.2.1 Participants

As noted in the introduction, the complicated socio-political context of Scots means that many Scots speakers have a nuanced relationship with it, and in fact may not identify themselves as speakers of Scots. For this reason, participants were allocated to three different classes/groups based on the linguistic environment (LE) in which they grew up: individuals who grew up exposed to the Scots phonological environment, in an English phonological environment, or in a non-native English phonological environment, referred to as the "Scots LE", "English LE", and "Non-native English LE" groups, respectively. Asking participants outright if they are speakers of Scots was not deemed sufficient. Instead, based on their answers to the following screening questions participants were allocated to one of the three classes: (a) What is your mother tongue(s)? If English, which variant of English? (b) Which other languages do you know/speak? (c) Have you grown up or been surrounded by the Scottish or Northern Irish variant of English[3]? and (d) Do you have a history of dyslexia? Two rounds of classes, each including 3 groups, were conducted (Table 6.1).

As the language courses offered were bespoke to the present study, the timeline of data collection was largely determined by the availability of funding. Therefore, as many classes were run as possible within the range of time that both funding and the tutor was available. Stopping criteria was driven by the most part by this logistic factor. A total of 69 participants were recruited from 20 countries (Figure 6.1). Notably, while indeed many native English

[3] Northern Irish variant of English was grouped with Scottish English due to Ulster Scots, a dialect of Scots spoken in Ulster in Northern Ireland. However, no participants in this study identified as having grown up exposed to Northern Irish English or Ulster Scots.

speakers were from England, the participant group consisted of participants from numerous other English-speaking countries such as Canada, the United States, and Australia, as seen in Figure 6.1. Participants were recruited via social media channels including Twitter and Facebook. Inclusion criteria was being 18 years or older. Across all six classes, a total of 62 completed the course (M_{age} = 40.89, *SD* = 14.18) and 58 completed the research sessions (90% retention rate for the course and 84% for the study; see Table 6.1). The demographics of the students who completed the course and accompanying research sessions are shown in Table 6.2 (students who did not complete the course are discussed in Section 6.3.1). Overall, the participant pool in this study was quite multilingual with 23 different native languages and 48 unique additional languages reported. Indeed, particularly in the second round of classes (largely joining the English group) many members of the polyglot community joined the course after the advertisement was shared by a member in the "Polyglots" Facebook group.

Table 6.1 Participant numbers by LE group

	Scots LE		**English LE**		**Non-native English LE**	
	Started	Completed	Started	Completed	Started	Completed
Round 1	8	8	13	9	10	10
Round 2	9	9[1]	17	13	12	10
Total Course	17	17	30	24[2]	22	21[1]
Total Research	17	16	30	22	22	20

[1] one student completed the course but not the research session

[2] two students completed the course but not the research session

Figure 6.1 World map of participants

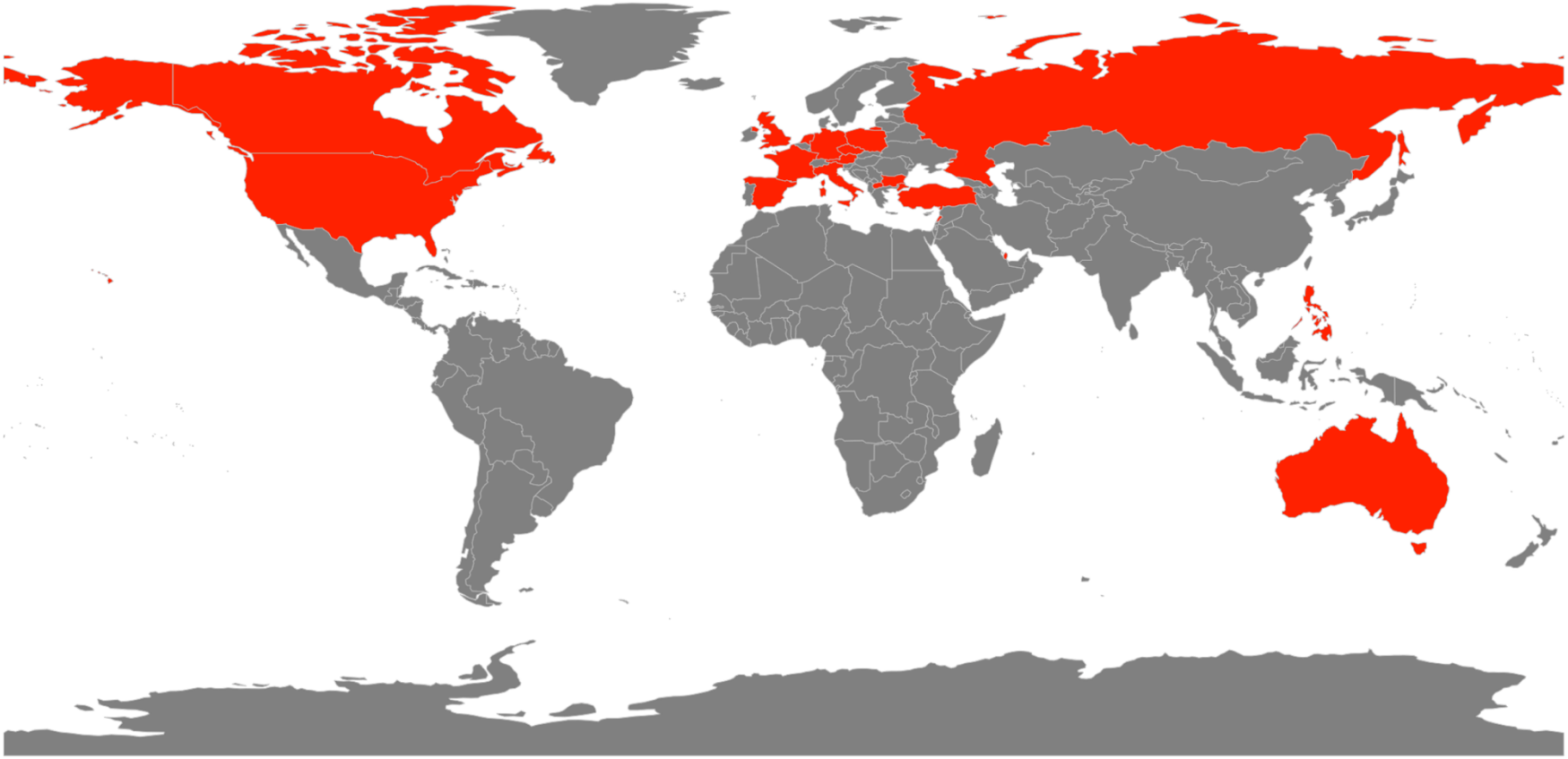


Table 6.2 Demographics of students who completed the course and research sessions

Characteristic	**Overall**, N = 61[1]	**Scots LE**, N = 16[1]	**English LE**, N = 22[1]	**Non-native English LE**, N = 20[1]
Round				
Round 1	27 (47%)	8 (50%)	9 (41%)	10 (50%)
Round 2	31 (53%)	8 (50%)	13 (59%)	10 (50%)
Participant Age	40.89 (21, 74)	47.31 (21, 74)	39.72 (21, 71)	37.05 (22, 68)
Gender				
Female	30 (52%)	9 (56%)	8 (36%)	13 (65%)
Male	24 (41%)	5 (31%)	13 (59%)	6 (30%)
Prefer to self-describe	1 (1.7%)	—	—	1 (5.0%)
Prefer not to answer	3 (5.2%)	2 (12%)	1 (4.5%)	—
Handedness				
Right	47 (81%)	13 (81%)	15 (68%)	19 (95%)
Left	8 (14%)	2 (12%)	5 (23%)	1 (5.0%)
Both	3 (5.2%)	1 (6.2%)	2 (9.1%)	0 (0%)
Education Level				
Primary school	—	—	—	—
Secondary/High school	—	—	—	—
Vocational degree/ Professional training	4 (6.9%)	2 (12%)	1 (4.5%)	1 (5.0%)
Undergraduate Student	2 (3.4%)	—	1 (4.5%)	1 (5.0%)
University Degree	25 (43%)	8 (50%)	9 (41%)	8 (40%)

Characteristic	**Overall**, N = 61[1]	**Scots LE**, N = 16[1]	**English LE**, N = 22[1]	**Non-native English LE**, N = 20[1]
Postgraduate Student	7 (12%)	1 (6.2%)	3 (14%)	3 (15%)
Master's Degree	17 (29%)	5 (31%)	7 (32%)	5 (25%)
Doctorate	3 (5.2%)	—	1 (4.5%)	2 (10%)
No. Languages Before Class	7.7 (2, 50)	4.7 (2, 10)	10 (2, 50)	7.6 (2, 35)

[1] n (%); Mean (Range)

6.2.2 Materials and procedure

The study is mixed methods, adopting an explanatory sequential design (see Section 2.5.1.1). Following a repeated measures (pre- and post) design, participants attended research sessions consisting of questionnaires, an auditory cognitive test, and an auditory task before the start of the first class and after the last class. Questionnaires were completed online and independently by each participant, without any influence from the researcher or teacher (i.e., neither were present). The auditory cognitive test was the Test of Everyday Attention (Robertson et al., 1994), described in detail in Section 2.6.5.1 which is the primary outcome variable of the cognitive research question present study. While not within the scope of the present study, participants also completed an online auditory oddball task exploring vowel length discrimination, which is a question of interest in Dr Dempster's research around Scots. In the pre-course testing session participants also completed Raven's Advanced Progressive Matrices, described in detail in Section 2.6.5.2 (Raven et al., 1998), as well as a series of VAS style questions about the online learning experience in the post-course testing session. Additionally,

participants completed the Language History Questionnaire (see Section 2.6.4.1), a free response question about their motivation for learning Scots, as well as a series of language attitude questions (see Appendix M). The language attitude questions, as the primary outcome of interest, were adapted from the 2010 government-commissioned survey into public attitudes towards the Scots language (TNS-BMRB, 2010). The majority of questions were asked on a horizontal fully labelled ascending 6-point Likert-style scale, though 5 statements were presented as a VAS-style question:

- Scots is a language in its own right.
- Scots is difficult to learn.
- It's really good for your brain to use/speak Scots
- Scots should be spoken but not written.
- Scots is appropriate to use in official documents/ communications.

As these 5 questions in particular are relevant to the primary research question, they were presented as VAS statements to allow for a more fine-grained analysis of change in opinions before and after the course (see Section 2.6.2.2 for further discussion of VAS). After completing the course, participants were invited to attend a feedback session. These feedback sessions were post-course focus groups to gather the students' opinions and experiences from the course, including questions around their attitudes towards Scots, the learning experience, intensity, frequency, and duration of the course (see Appendix N). The focus groups were conceptualized as lasting for around 30 minutes, though in practice ranged from 35 minutes to 54 minutes. Other qualitative data collected include discussions during the video calls, emails, and free response questions in the pre- and post-course questionnaires.

Ethics was approved by the University of Edinburgh PPLS Research Ethics Committee (ref: 440-1819/8).

6.2.3 Analysis

In order to answer the two research questions of the present study, a mixed method analytic approach was adopted. As the outcome measures of the primary research question are both quantitative, in the form of Likert-style and VAS language attitude questions, and qualitative, from post-course focus groups and free response questions. The quantitative data was analysed using descriptive statistics, and where the descriptive statistics suggested potential differences between groups or testing sessions, using inferential statistics in the form of *t*-tests or ANOVAs. Where parametric test assumptions are violated, non-parametric alternatives were used (e.g., Wilcoxon signed-rank test in lieu of a paired-sample *t*-test, Kruskal-Wallis test in lieu of a one-way ANOVA, etc.). As described in Section 3.2.4, Holm-Bonferroni corrections were applied as a correction for multiple testing to control the FWER.

For the qualitative data, the free-response questionnaire data and the feedback session data was transcribed orthographically/verbatim and analysed using reflexive thematic analysis (Braun & Clarke, 2006, 2021a) (see Section 2.3.1.5.3). Qualitative data was analysed and coded using NVivo 12 Pro (released January 2022) (*NVivo*, 2022). Prior to the initial coding stages, all participant names were anonymised (see Section 2.6.1 for further details on the anonymisation procedure). As the primary researcher throughout this project, I ensured prolonged engagement with the data and individual participants prior to analysis. I noted my initial impressions on potential analytic interests after each feedback session as well as after transcription. A critical realist epistemological stance was taken in this study, recognising the human influence on reality and truth which is situated and interpreted. As such, this epistemological stance allows for the analytic process to focus on the lived experience of the participants while considering the social and cultural

influences on these experiences (Braun & Clarke, 2021a). Therefore, both a deductive (based on the interview schedule) and inductive approach to coding was adopted, using both semantic (explicit or overt) and latent (implicit or underlying) coding frameworks. While both coding approaches were planned, in practice, far more inductive coding was employed overall after initial data familiarisation stages and after deductive coding proved to provide too shallow of an analysis. RTA allows for both inductive and deductive theme generation, which particularly suited the qualitative arm of the current research. This flexibility allowed for the analytic process to be shaped not only by pre-existing questions but also to be largely informed by the data and experiences of the participants themselves. Quotes are reported in the results section include both the participant's age and anonymised name.

The secondary research question is quantitative, with the cognitive scores as the primary outcome variable(s). The same analytic approach was taken as described in Chapter 4 and Chapter 5 (see Sections 4.2.4 and 5.2.4), including descriptive statistics followed by inferential statistics where appropriate with Holm-Bonferroni correct *p*-values being reported. As the data is clustered by participants within classes, linear mixed effects modelling was selected as the method of data analysis (see Section 2.4.4). Models were fitted with participants nested within class, however, failed to converge with this random effects structure. As such, nesting within class was removed and the random intercept by participant grouping structure remained for all models reported. Consistent with previous research using the TEA, a trial random effect was not included as the models were intended to measure overall score on each TEA task rather than the change in scores (Long et al., 2020). In terms of fixed effects, categorical covariates/predictors were included using treatment coding while numeric covariates/predictors were mean-centred for

interpretability (age of participants *M* = 40.89, years of education *M* = 19.02, average hours per week practicing Scots *M* = 15.96, composite language knowledge score before the course *M* = 70.15, and Raven's APM score *M* = 76.13). An interaction (fixed effects) between testing session (i.e., pre-course or post-course) and group (i.e., Scots LE, English LE, or Non-native English LE) well as testing session and age of the participant was investigated for each model within the model building strategy. Overall, the approach to model building adopted follows Barr et al. (2013)'s suggestion of beginning with the maximal random effects structure justified for the study design, or "keeping it maximal". All models were fit with Restricted Maximum Likelihood estimation and the bobyqa optimiser. As described in Section 4.2.4, inference criteria of these models use included case-based bootstrapping with 5,000 resamples to compute 95% confidence intervals around the model parameter estimates.

Statistical analyses and data visualisations were conducted using R Version 4.1.3 (R Core Team, 2022), the lme4 (Bates et al., 2015), lmerresampler (Loy et al., 2022), rstatix (Kassambara, 2021), effectsize (Ben-Shachar et al., 2020), ggplot2 (Wickham, 2016), and patchwork (Pedersen, 2020) packages.

6.3 Results

6.3.1 Dropouts

The retention rate across the course was high (90% completed the course) and slightly lower for the study itself (84% completed the research sessions). Of the 11 participants without post-course data, 4 attended the classes though did not complete the post-course research session. Therefore, in terms of dropout analysis, 7 participants withdrew from the course (Table 6.3). One participant withdrew from the course and study requesting their data

being deleted, as such they are included only in number, explaining the discrepancy between 6 and 7 in Table 6.3 and Table 6.4. Of those who withdrew from the course, 71% of those who withdrew cited scheduling difficulties or unexpected changes in their work schedule. As seen in Table 6.4, the baseline TEA scores as well as the majority of demographic variables are similar between the participants who completed the course and those who dropped out. Interestingly, no student in the Scots LE group withdrew from the study, although one did fail to complete the second research session.

Table 6.3 Demographics of students who dropped out

Characteristic	**Overall**, N = 6[1,2]
Round	
Round 1	1 (17%)
Round 2	5 (83%)
Group	
Scots LE	—
English LE	5 (83%)
Non-native English LE	1 (17%)
Participant Age	42.67 (32, 57)
Gender	
Female	5 (83%)
Male	1 (17%)
Handedness	
Right	5 (83%)
Left	—
Both	1 (17%

Education Level	
Primary school	—
Secondary/High school	—
Vocational degree/ Professional training	—
Undergraduate Student	—
University Degree	1 (20%)
Postgraduate Student	—
Master's Degree	4 (80%)
Doctorate	—
(Missing)	1
No Languages before Course	10.67 (4, 21)

[1] n (%); Mean (Range

[2] Total dropouts n = 7, however n = 6 reported due to data withdrawal

Note: due to only 1 participant withdrawing from the non-native English group, demographics are not shown by group but rather overall to maintain anonymity and minimize potential data linkage/identifiability

Table 6.4 Demographic and baseline scores of participants who completed the course versus dropped out

Characteristic	**Completed Course**, N = 58[1]	**Dropped Out**, N = 6[1, 2]
Participant Age	41 (21, 74)	43 (32, 57)
Gender		
Female	30 (52%)	5 (83%)
Male	24 (41%)	1 (17%)
Prefer to self-describe	1 (1.7%)	—
Prefer not to answer	3 (5.2%)	—
Handedness		
Right	47 (81%)	5 (83%)
Left	8 (14%)	—
Both	3 (5.2%)	1 (17%)
Number of Languages	7.7 (2, 50)	10.7 (4, 21)
Education Level		
Primary school	—	—
Secondary/High school	—	—
Vocational degree/ Professional training	4 (6.9%)	—
University Degree	25 (43%)	1 (20%)
Postgraduate Student	7 (12%)	—
Master's Degree	17 (29%)	4 (80%)
Doctorate	3 (5.2%)	—
Ravens's APM	76 (11, 100)	73 (35, 95)
Pre-course Elevator Task	97.78 (85.7, 100)	100 (100, 100)

Pre-course Elevator Task with Distraction	87.07 (10, 100)	88.33 (40, 100)
Pre-course Elevator Task with Reversal	64.48 (0, 100)	73.33 (20, 100)

[1] Mean (Range); n (%)

[2] Total dropouts n = 7, however n = 6 reported due to data withdrawal

6.3.2 Language attitudes

After the course in the post-course questionnaire, students were asked to describe their experience in the course in 1 to 2 words, shown in Figure 6.2. Reflecting the unique linguistic and cultural status of the language, the most common descriptors are “informative” and “interesting” followed by “fun” and “engaging”, which highlight the positive overall learning experience. As mentioned in the methods, participants responded to 15 language attitude questions before and after the Scots course, which were presented as Likert-style scales, multiple choice, or VAS questions (Appendix M). Results will be presented according to question response type below followed by the qualitative analysis.

Figure 6.2 Course described in 1 word

6.3.2.1 Likert style questions

Participants responded to eight statements via a 6-point Likert-style scale:

1. It is important for members of my community to know Scots.
2. Schoolchildren here should have the opportunity to study Scots.
3. I would encourage others like me to speak Scots.
4. Learning Scots demands less intelligence than other languages I know.
5. There is no good reason for anything to use Scots in the 21st century.
6. Scots has a community with its own culture.
7. Scots is important for the cultural heritage of Scotland.
8. Scots is important for my own cultural heritage.

Responses overall can be seen in Appendix O (Figure O1), with group level responses shown in Figure 6.3 (Scots LE group), Figure 6.4 (English LE group), and Figure 6.5 (Non-native English LE group).

Figure 6.3 Scots LE group responses to Likert-style questions before and after the course

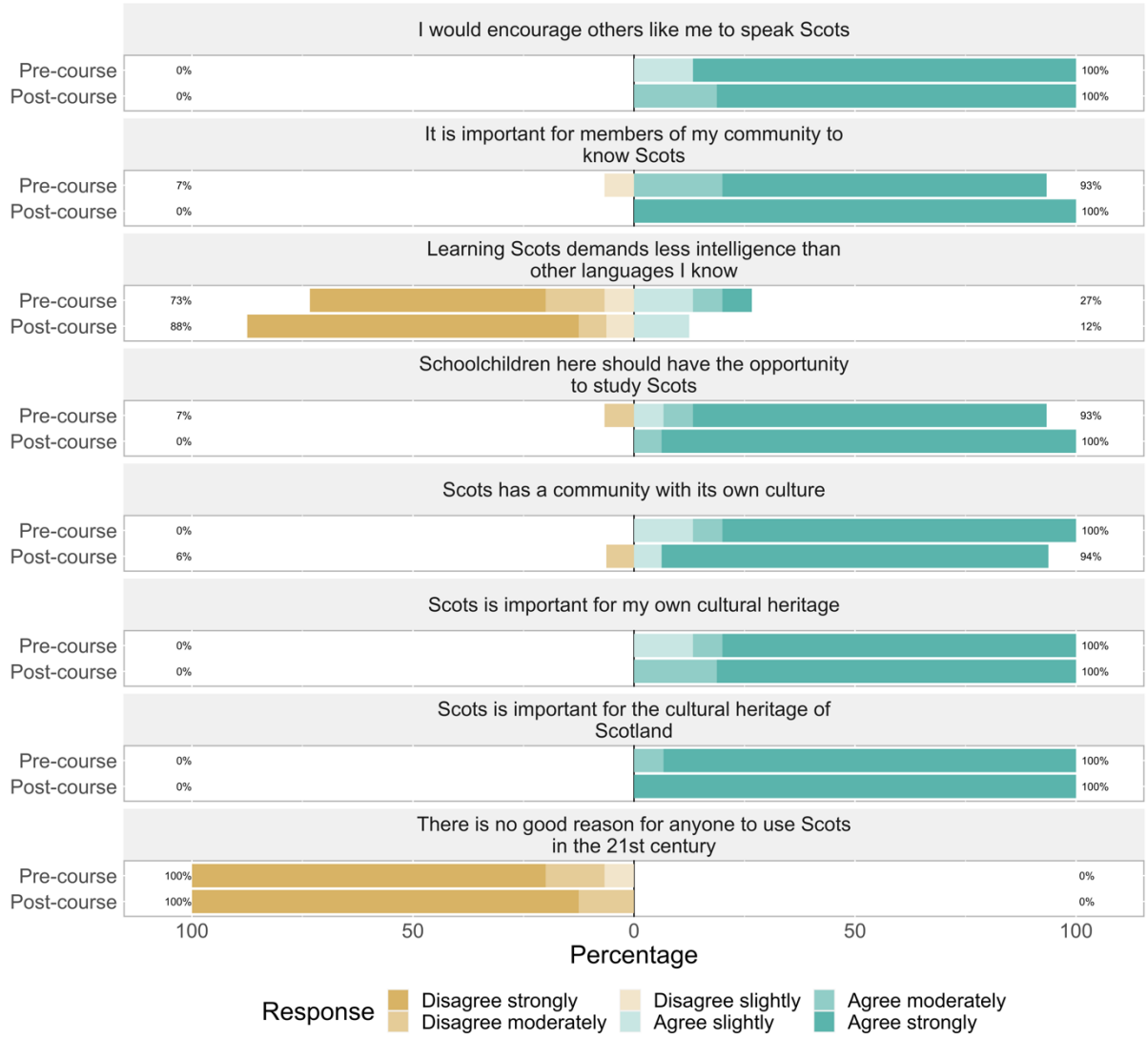


Figure 6.4 English LE group responses to Likert-style questions before and after the course

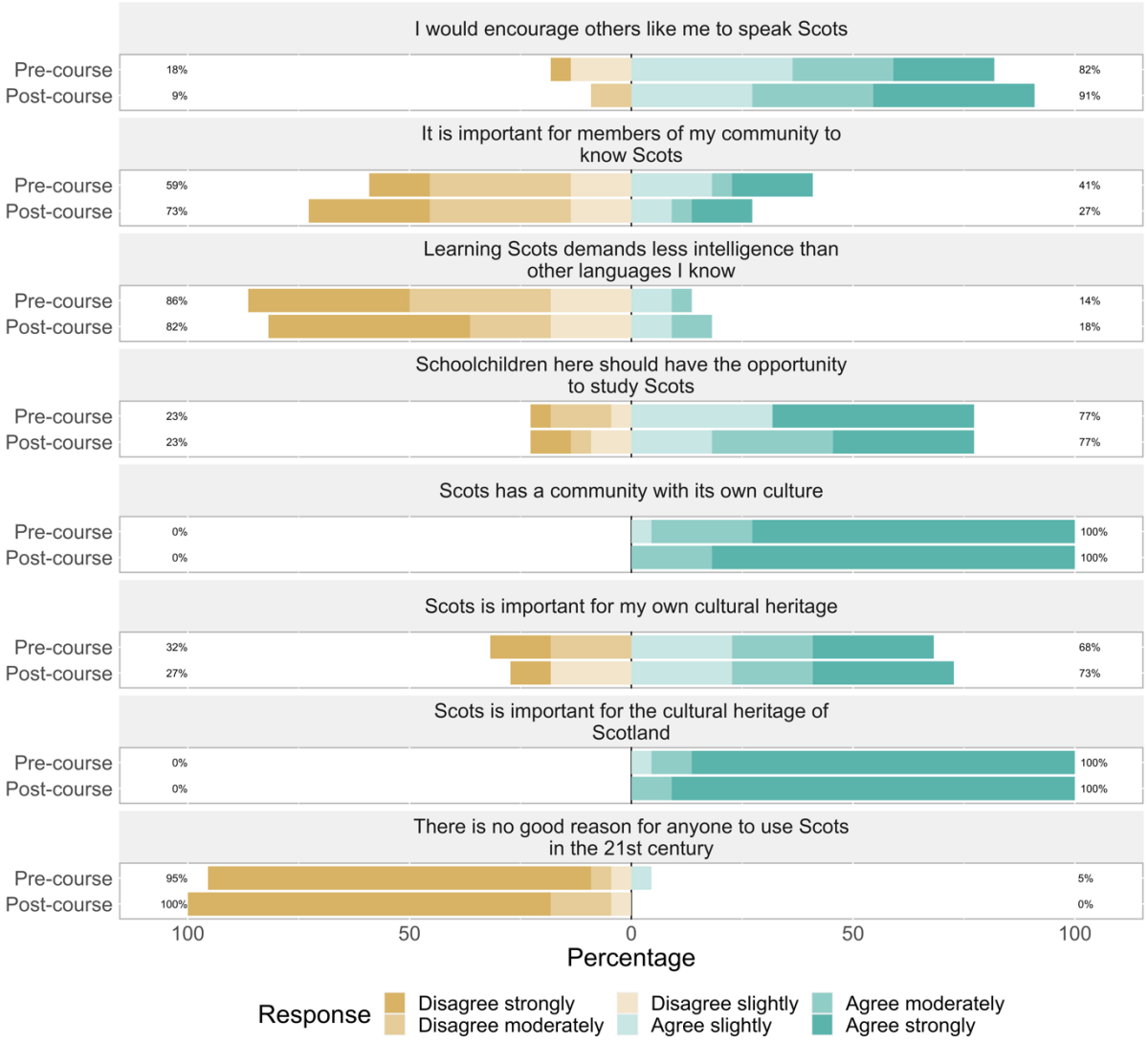
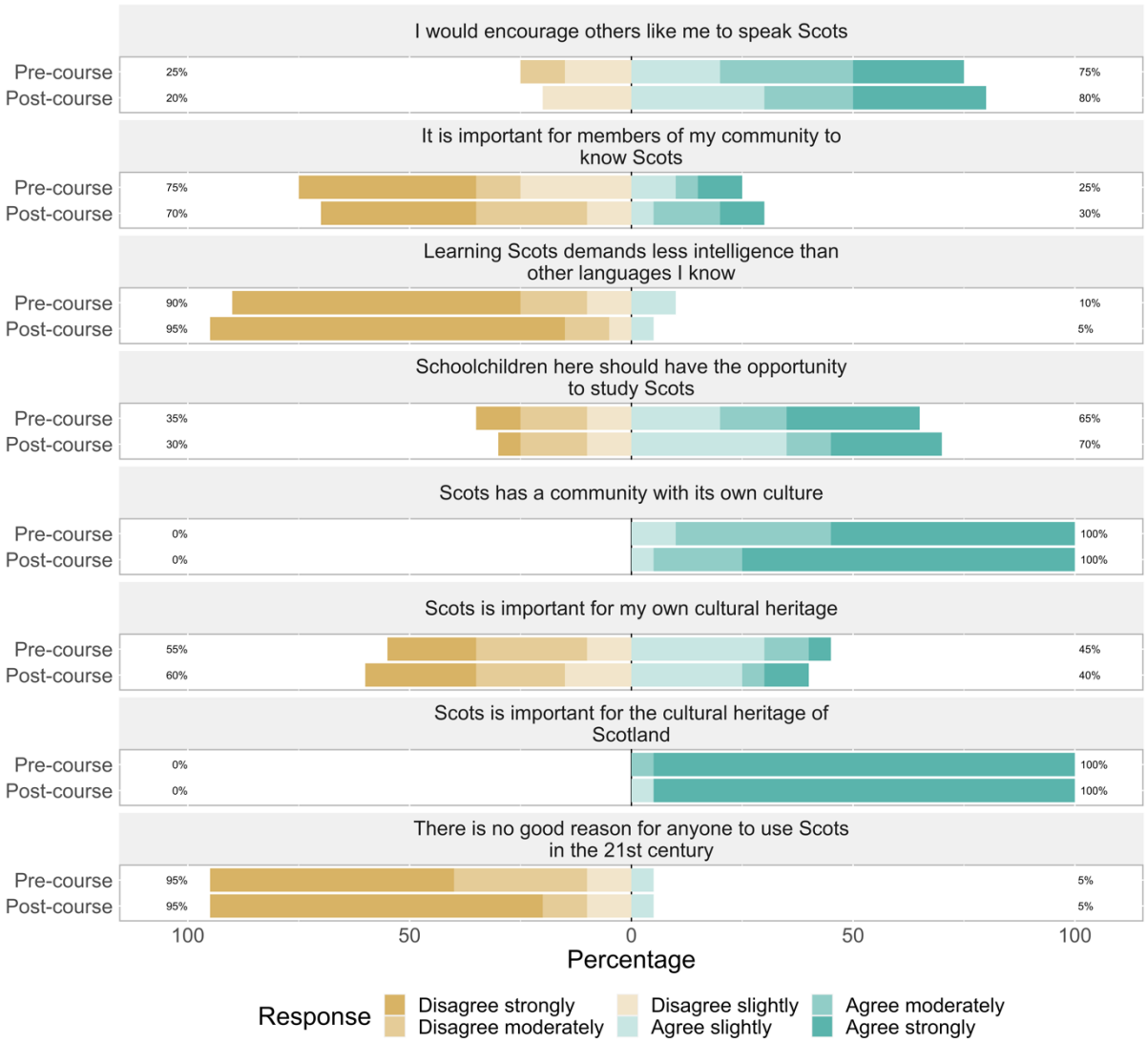


Figure 6.5 Non-native English LE group responses to Likert-style questions before and after the course



As expected, given the different levels of identification with the language, the Scots LE group and non-native Scots speaking groups (i.e., English LE and Non-native English LE groups) show a different pattern of results after the course around questions of identity. *"I would encourage others like me to learn Scots"* was at ceiling with 100% agreement in the Scots LE group before the course, remaining so after the course though with stronger

responses (i.e., higher percentage of "agree moderately" and "agree strongly"). For the English and non-native English LE groups after the course there was a trend of further agreement (English LE group agreement increased 9%, non-native English LE group agreement increase 5%), though in both groups some level of disagreement to the statement remained. A higher percentage of the non-native English LE group (20%) disagreed after the course, while only 9% of the English LE group disagreed.

Similarly, while the Scots LE group increased 7% in agreement after the course and universally agreed strongly to the statement, *"It is important for members of my community to know Scots"*, the English LE and non-native English LE groups had comparatively lower levels of agreement after the course (27% and 30% respectively). While the non-native English LE group increased 5% in their level of agreement, the English LE group decreased in agreement 14% after the course.

In terms of the community of Scots speakers, both the English LE and non-native English LE groups had 100% agreement before and after the course. Interestingly, the Scots LE group shifted from 100% agreement before the course to 94% agreement after the course, though of those in agreement more strongly agreed with the statement.

Relating to questions of the role Scots plays in personal and Scottish cultural heritage, the Scots LE group before the course reported 100% agreement to both statements, which remained after the course. Scots was universally seen as important for the cultural heritage of Scotland with 100% agreement across all LE groups at both testing sessions. Interestingly, Scots LE group students responded to the statement *"Scots is important for the cultural heritage of Scotland"* after the course with 100% strong agreement, and a mix of "agree moderately" and "agree strongly" after the course to Scots being

important to their personal cultural heritage. Both the English LE and non-native English LE groups showed 100% agreement to Scots being important to the cultural heritage of Scotland before the course, which remained at the same levels after the course. As neither group was comprised of Scottish participants, the level of agreement to Scots being important to their personal cultural heritage was surprising. The English LE group increased their level of agreement 5% after the course for a total of 73% of participants agreeing to Scots being important to their personal cultural heritage. The non-native English LE group showed a slightly different pattern with a 5% decrease in agreement in the second testing session, though still by the end of the course 40% of participants agreed to the statement.

Relating to opinion towards Scots as a language, the statement *"Learning Scots demands less intelligence than other languages I know"* decreased 15% in levels of agreement in the Scots LE group after the course. However, 12% of participants still agreed slightly at the second testing session. The English LE group had high levels of disagreement before the course (86%) and after the course (82%), though increased 4% in levels of agreement. The non-native English LE group decreased 5% in levels of agreement after the course (resulting in 95% disagreement), though 5% of participants still agreed slightly to the statement.

The final two Likert-style questions concern the perceived utility of Scots: "There is no good reason for anyone to use Scots in the 21st century" and "Schoolchildren here should have the opportunity to study Scots". Relating to its use in the 21st century, both the English LE and Scots LE groups disagreed 100% after the course. The Scots LE group had the same level of agreement as before the course, though the English LE group decreased their level of agreement by 5%. The non-native English LE group remained stable in their

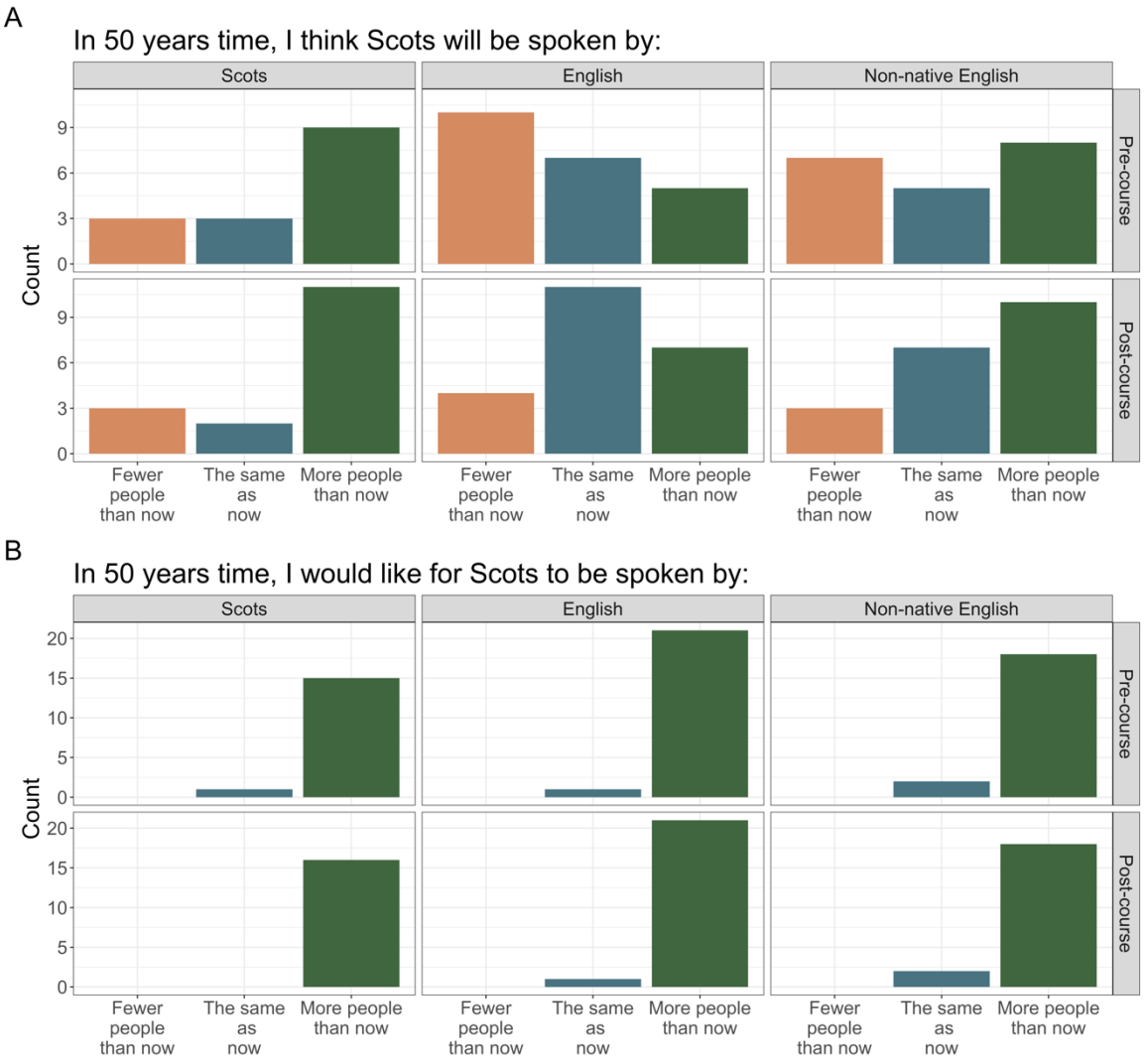
opinions overall with 95% disagreement, though within that 95% had higher rates of strong disagreement after the course. In terms of the final statement, 'here' was intentionally included to capture participant opinions how accessible Scots should be in schools in their local context. Interestingly, the non-native English LE group, increased 5% in agreement after the course, resulting in 70% agreement though the majority of this was slight agreement. The English LE group opinions remained stable before and after the course with 77% agreement. Finally, the Scots LE group increased 7% in agreement after the course to 100% agreement, with the majority strongly agreeing that schoolchildren in Scotland should have the opportunity to learn Scots.

Relating to the age profile of participants, age was not significantly correlated with any of the Likert-scale attitude questions.

6.3.2.2 Multiple choice questions

The two multiple choice questions relate to opinions around the future of Scots, both what the participants think is likely to happen in terms of number of speakers and what they personally would like to be the case. The dissociation between what participants *thought* would be the case in 50 years (Figure 6.6A) and what they *would like* to be the case (Figure 6.6B) is quite striking before the course, and even more extreme after the course. While the English LE and non-native English LE groups shifted post-course to most positive responses (i.e., "the same as now" and "more people than now") in response to what they think will be the case, the Scots LE group was more stable with only 2 participants altering their response post course (i.e., from "the same as now" to "more people than now").

Figure 6.6 Language attitude multiple choice questions by LE group and session



6.3.2.3 VAS questions

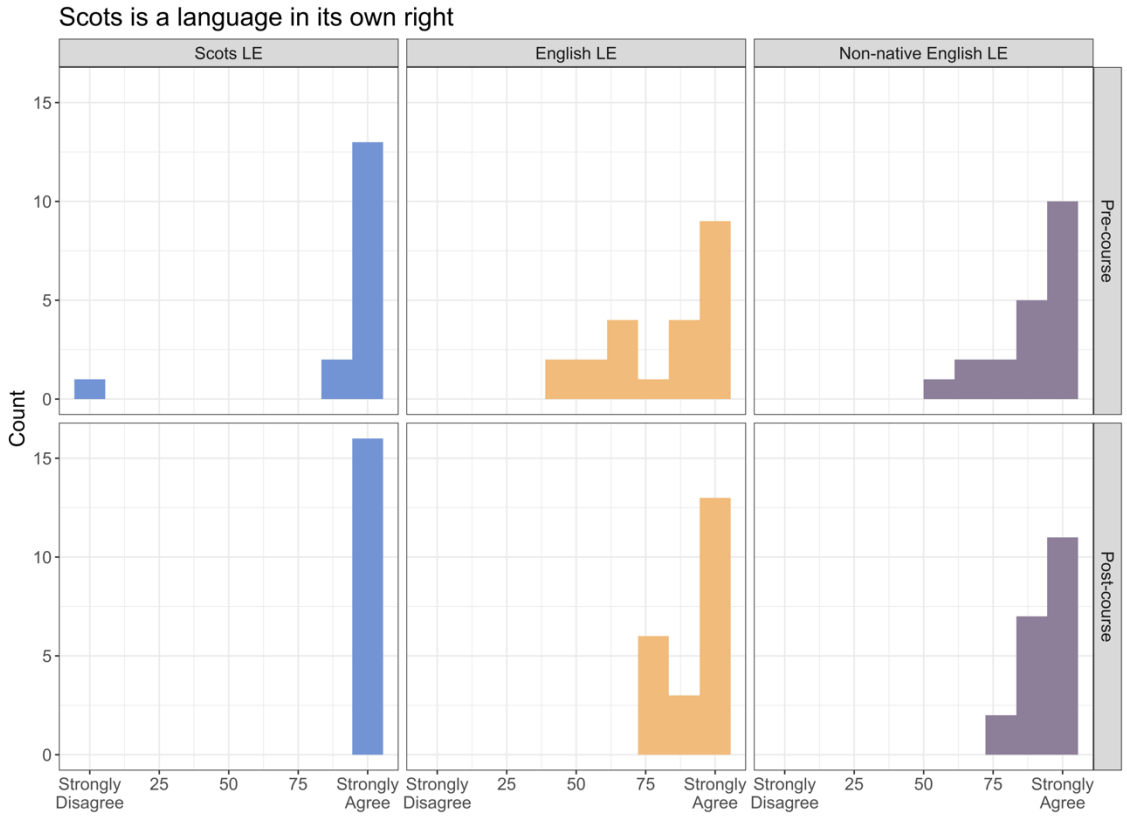
Participants responded to 5 VAS statements before and after the course:

1. Scots is a language in its own right. (Figure 6.7)
2. Scots is difficult to learn. (Figure 6.8)
3. It's really good for your brain to use/speak Scots. (Figure 6.9)
4. Scots should be spoken but not written. (Figure 6.10)

5. Scots is appropriate to use in official documents/communications. (Figure 6.11)

The change in opinions scores before and after the course are visualised by group in Appendix O (Figure O2).

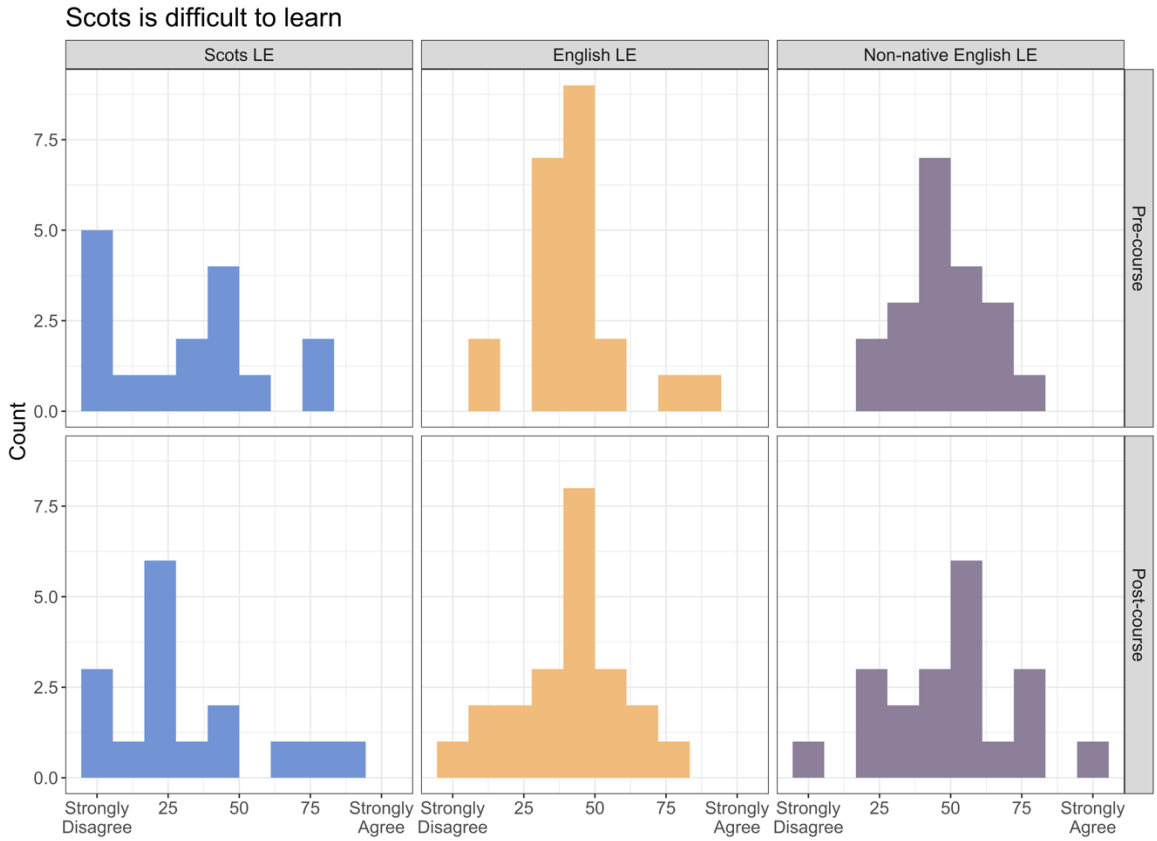
Figure 6.7 Language attitude VAS statement 1 by LE group and session



As seen in Figure 6.7, across all LE groups there appears to be stronger agreement to Scots being a language in its own right after the course. In the Scots LE group after the course there was 100% strong agreement with the statement, where all participants rated 100 out of 100. Notably, the Scots LE group is the only group to have a response below a rating of 50 before the course, with one extreme outlier strongly disagreeing (i.e., rating of 0).

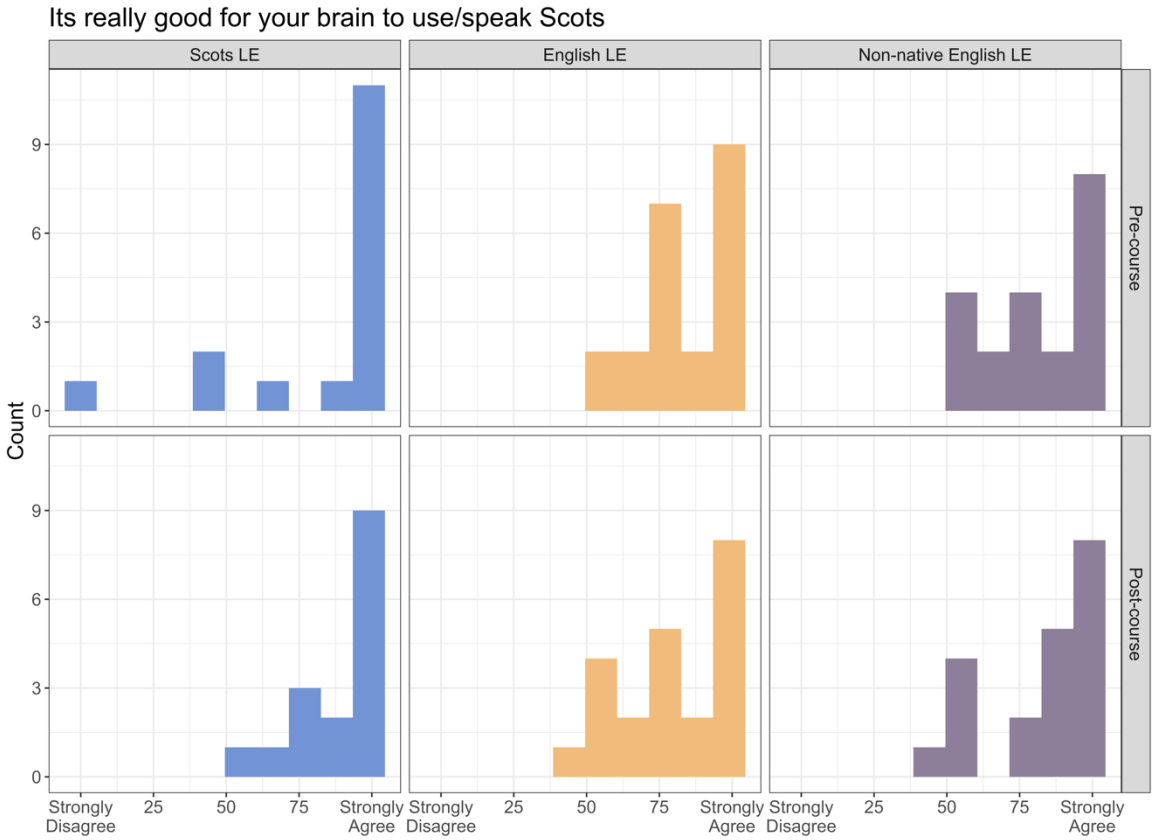
This change in opinions was further investigated inferentially. Overall, across all LE groups, opinions towards this statement significantly increased in levels of agreement before (M = 87.88, SD = 19.02, Mdn = 99) and after the course (M = 95.19, SD = 7.69, Mdn = 100), W = 57, p < .001, with a large effect size (r = .51). Subgroup analysis revealed that the English LE group in particular experienced the highest level of change in their opinion after the course ($M_{Post-course}$ = 92.32, $SD_{Post-course}$ = 9.58, $Mdn_{Post-course}$ = 99.5) compared to before ($M_{Pre-course}$ = 82.73, $SD_{Pre-course}$ = 18.6, $Mdn_{Pre-course}$ = 90) and in fact is largely driving the overall trend. Both the Scots LE group and non-native English LE groups did have higher levels of agreement with the language status of Scots after the course ($M_{Scots\ LE}$ = 100, $SD_{Scots\ LE}$ = 0, $Mdn_{Scots\ LE}$ = 100; $M_{Non-native\ LE}$ = 94.5, $SD_{Non-native\ LE}$ = 6.83, $Mdn_{Non-native\ LE}$ = 100) compared to before the course as well ($M_{Scots\ LE}$ = 91.81, $SD_{Scots\ LE}$ = 24.79, $Mdn_{Scots\ LE}$ = 100; $M_{Non-native\ LE}$ = 90.4, $SD_{Non-native\ LE}$ = 12.93, $Mdn_{Non-native\ LE}$ = 96). However, the English LE group change in opinion scores is the only subgroup analysis with a statistically significant difference before and after the course, W = 90, p = .002 with a large effect size (r = .67).

Figure 6.8 Language attitude VAS statement 2 by LE group and session



Opinions toward the difficult of Scots as a language to learn overall remained quite stable before (*M* = 42.52, *SD* = 21.42, *Mdn* = 48.5) and after the course (*M* = 41.74, *SD* = 24.39, *Mdn* = 40.5), with a small trend towards disagreement in the second testing session (Figure 6.8). The Scots LE group trended towards more disagreement around the difficulty of Scots after the course (*M* = 30.44, *SD* = 27.39, *Mdn* = 20) compared to before (*M* = 33.44, *SD* = 27.98, *Mdn* = 33). The English LE group showed the same pattern of responses before (*M* = 43.18, *SD* = 18.64, *Mdn* = 40) and after the course (*M* = 42.09, *SD* = 18.78, *Mdn* = 42.5). The non-native English LE group was the only one to slightly agree more overall to the difficulty of Scots after the course (*M* = 50.4, *SD* = 24.8, *Mdn* = 51.5) compared to before (*M* = 49.05, *SD* = 16.12, *Mdn* = 50).

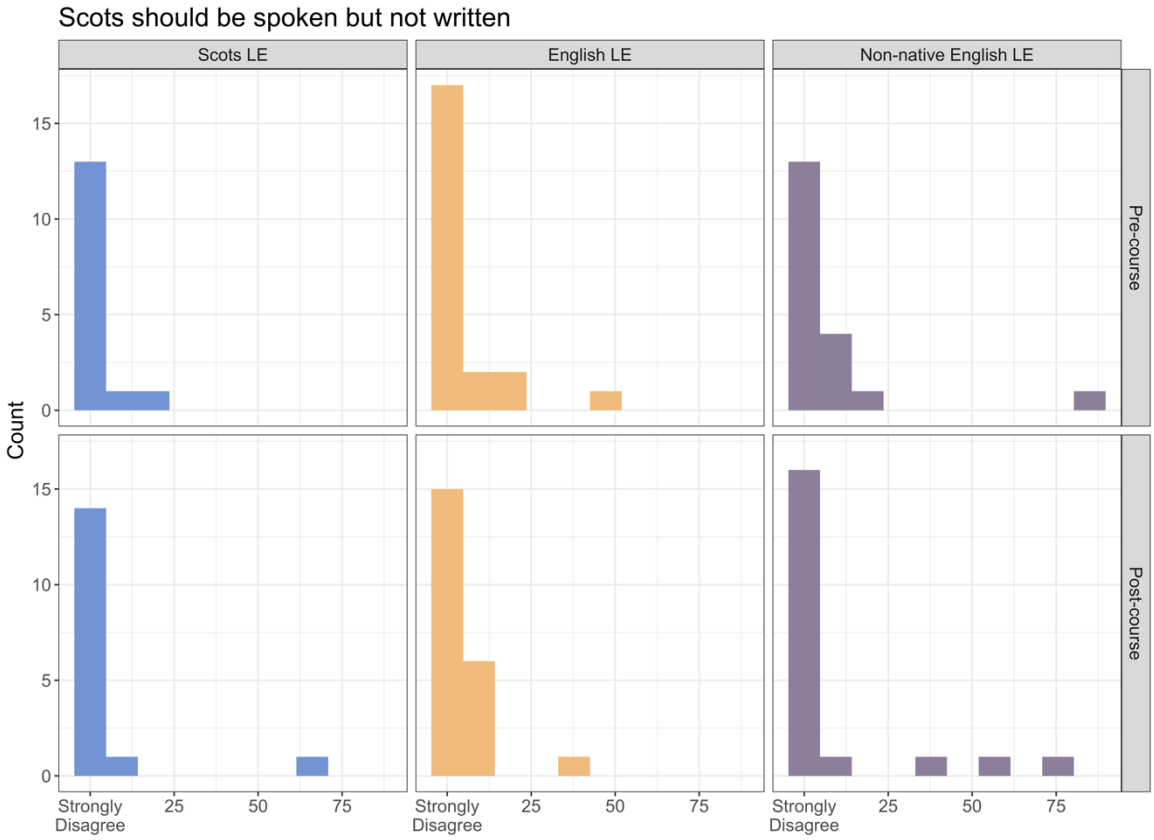
Figure 6.9 Language attitude VAS statement 3 by LE group and session



Previous studies looking at adults across the lifespan learning majority languages (e.g., German, French, Spanish, Danish, Japanese, Chinese, etc.), not reported as part of this thesis, have shown that participant typically strongly agreement with the statement *"It's really good for your brain to use/speak [target language]"*. In the case of Scots, this statement was posed to further parse apart student opinions towards the status of the language. Overall, as seen in Figure 6.9, opinions remained surprisingly stable before (*M* = 83.57, *SD* = 20.86, *Mdn* = 91) and after the course (*M* = 84.07, *SD* = 18.32, *Mdn* = 90). While not statistically significant, the English LE group tended towards more disagreement after the course (*M* = 80.86, *SD* = 18.62, *Mdn* = 80) compared before (*M* = 85.32, *SD* = 15.08, *Mdn* = 83.5). The Scots LE and non-native English LE groups showed the opposite trend, with more agreement after the

course ($M_{Scots\ LE}$ = 89.38, $SD_{Scots\ LE}$ = 15.54, $Mdn_{Scots\ LE}$ = 100; $M_{Non\text{-}native\ LE}$ = 83.35, $SD_{Non\text{-}native\ LE}$ = 19.91, $Mdn_{Non\text{-}native\ LE}$ = 91) compared to before the course ($M_{Scots\ LE}$ = 84.25, $SD_{Scots\ LE}$ = 28.77, $Mdn_{Scots\ LE}$ = 100; $M_{Non\text{-}native\ LE}$ = 81.10, $SD_{Non\text{-}native\ LE}$ = 19.75, $Mdn_{Non\text{-}native\ LE}$ = 85). Notably, as with the first statement about the status of Scots as a language, the Scots LE group is the only one wherein a participant strongly disagreed pre-course (i.e., rating of 0). By the end of the course, though, the lowest rating was 50, reflecting a neutral sentiment.

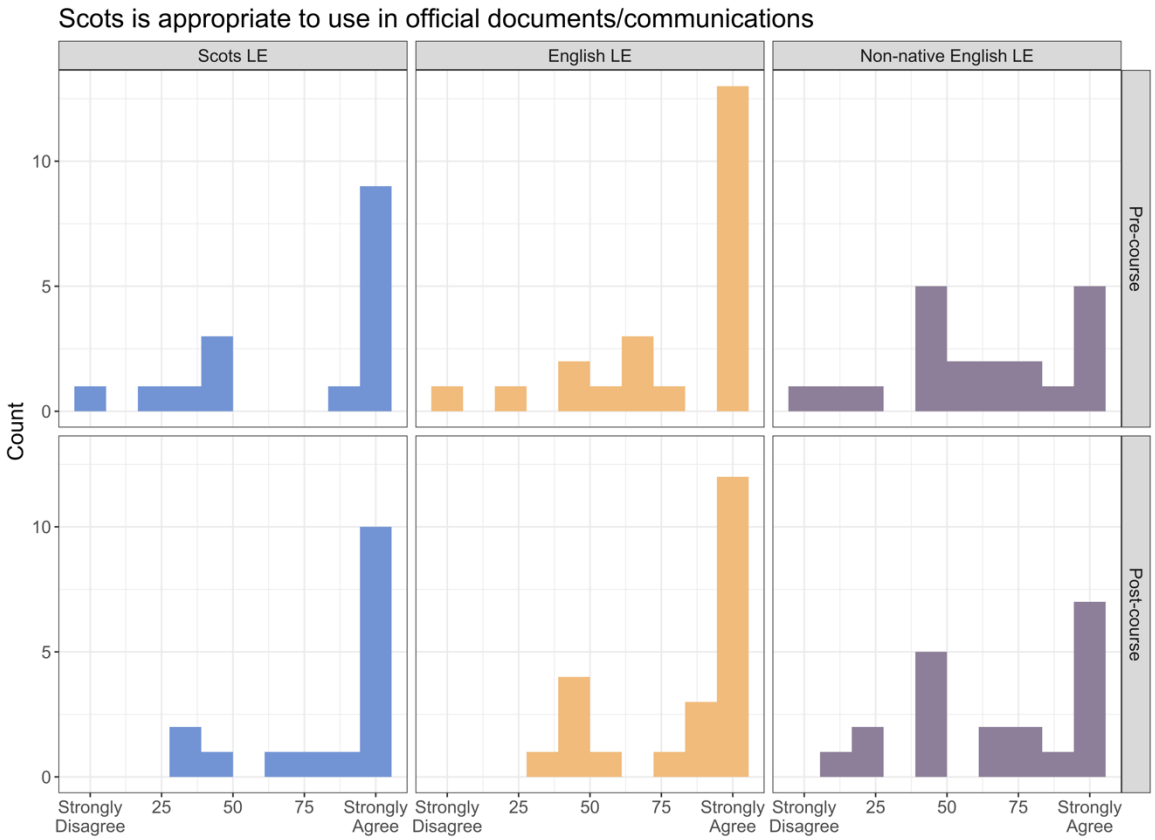
Figure 6.10 Language attitude VAS statement 4 by LE group and session



The latter two questions focus on the perceived (appropriate) utility of Scots. Opinions on the fourth VAS statement, *"Scots should be spoken but not written"* remained stable overall before (M = 5.21, SD = 13.64, Mdn = 1) and

after the course (*M* = 6.33, *SD* = 16.23, *Mdn* = 0). Indeed, the vast majority of participants across LE groups strongly disagreed with this statement both before and after the course (selecting 0, hence the median of 1; see Figure 6.10). Similarly, all LE group opinions towards written Scots being appropriate remained quite stable across the sessions, due to high levels of disagreement (and therefore agree with Scots being appropriate to write) to the statement: Scots LE before (*M* = 2.66, *SD* = 5.90, *Mdn* = 1) and after the course (*M* = 5, *SD* = 17.45, *Mdn* = 0); English LE before (*M* = 5.09, *SD* = 11.3, *Mdn* = 1) and after the course (*M* = 4.04, *SD* = 7.88, *Mdn* = 1); non-native English LE before (*M* = 7.37, *SD* = 19.52, *Mdn* = 1) and after the course (*M* = 9.9, *SD* = 21.47, *Mdn* = 1).

Figure 6.11 Language attitude VAS statement 5 by LE group and session



Overall, opinions towards the appropriateness of Scots in official capacities shifted towards more agreement at the end of the course (*M* = 78.55, *SD* = 27.11, *Mdn* = 93) compared to before the course (*M* = 72.69, *SD* = 31.44, *Mdn* = 86.5). As shown in Figure 6.11, within the three LE groups there was also general trends of slightly higher levels of agreement at the second testing session: Scots LE before (*M* = 74.63, *SD* = 34.61, *Mdn* = 100) and after the course (*M* = 84.38, *SD* = 25.55, *Mdn* = 100); English LE before (*M* = 79.5, *SD* = 29.28, *Mdn* = 100) and after the course (*M* = 83.5, *SD* = 23.42, *Mdn* = 100); non-native English LE before (*M* = 63.65, *SD* = 30.50, *Mdn* = 62.5) and after the course (*M* = 68.45, *SD* = 30.32, *Mdn* = 73).

Relating to age, across the 5 VAS style questions respondent age was not significantly correlated with the response patterns.

6.3.2.4 Qualitative analysis of language attitudes

While at the surface, the quantitative measures of language attitudes show an interesting trend where some opinions remain more or less stable and others (in some cases quite dramatically) shift after the courses, the qualitative data analysis further illuminates a nuanced and complex set of experiences unique for the two primary groups: Scots LE and non-Scots LE. The three developed themes are discussed below: (a) "All signs point to a language by any other name", (b) "Identity metamorphosis: from cringe to unleashing the unicorn", and (c) "Are you mocking me? Am I mocking you?".

6.3.2.4.1 All signs point to a language by any other name

The theme "All signs point to a language by any other name" relates to the data from the non-Scots LE groups (i.e., English LE and non-native English

LE). Reflecting the status and prestige of Scots and the higher levels of awareness of Scottish Gaelic, some participants (both living in and outwith Scotland) shared that they were not actually sure of what Scots was before starting the course and/or had confused it with Scottish Gaelic *("I didn't realise the Gaelic-Scots divide in Scotland. I think for a while I thought they were one [language]"* – Freya, 26). As Harrison (37) explains,

> *"I knew nothing about Scots, I didn't even really know it was a language to be honest. I've definitely learned a lot about Scots. I mean, I wouldn't really have known anything about Scots, apart from ironically the Wikipedia. [general laughter] I remember where I think I'm sure I used to show it to friends as almost a joke. And it's like, 'Hey, look, you can understand Scots, just like English and funny spelling.' "*

All participants shared that after the course, whether they had not known about Scots or had considered it a dialect of English, they now conceptualise it as a language. One student shared that the status of Scots as a language in its own right was reinforced by sharing with friends what was being taught in the course by *"exposing say friends to it, to the work I was doing and seeing their complete lack of comprehension of it was like, 'Okay, yeah, this is this has to be a language because they're not, they're not processing any of this.'"* (Lily, 40).

At different stages of working through the course materials and covering such a depth of content around Scots, points of reflection were triggered on the status and prestige of the language in relation to English. In particular, the grammatical and syntactic differences between Scots and English were highlighted by participants: *"So, like at first, I thought that, 'Ah like Scots is probably rolling your /r/s, and then like changing some vowels up and stuff' and then, but then I noticed that there are actually grammatical differences"* (Olivia, 22). Additionally, taking the sentiment further, Dylan (26) explained that after taking the class and, *"actually study[ing] it like in an actual*

focused, serious manner. And now I'm really like, 'Oh my God, people think this is really not a language, like, have you, have you looked at it? Have you seen it? Like, come on, just like really?'".

Further, not only were student perceptions of the language status challenged, but as Ivy (40) shared, also the community of its speakers: *"Scots being a language in its own right… I already knew that, but in terms of the community speaking it. At first, I thought it was mainly spoken by the elderly, but I realized it's not the case."* In particular, students who had Scottish relatives shared their experience through the course forced them to confront previous bias towards Scots and speakers of Scots, resulting in feelings of retrospective embarrassment:

> *"[the course] changed my views in in many ways…I probably did, I probably was inclined to think that Scots was a dialect of English. I rapidly probably from day one, first lesson almost I was persuaded – Scots is a language not a dialect… I had some prejudices, I would say, quite embarrassing prejudices, for example, I thought [a distant family member] had a speech defect of some kind. I didn't realize…that's how you say it in Scots… I guess I've learnt more respect for Scots as a language and for the way it's spoken. It is not people making mistakes or being uneducated, it's just because it has a different grammatical structure"* (Arthur, 67).

6.3.2.4.2 Identity metamorphosis: From cringe to unleashing the unicorn

The theme "Identity metamorphosis: from cringe to unleashing the unicorn" encapsulates the sometimes painful and uncomfortable discovery and revelation that the Scots LE group participants experienced across the course in just 3 weeks. I chose "metamorphosis" for this theme name as it entails a process, often sudden and extreme, but a process wherein there are different (in some cases non-linear) stages. Ubiquitous throughout the Scots LE group discussions was the experience of *"being told in school that Scots was bad*

English and reprimanded for speaking it" (Henry, 74) or *"believ[ing] people when they said Scots was 'only a dialect'"* (Charlotte, 36). Indeed, as Sophia (50) explained, being consistently told not to speak Scots resulted in *"a natural inhibition for me not to speak Scots, as I was always taught to speak 'proper'".*

Numerous participants reported a sense of discomfort speaking Scots initially as it went against their *"internal soliciting around speaking and using Scots"* (Violet 55). While not all Scots LE participants completed the course at the same stage of this identity metamorphosis, all participants shared that the course triggered the process in some way by *"validat[ing] my own experience of Scots"* (Eliza, 42). As Charlotte (36) describes,

> *"It was funny to find that the others in the course shared that same feeling almost of shame about our own language, accent, and the way we speak in general, even though we were from all over the country and of different age groups. It was a relief to learn that all those words/pronunciations we were told off for at school were actually normal and to learn how our language has evolved and changed across the country."*

Orla (55) not only echoed this sentiment, but also furthered highlighted the emotional experience of the process of accepting and recognising Scots as a language which *"did come as a revelation. Actually, there were quite a few things I found really revolutionary. And I find myself having almost an emotional response on a couple of occasions, to realize that, that… to realize just how much the cringe has been battered into us about our own culture and our own language."*

Indeed, after the course all participants in the Scots LE group referred to Scots with personal possessive pronouns (i.e., "my language", "our language"). I noted this variation in language during the focus groups. This shift from more distantly referring to Scots as "Scots" before the course to referring to Scots as a personal language after the course became even more

apparent throughout the analytic process, which illuminates the internal unfolding of the participants' personal consideration and categorisation of Scots and acceptance of it as a language bearing shared cultural identity and meaning. Orla (55) succinctly outlined the effect that the long-term systematic ideological undercutting of Scots has had by *saying "I have spent most of my life being actively disconnected from my language in formal and literary settings through an imposition of the primacy of English language on generations of Scots."* Further, Imogen (39) highlighted the negative impact societal disregard for Scots and its speakers can and has had by explain that *"and if the words we use are labelled as 'slang'… that will inevitably have a negative impact on your self-esteem and your identity. I don't want children to feel like the way they speak isn't valued and therefore, their thoughts and opinions aren't important."*

Poppy (56) describes her experience on the course an internal process of recategorising Scots from just slang to a language, as well as accepting the language generally as a means of becoming more comfortable speaking it:

> *"And it was trying to work out in my own head what slang, what was Scots, and what was somewhere in-between. It is almost accepting that, yes, I wasn't allowed to speak in a certain way, but when I am down in [an area on the West Coast of Scotland] so many people speak to me in Broad Scots. I understand them, but soon as I tried to speak to them, I would just get so tongue tied. And it just because it was drummed out of me. So maybe that might ease off? Because I am now more accepting of it. I am less… I am less concerned about sounding slang, if that makes sense".*

Violet (55) explained her experience in the course was not only that of a metamorphosis, but an unfolding, illumination, or rediscovering a part of herself: *"I had numerous realisations as to why I construct sentences incorrectly in English. I discovered I use Scots words that don't exist in English… I found a*

dormant part of myself - more poetic and passionate when I think in Scots...It's like recovering part of myself that went to sleep or was cut off." For other participants the internal changes resulting from the course were a more direct realisation and acceptance of being Scots speakers and resulting desire for the language to escape the negative stigma. Henry (74) directly outlines,

> *"[this course] made me realise that my childhood language up to 5 years old was Scots, not Scottish English…One of the questions on the first questionnaire was what was the household language in childhood. And I put Scottish English, but in reality, and listening through the class, I realised that I should have put Scots. We were speaking Scots, purely, within the home. It was only that I got to school that I had to 'speak proper' as it was called… there was that pressure that English was proper, and Scots was slang. Now, I got along partly with that for many, many years, although I still thought it's a language. But what became clear is that, in fact, I was speaking Scots all the time! Like the man who discovered he was speaking prose when told about the difference between prose and poetry. So that was a real revelation to me and just reaffirmed to me that Scots is a language… We should be able to just do contemporary politics or whatever, in Scots, without apology, without embarrassment. We're not all Groundkeeper Willie type characters from the Simpsons."*

Participants in further stages of the metamorphosis process reported already altering the way in which they go about their daily life as a result of the course. For example, Orla (55) reflected on the ways in which the course has tangibly impacted her life in just 3 weeks, resulting in her leaning into the language as an expression of identity and culture:

> *"I found [the course] actually really liberating. And I've stopped correcting [her child] when [they] talks about 'them ones', because now I realize that there is a historic prominence to that grammatical structure [they're] using, and I now know that I can unleash the unicorn much more in language…this year for the first time, after doing Michael's course, I filled out that we spoke Scots at home when I filled out my [child's] census."*

6.3.2.4.3 Are you mocking me? Am I mocking you?

While both LE groups at some level found the course to be challenging of their opinions and attitudes towards Scots, an interesting pattern of diverse opinions on non-Scots speaking Scots further highlights the different stages of metamorphosis for the Scots LE group and further challenges the non-Scots LE groups classification of Scots in relation to English. Some participants in the Scots LE group found the notion of non-Scots speaking Scots to be entirely unacceptable, highlighting the deeply embedded feelings of cringe and discomfort around the language, even in the face of starting the process of accepting it as a language and important element in identity. For example, Ruby (49) explained that

> *"if I was in the mixed class [mixed native and non-native Scots speakers] I probably would not have stayed [on the course]. I had to go to the beginner's class, and I could nae sit there when they were reading… put on accents was so distracting and I could not handle it."*

Underlying being unable to handle it is likely a sense that in "putting on accents" these speakers where not doing so in earnest, but rather with malicious intendent and mockery. Interestingly, considering the other classes as "beginner's" was not only something from the design of the study that we on the research team considered, but also a sentiment shared by other members of the Scots LE group who did not feel the same about non-native Scots speakers. Sophia (50) thought that *"it would be good for us to hear people of other nationalities trying to speak Scots because there's a differentiation then between Scots and English, and an equality, an assumed equality, which we know is true"*, however it may be pedagogically difficult to have these students in the same class due to mixed proficiency levels. Indeed, Lucas (29) reflected on the structure of the study saying, *"I think you did it in*

an ideal way actually, splitting people in terms of the questionnaire and the that questions you asked." Further, Henry (74) elaborated that it would be a complement in fact if other people tried to speak Scots, and he did not find different accents to be an issue:

> *"If I were in a class of mostly foreigners, non-native speakers, it wouldn't worry me one bit if their accent wasn't perfect. It is the fact that they are making the effort that is flattering in a way, you know, not flattering to deceive, but it's courteous and friendly and warm. So not speaking it with a perfect accent doesn't matter".*

However, the same could not be said of the non-Scots LE students themselves. Many participants in both the English LE and non-native English LE group shared that they had difficulties speaking Scots largely due to feeling unease around appearing to mock the Scottish accent. This highlights that while most participants completed the courses feeling Scots was a language in its own right, its status as a minority language nonetheless makes it vulnerable to a sense of mockery from majority language speakers. For example, Olivia (22) shared that

> *"I was not comfortable imitating the accent because it's so close to English, that I felt like I was imitating him [the teacher] personally. It just did not make me feel comfortable at all. I mean, when learning languages that are completely different, like Finnish, for example, I just say that 'Okay, this is the language, and this is how it's pronounced, and I will pronounce it that way.' But because it's so close to English, I just felt weird."*

Arthur (67) shared a similar sentiment underscoring a shift in his feelings about speaking Scots based on the reassurance from and perspective of the teacher:

> *"Just like you can speak English with a variety of different accents, you can speak Scots with a variety of different accents. Whereas I thought*

speaking Scots with an English accent would really be, you know, unacceptable."

On the other hand, for other students the discomfort around speaking Scots did not have to do with its status or relationship to English, so much as general self-consciousness that comes from learning to speak a new language:

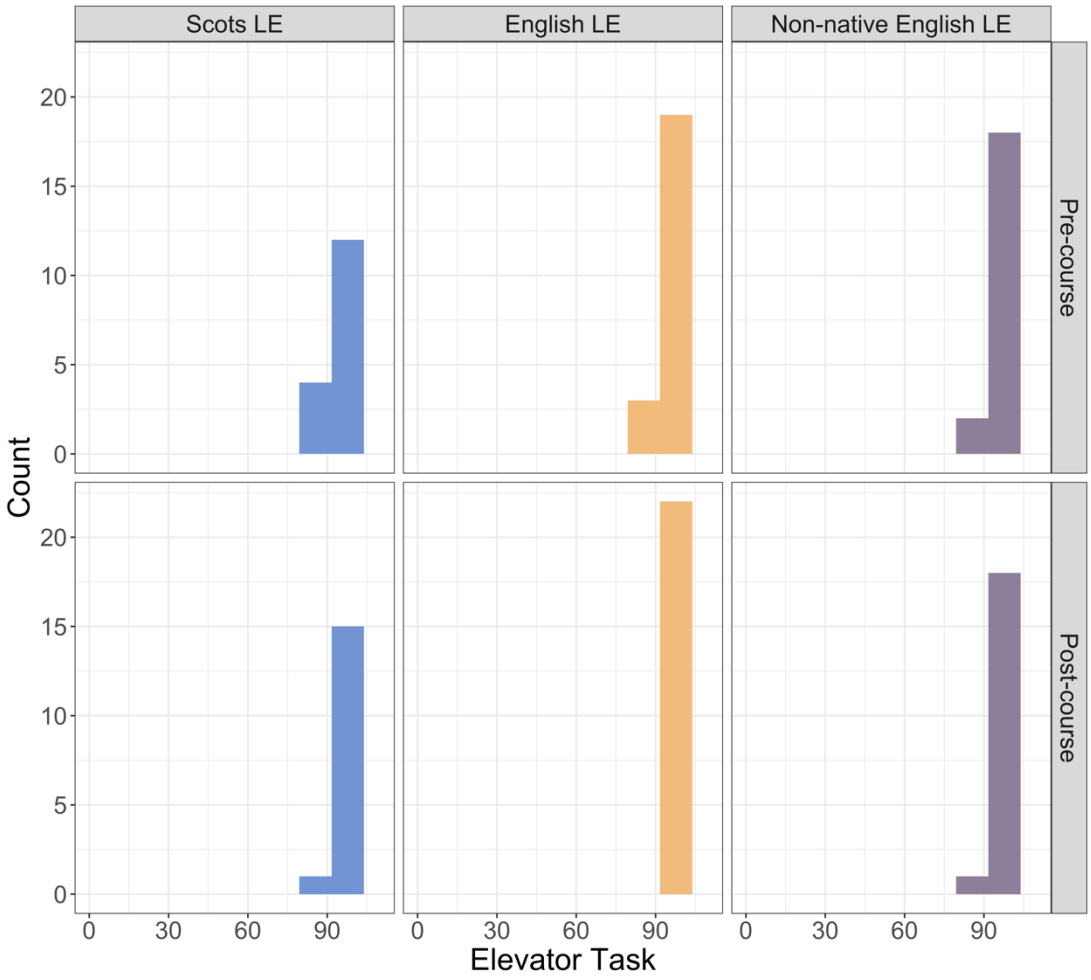
> *"I was a little self-conscious about the pronunciation, you know, and I personally try to get myself over that because I taught ESL [English as a Second Language], so I always used to tell my students 'Oh, don't worry about it. Don't be embarrassed, make your mistakes' and I was trying to remind myself that and yet I still felt self-conscious because I said 'Oh, I know I'm not saying this right at all'. [laughter] it's… it's a tough hurdle for people to get over, with learning a language and I'm not immune to it."* (Amelia, 49).

6.3.3 Cognitive outcomes

6.3.3.1 Elevator Counting

The Elevator Counting task, as expected given previous literature and that it is the least computationally difficult of the auditory subtests, had a ceiling effect, shown in Figure 6.12 (Bak et al., 2016c; Long et al., 2020; Vega-Mendoza et al., 2015). As such, it was not further analysed as part of this study.

Figure 6.12 Elevator Counting scores by LE group and session



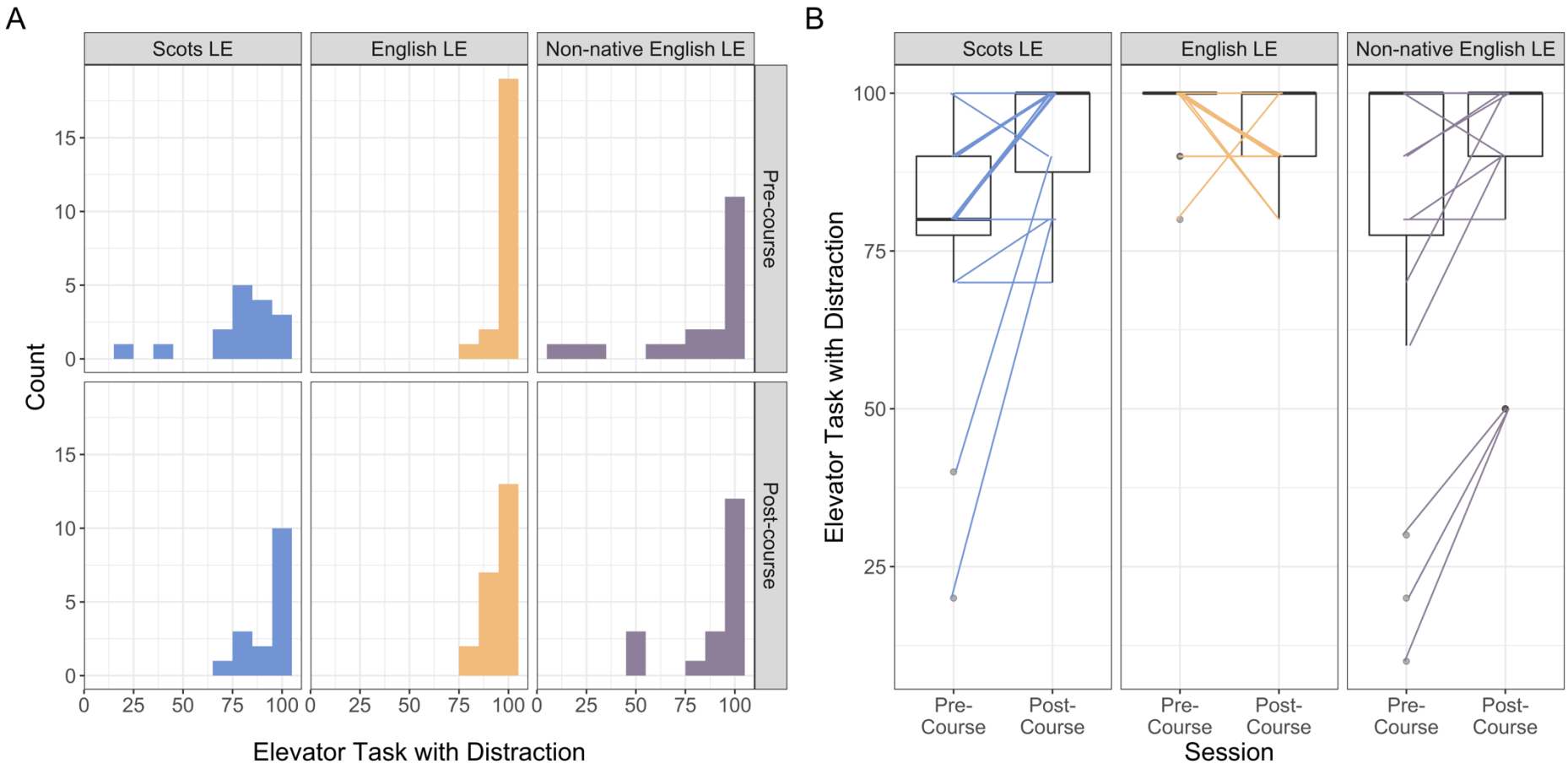
6.3.3.2 Elevator Counting with Distraction

The Elevator Counting with Distraction task showed more variability across sessions, though the English group scored at ceiling both before and after the class (Figure 6.13). On average overall, participants scored higher on the ETD task after the course (M = 92.63, SD = 12.61, Mdn = 100) than before the course (M = 87.07, SD = 22.16, Mdn = 100), W = 340, p = .007, with a moderate effect size (r = .31). When looking by LE group, Scots LE and non-native LE both improved on the ETD task after the course ($M_{Scots\ LE}$ = 93.13,

$SD_{Scots\ LE}$= 10.15, $Mdn_{Scots\ LE}$ = 100; M_{Non}-native LE = 89.47, $SD_{Non\text{-}native\ LE}$ = 18.40, $Mdn_{Non\text{-}native\ LE}$ = 100) from baseline ($M_{Scots\ LE}$ = 78.75, $SD_{Scots\ LE}$ = 21.56 $Mdn_{Scots\ LE}$ = 80; $M_{Non\text{-}native\ LE}$ = 81.5, $SD_{Non\text{-}native\ LE}$ = 29.07, $Mdn_{Non\text{-}native\ LE}$ = 100), ($W_{Scots\ LE}$ = 74.5, $p_{Scots\ LE}$ = .005; $W_{Non\text{-}native\ LE}$ = 42.5, $p_{Non\text{-}native\ LE}$ = .019), with a large effect size ($r_{Scots\ LE}$ =.73; $r_{Non\text{-}native\ LE}$ =.55). However, the English LE did not ($M_{Pre\text{-}course}$ = 98.18, $SD_{Pre\text{-}course}$ = 5.011, $Mdn_{Pre\text{-}course}$ = 100; $M_{Post\text{-}course}$ = 95, $SD_{Post\text{-}course}$ = 6.73, $Mdn_{Post\text{-}course}$ = 100). However, at baseline before the course, the English LE group scored significantly higher than the other two LE groups, $H(2) = 16.32$, $p < .001$, with a large effect size ($\eta^2 = .26$) according to (Field, 2013).

Therefore, a linear mixed model with random intercepts by participant was fitted, see Table 6.5. Controlling for LE group, participant age, language knowledge before the class, number of classes missed, hours spent practicing Scots, gender, and education, the testing session (β = 10.00, cluster sample bootstrap 95% CI: 2.45 – 17.57) and participant score on Raven's APM (β = 0.34, cluster sample bootstrap 95% CI: 0.23 – .044) were significant predictors of ETD scores. Further, there was a significant interaction between LE group and session wherein the English LE group did not improve the post-course testing session (β = -12.38, cluster sample bootstrap 95% CI: -20.73 – -4.24), though the Scots LE and non-native English LE groups showed a different pattern.

Figure 6.13 Elevator Task with Distraction scores by LE group and session



Note: Figure A shows histograms of ETR scores by group (Scots LE, English LE, Non-native English LE) and testing session (pre-course and post-course). Figure B shows 3 boxplots of ETR scores by testing session with the lines representing individual participants, one for each LE group.

Table 6.5 ETD model

Predictors	**Beta**	**95% CI**[1]
(Intercept)	71.93	**56.61, 90.21**
Session		
Pre-course	—	—
Post-course	10.00	**2.45, 17.57**
Group		
Scots LE	—	—
English LE	9.09	-3.56, 14.79
Non-native English LE	-9.57	-15.97, 2.55
Age of Participant	-0.06	-0.19, 0.08
Composite Score	-0.03	-0.07, 0.01
Raven's APM Score	0.34	**0.23, 0.44**
Classes Missed (out of 9)	1.55	-1.56, 2.19
Hours Practicing Scots	0.04	-0.02, 0.08
Gender		
Female	—	—
Male	1.38	-8.34, 10.3
Prefer to self-describe/ not to answer[2]	-5.20	-13.08, 19.87
Years of Education	-0.43	-0.86, 0.03
Class x Session		
English x Post-course	-12.38	**-20.73, -4.24**
Non-native English x Post-course	-1.13	-11.49, 8.50
Observations	107	

ICC	.70
Marginal R^2 / Conditional R^2	.32 / .79

[1] CI = Confidence Interval, case-based bootstrap with 5,000 samples

[2] prefer not to answer and prefer to self-describe combined due to small group sizes

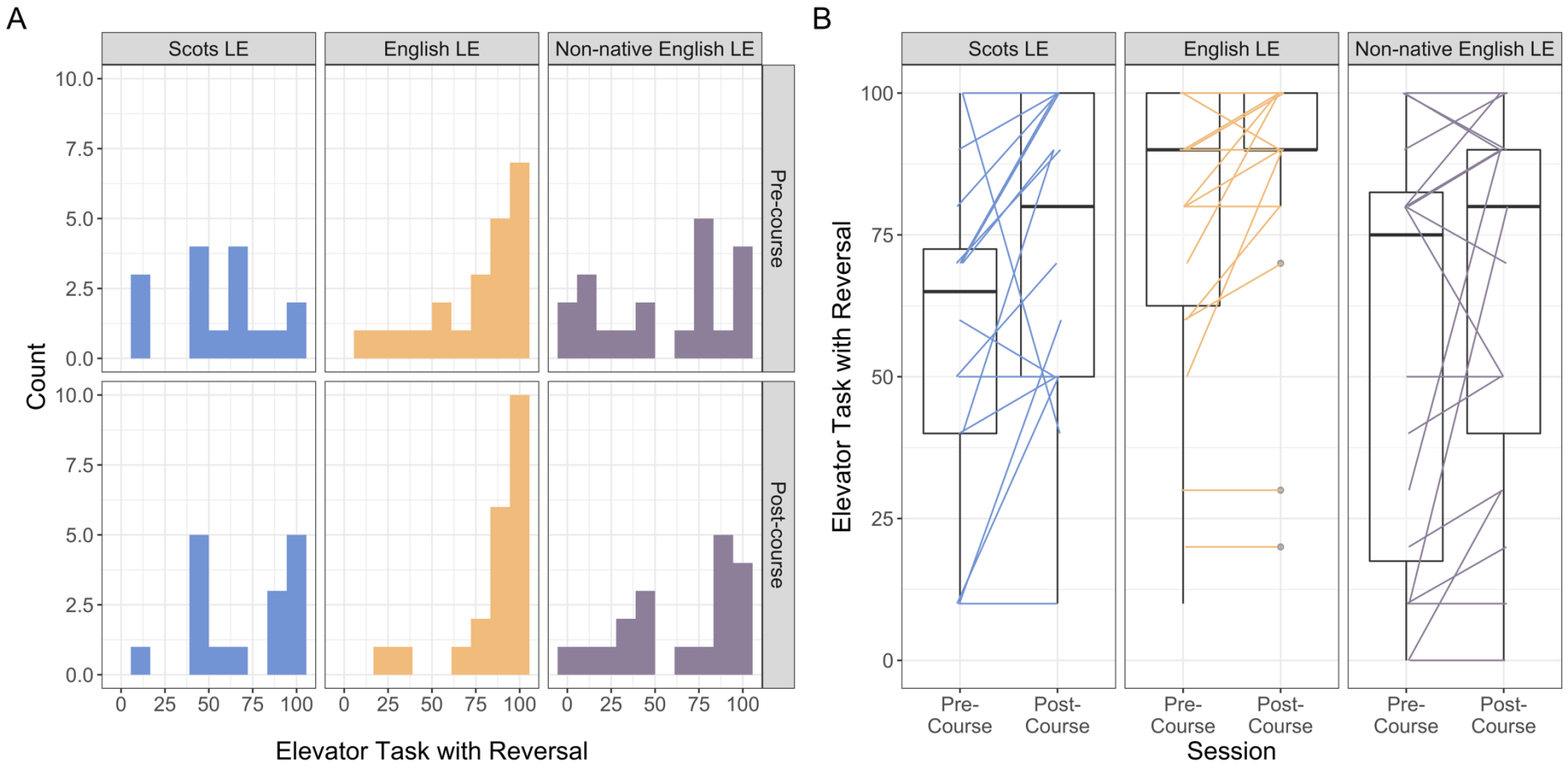
6.3.3.3 Elevator Counting with Reversal

Similarly, the ETR scores showed variability across testing sessions (Figure 6.14). Overall, participants scored higher on the ETR task after the course (M = 75.18, SD = 29.29, Mdn = 90) than before the course (M = 64.48, SD = 32.83, Mdn = 75), W = 559.5, p < .001, with a large effect size (r = .50). When looking by LE group the change in ETR scores after the course (M = 65.26, SD = 33.89, Mdn = 80) from before the course (M = 56.5, SD = 37.6, Mdn = 75) was not significant for the non-native English LE group. However, the Scots LE and English LE groups both significantly improved in ETR scores after the course ($M_{Scots\ LE}$ = 71.88, $SD_{Scots\ LE}$ = 28.1, $Mdn_{Scots\ LE}$ = 80; $M_{English\ LE}$ = 86.67, $SD_{English\ LE}$ = 22.21, $Mdn_{English\ LE}$ = 90) compared to baseline ($M_{Scots\ LE}$ = 57.5, $SD_{Scots\ LE}$ = 29.78, $Mdn_{Scots\ LE}$ = 65; $M_{English\ LE}$ = 76.82, $SD_{English\ LE}$ = 27.5, $Mdn_{English\ LE}$ = 90) ($W_{Scots\ LE}$ = 76, $p_{Scots\ LE}$ = 0.03; $W_{English\ LE}$ = 42, $p_{English\ LE}$ = 0.02), with large effect sizes ($r_{Scots\ LE}$ = .57, $r_{English\ LE}$ = .52).

Therefore, a linear mixed model with random intercepts by participant was fitted, see Table 6.6. Unlike the ETD model, the group by session interaction was not significant and did not improve the model fit, thus it is not reported here. Controlling for LE group, number of classes missed, hours spent practicing Scots, gender, and education, the testing session (β = 8.60, cluster

sample bootstrap 95% CI: 3.19 – 13.70), participant age (β = -0.71, cluster sample bootstrap 95% CI: -0.88 – -0.52), language knowledge before the class (β = 0.22, cluster sample bootstrap 95% CI: 0.14 – 0.29), and participant score on Raven's APM (β = 0.22, cluster sample bootstrap 95% CI: 0.08 – 0.36) were significant predictors of ETR scores.

Figure 6.14 Elevator Task with Reversal scores by LE group and session



Note: Figure A shows histograms of ETR scores by group (Scots LE, English LE, Non-native English LE) and testing session (pre-course and post-course). Figure B shows 3 boxplots of ETR scores by testing session with the lines representing individual participants, one for each LE group.

Table 6.6 ETR model

Predictors	**Beta**	**95% CI[1]**
(Intercept)	82.35	**57.71, 100.04**
Session		
Pre-course	—	—
Post-course	8.60	**3.19, 13.70**
Group		
Scots LE	—	—
English LE	-2.65	-7.47, 2.26
Non-native English LE	-14.58	-23.40, 8.77
Age of Participant	-0.71	-0.88, -0.52
Composite Score	0.22	**0.14, 0.29**
Raven's APM Score	0.22	**0.08, 0.36**
Classes Missed (out of 9)	1.06	-0.14, 2.47
Hours Practicing Scots	-0.29	-0.41, 0.18
Gender		
Female	—	—
Male	0.43	-3.10, 4.16
Prefer to self-describe/ not to answer[2]	-13.34	-29.53, 2.85
Years of Education	-1.21	-1.89, 0.51
Observations	107	
ICC	.70	
Marginal R^2 / Conditional R^2	.33 / .81	

[1] CI = Confidence Interval, case-based bootstrap with 5,000 samples

[2] prefer not to answer and prefer to self-describe combined due to very small group sizes

6.4 Discussion

The previous two language learning chapters focussed on two of the most widely spoken languages in the world, Spanish and Chinese (incidentally, also the two languages which show an increase in the number of learners in the UK, in contrast to the French and German, which have been the main languages traditionally thought at British schools (Tinsley & Board, 2017)). In this study, the focus shifted to a minority language, Scots, which has historically been undermined and disregarded as a slovenly dialect of English, drawing more attention to the questions of language attitudes and perceptions. This study set out to explore the impact of 2-hour Scots classes 3 times a week for 3 weeks on cognition as well as student attitudes towards the target language.

Participants joined the courses from 4 continents (Figure 6.1), demonstrating a genuine interest in Scots well beyond Scotland and the Scottish diaspora (e.g., Canada, the United States, Australia, New Zealand). For pedagogical reasons, students were separated into different classes based on their language background (i.e., Scots LE, English LE, and non-native English LE groups). The relevance of such a distinction was confirmed by participant comments as well as the language attitude results, which differ by group. Notably, across the three language learning studies presented within this thesis, this course had the lowest dropout rate, retaining 90%of participants who started the course. Using a mixed methods approach to data collection and subsequent analysis, participants were found to have undergone a general process of unlearning and relearning, or in the case of some participants discovering for the first time, that Scots is a language in its own right after just

3 weeks. The perceived mechanism as well as the quantity and depth of these changed opinions, however, vary by group.

The Scots LE group underwent an identity metamorphosis as a result of the courses, with participants reporting varied levels of surprise and discovery in confronting their socially influenced internal monitoring of Scots (reflected in the theme "Identity metamorphosis: from cringe to unleashing the unicorn"). The non-Scots LE groups also went through a process wherein their opinions changed, but it was less of a personal journey and more of a synthesis of facts and information leading to the conclusion that it is indeed a language separate, though related to, English (reflected in the theme "All signs point to a language by any other name"). Some students in particular highlighted newfound retrospective embarrassment in confronting their previous opinions towards Scots and the speakers of the language.

The historical and current socio-political factors influencing the status and prestige of Scots were found to drive individual differences in student's levels of comfort speaking Scots and/or hearing non-Scottish people speak Scots (reflected in the theme "Are you mocking me? Am I mocking you?"). While some non-Scots LE participants found speaking Scots uncomfortable for fear of coming across as mocking (due to the similarities to English), others attributed this discomfort to the general unease of first learning to speak a new language, regardless of the target language's status. The status of Scots as a minority language and its lateral relationships to English means that it is not as simple for participants as saying "it's a language therefore I will treat it as any other language" as there are still socio-political hurdles to overcome. Similarly, there were conflicting opinions among native-Scots LE speakers on their feelings towards hearing non-native speakers use the language. Responses fell on a continuum where at one extreme it was perceived as unbearable and

unacceptable, at the other it was seen as complementary and validating. Falling somewhere in the middle are those who did not see a difference between new adult speakers of Scots and new speakers of another language (i.e., *"That is not different than me going to France and trying to speak French with the French accent. Why not Scots?"* – Zachary, 49).

The quantitative data measuring change in language attitudes before and after the course further support the qualitative findings. Through a variety of question types (i.e., Likert, multiple choice, and VAS) three larger components of language attitudes were explored: Scots as a language, the (appropriate) utility of Scots, and Scots and identity. Interestingly, the Scots LE group overall agreed more strongly on the role of Scots in Scottish cultural heritage than in their personal heritage after the course. Though in both cases there were very high levels of agreement, this supports the qualitative findings that participants were on their own individual journey of metamorphosis after the course, falling at difference stages.

Another notable finding was that compared to the non-native English LE group, the English LE group reported higher levels of agreement to statements around personal identity and Scots. This is likely due to the many participants within the English LE group having Scottish family members or a Scottish lineage, and thus feeling a greater sense of connection as a result to Scottish culture and the Scots language. Alternatively, while some participants within the non-native English LE live or lived in Scotland, the majority had never been to the country or indeed heard the Scottish English accent. Therefore, these students were more likely to be attending the course due to general interest rather than a personal connection to the language or culture.

Around the opinion of Scots as a language, the most direct statement also proved to be the only item which changed after the course to a statistically

significant degree (in either direction – more positive or more negative): *"Scots is a language in its own right."* This change in opinion to statistically higher levels of agreement was largely driven by the English LE group. Further, this question is the only statement where in there was no further disagreement from participant's opinion before the course, regardless of LE group. Which is to say, a 3-week course on Scots unidirectionally (positively) altered all student's opinion of Scots as a language. Remarkably, the only group in which a participant before the course gave a rating below 50 (representing a neutral opinion) was in the Scots LE group, where a participant rated 0 (representing strong disagreement to Scots being a language). After the course, all students in the Scots group agreed strongly (rating of 100) to Scots being a language in its own right. This means that after just 3 weeks this participant entirely changed their opinion towards Scots.

While in the Scots LE group the question about Scots' status as a language resulted in 100% agreement after the course, questions around the appropriate utility of the language continued to show more variance. Before the course, a handful of participants in all 3 LE groups rated the statement *"Scots is appropriate to use in official documents/communications"* with a 0 (strongly disagreeing). Overall, the majority of participants maintained or shifted their opinions to higher levels of agreement after the course, the individuals in all 3 LE groups who initially strongly disagreed changed their opinions to strongly agreeing (a 100-point shift) after the course (Appendix O).

Taken together, these three components of language attitudes can then be translated into the change in opinions on statements to the future of the language in 50 years' time. Across all LE groups before the course most participants reported that they would like for Scots to be spoken by more people than now and this remained the case after the course. There was a

notable shift in the English LE and non-native English LE group responses after the course around how many people they think will be speaking Scots. In the English LE group before the course, "fewer people than now" was the most common answer, though after the course this was replaced by "the same as now" followed by "more people than now." In the non-native English LE group, the most common answer after the course was "more people than now." While it cannot be tested experimentally, this shift in personal opinion and optimism around the future of Scots is likely due to learning through the course more about the language and its speakers, both past and present.

Notably, across the impact of the courses on attitudes towards Scots, there was a similar pattern of results across age groups. That is to say, including both the qualitative and quantitative data, participant age did not result in different patterns of responses. The most distinguishing between-subjects factor in language attitudes was the groups based on the linguistic environment in which the participants grew up (i.e., growing up in a Scots LE or non-Scots LE). In the other language learning studies within this thesis, participants would often email me or mention in the feedback session their gratitude for being provided the opportunity to participate and learn the target language. However, no one had thanked me for choosing to study the target language specifically, rather it was the opportunity and experience. In this study, a handful of participants from the Scots LE group individually thanked me personally over email (entirely unprompted) or in the questionnaire for finding value and importance in Scots and dedicating a portion of my PhD research to it. The decision to study Scots itself was seen as something unique and worthy of recognising, outwith the opportunity of the classes being provided, further evidencing the unique status of Scots as a source of cultural

identity not shared by globally popular majority languages like Spanish or Chinese.

Relating to the cognitive research question, overall participants scored higher in measures of inhibitory attention and attentional switching after the course. Inhibitory attention scores improved overall in the second testing session (i.e., after the course) and those who scored higher on the Raven's APM before the course also scored higher at the first testing session. As mentioned in Section 6.2.1, the English group in particular was multilingual with many polyglots, though language knowledge was self-reported. As such, it is not surprising for the pre-course ETD scores to be near ceiling in this group, and indeed significantly higher than the other groups (see Figure 6.13). Further, due to being at ceiling before the course, the significant class by session interaction in the ETD model is due to some participants in the English group not receiving a perfect score in the second testing session. The English LE group median remained at 100, though the mean shifted 3.18 points (from 98.18 before the course to 95 after the course). Therefore, the group overall showed a different pattern from the Scots LE and non-native English LE groups who both improved after the course, with the Scots LE group improving the most. In terms of attentional switching, as measured by the ETR, scores were also significantly improved after the course. In this measure, older participants scored lower in the first session, though improved after the language learning course in the same pattern as the younger participants. Additionally, in line with prior literature, more multilingual participants scored higher at the first testing session.

While both an effect was found for both inhibitory attention (ETD) than attentional switching (ETR) in line with the two preceding studies presented within this thesis, the strength of the effects was not as strong as those found

in the Spanish study (Chapter 4). The two key differences between these studies are (a) the duration where the Spanish study was 1 week longer resulting in 6 more hours of language training (Spanish classes 3 times a week for 4 weeks, Scots classes 3 times a week for 3 weeks) and (b) the status of the language. The additional 6 hours of language learning may explain the differences, which could be explored in future research using a majority language such as Spanish at the 3 times a week for 3 weeks interval.

Another, not mutually exclusive possibility, however, is that the status and prestige of Scots, its close proximity to English, and resulting attitudes towards the language could explain these findings. Within the classroom setting, inhibitory attention as measured by the ETD maps onto selective listening while attentional switching relates to speaking and switching between languages. If students are reluctant to speak the language, as suggested by the qualitative findings, they may spend more time listening than verbally using the language. That is, training in a more passive way. While data was collected around the time spent practicing Scots outwith the classroom and using what materials, we did not measure *how* students reported practicing (reading, writing, speaking, or listening) within or outwith the classroom. Future studies could further explore this hypothesis by comparing more passive learning, such as a Latin translation course, with a more active speaking-focused course.

Taken together, these findings contribute to the wider literature on the cognitive effects of language learning as a novel investigation of mixed-age minority language classes for 9 hours a week across 3 weeks. The findings of the present study suggest that a thorough study into the cognitive effects of language learning should also consider student attitudes towards the language as they may influence the amount of time or manner of practicing the language, which in turn may influence cognition. Furthermore, this study

demonstrates that language status and prestige most certainly influence language attitudes and the individual experience of participants taking the course. The findings on language attitudes towards Scots paint a hopeful picture for the future prestige, function, and status of the language, suggesting that people of different backgrounds not only have an interest in Scots, but upon learning more about the language in a classroom setting are likely to improve their opinions (if originally negative) on the language and its speakers. As Eliza (42) states, "Scots needs to be spoken, used and promoted because otherwise it'll just become a curiosity, or will be viewed as a few words of localised slang, which would be awful". This study supports the work of Scots advocates, such as Dr Dempster, raising awareness of Scots, least of all amongst its speakers who must dismantle centuries of socio-political forces stigmatising and undermining the language.

Chapter 7 Online Language Learning in the Third Age: Recommendations to Improve Learning Experiences

The following chapter consists of two parts: a study which is completed and presented in full and another which is shorter as it is still in progress. Both studies are similar methodologically as well as in scope as they transition from the preceding applied language learning chapters into a translational applied contexts.

The first study explores perceptions, attitudes, and experiences which are not language specific by investigating one of the key unexpected variables introduced into the studies of this thesis by CoViD-19: online learning. In particular, this chapter wholistically explores the online learning experiences of older adults throughout the studies presented in previous chapters. It is the result of reflection on the feedback and experiences of older adults specifically across the database collected throughout this thesis of mixed age language learning courses from the preceding three chapters as well as data from language learning classes not reported as part of this PhD though that I collected during my PhD studies (Center for Open Learning Languages for All classes).

The chapter as it is presented here is published in the Journal of Gerontology and Geriatrics Education. The paper is a collaboration between myself and a fellow PhD student in the Netherlands also researching third-age language learning, whom I met at a conference in 2020. We have used our independent studies/sources of data and through qualitative analysis found converging themes around the challenges and benefits of older adults' online language learning experiences. Taken together, we provide concrete recommendations for online learning success with this age group, language and otherwise.

The second study, which is in progress, explores the perceptions, attitudes, and experiences of both learners and teachers in mixed-age language learning classrooms.

van der Ploeg, M. & Blankinship, B. (2022). Online Language Learning in the Third-Age: Concrete Recommendations to Improve Seniors' Learning Experiences. *Gerontology & Geriatrics Education.* https://doi.org/10.1080/02701960.2022.2143357

7.1 Introduction

The CoViD-19 pandemic has shown an increase in online activities which language learning could be a part of. This affects third-age language learners as well. As people grow older, there is more time and opportunity for leisure activities such as language learning. Seniors might travel more, want to learn a language to communicate with family members (e.g., children's spouses, grandchildren) that speak another language, or learn a language for personal enrichment purposes. Additionally, recent studies also seem to point in the direction of language learning providing older adults with cognitive benefits (e.g., Antoniou et al., (2013)). Despite the variety of reasons for language learning in older adulthood, very few studies have investigated the process of language learning itself with this age group. Further, there are very few studies on digital language learning in older adulthood that offer concrete pointers on how to go about online language courses. Ware and colleagues (2017) investigated the feasibility of technology-based courses for seniors in a four-month English course for French seniors. Their results showed communication with the outer world to be a main motivational factor for seniors, and technology presents a tool for that. More specifically, Boulton-Lewis and Tam (2018) suggest the use of blended language learning designs, even by simply using the internet. As the majority of seniors already use technology, they argue that technology can enrich seniors' learning environment. Crucially, however, the frequency of technological use tends to be less in seniors than for younger people. Seniors, moreover, have not been found to readily embrace new technology in general (Purdie & Boulton-Lewis, 2003; Wu et al., 2015). Still, several studies suggest technology-based senior language courses can be feasible, enjoyable and motivating (Djoub, 2013; Ware et al., 2017; for a review see Bosisio, 2019). Additionally, the use of

technology might even have additional benefits on top of learning new language skills: learning how to use technology can add to feelings of achievement (G. Lee et al., 2013) and has even been suggested to contribute to well-being and sense of empowerment (Shapira et al., 2007).

Although the literature so far has thus looked favourably towards the use of technology in the third-age classroom, there are no specific recommendations on how to go about this. Moreover, the studies to date are merely based on intuitions rather than grounded in actual third-age online language class observations. Djoub (2013) does offer a selection of factors that must be considered when designing an online/blended (language) learning course for seniors: (a) a needs analysis prior to the course onset (cf. van der Ploeg et al., submitted) with questions on the technical skills of participants; (b) relatively simple and intuitive computer programs/software; (c) no exercises that demand speedy processing and time-restricted responses (due to potential working memory problems); (d) the ability to interact with peers and share work and an online environment; and (e) initiation of a reflection process by teachers should about seniors' learning process specifically in relation to the technology used.

Despite the fact that Djoub (2013)'s recommendations offer a framework, they do not offer a concrete handbook on how to design an online (language) course for seniors and the recommendations are not rooted in observations either. In this paper we aim to do exactly this. Our study presents a bottom-up analysis of multi-site observational data, collected during the CoViD-19 pandemic, aiming to answer the research question: What should be considered when designing and implementing online language learning courses for seniors?

7.2 Methods

7.2.1 Data

The data for this paper originate from different independent language courses and focus groups in both the Netherlands and Scotland comprising a total of 73 learners. As both contexts include language classes, the combination of the Dutch and Scottish data will allow for a stronger generalisability of recommendations for online third-age language classes. The data were collected between May 2020 and August 2021. The Dutch study targeted healthy seniors (i.e., no cognitive deficits) of 65 and older for a three-month English class. Before the start of the course participants signed an informed consent letter explaining the purpose of the study and data management. The study was approved by the ethics committee of the University of Groningen (CETO; ref: 61890455). The Scottish dataset includes participants outside Scotland as the online delivery allowed for participants to join the classes where the time zone difference permitted. This dataset comprises three language learning studies: introductory Spanish and Scots classes as well as German, French, Italian, Japanese, and Polish "Languages for All classes" at the University of Edinburgh's Center for Open Learning. These studies targeted mixed-age classes of healthy adult learners (i.e., 18 plus). For the purposes of the present discussion, only data from learners of 55 and older are included. Similar to the Dutch study, participants signed an informed consent letter prior to the start of the course and the first research session. The study was approved by the University of Edinburgh PPLS Research Ethics Committee (ref: 440-1819/8).

Figure 7.1, Figure 7.2 and Table 7.1 below present the data as well as descriptive statistics in more detail. For clarity, we have delineated the dataset

as either the Dutch or Scottish data based on where the researchers are located.

Figure 7.1 Break down of Dutch data - English classes and focus groups

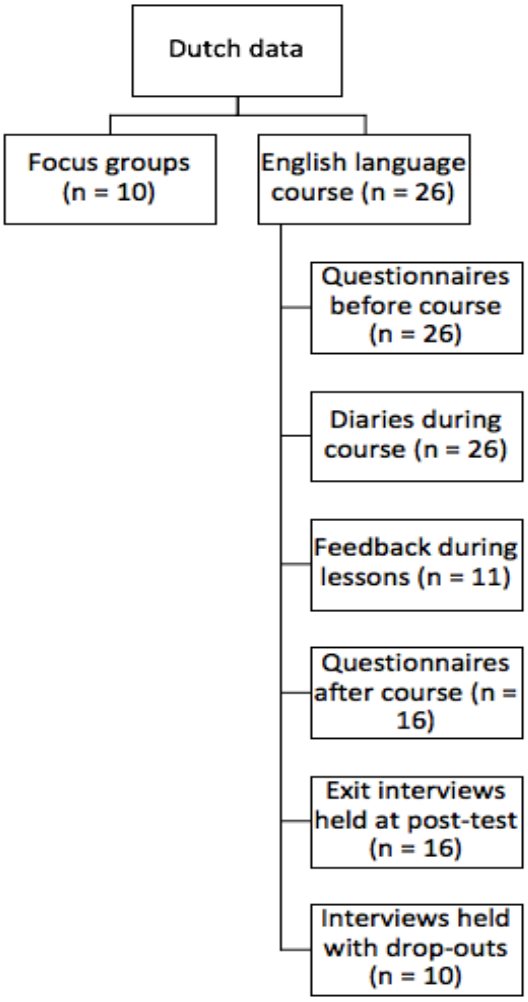


Figure 7.2 Breakdown of Scottish data

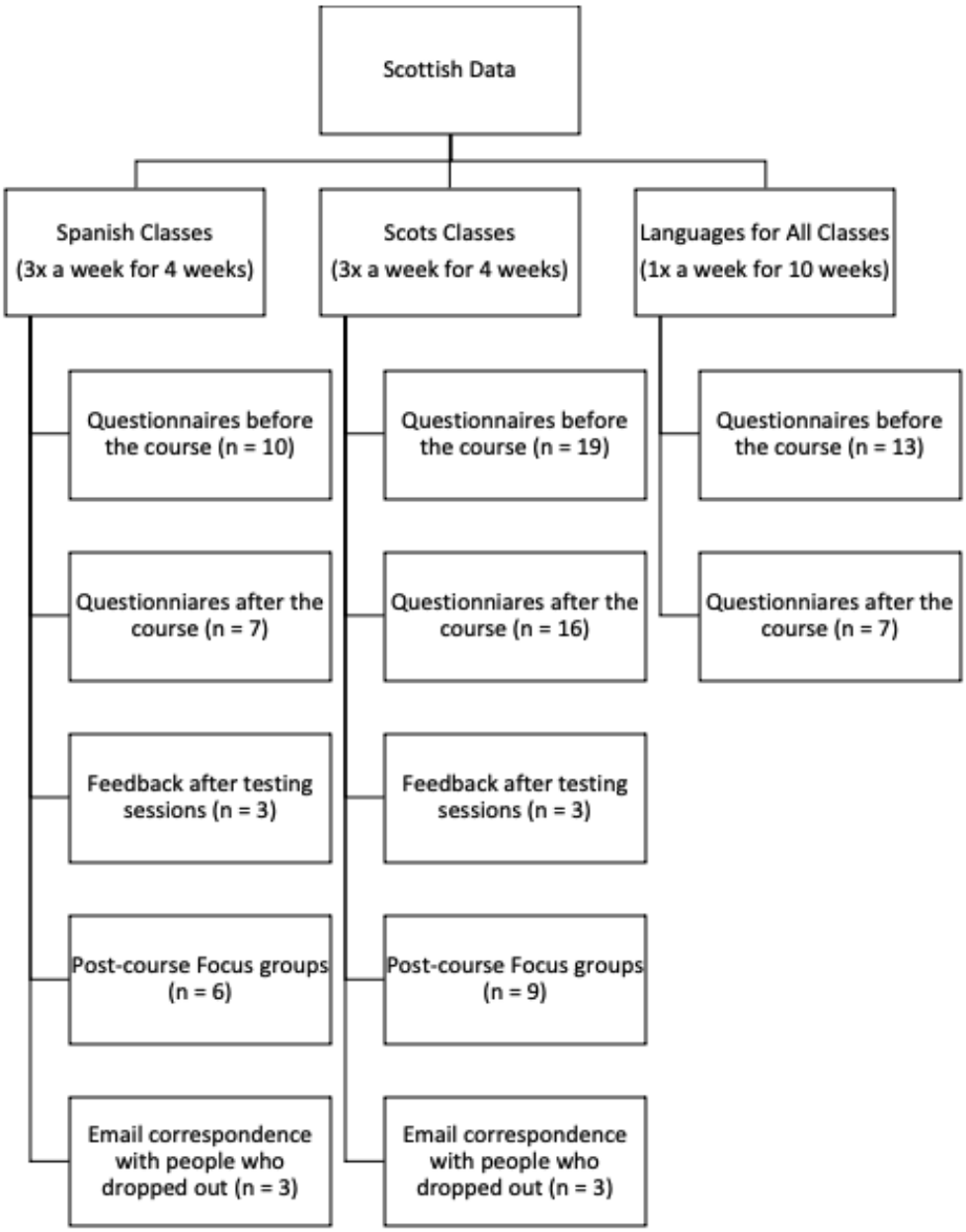


Table 7.1 Descriptive statistics for each dataset

Descriptive Statistic	Dutch	Scottish
Total number of learners	31	42
Gender	22 female, 9 male	24 female, 16 male, 2 prefer not to answer
Age range	65-87	50-74
Mean age	72.4	62.7
Educational level (Primary school - Doctorate)	Highschool - University	Highschool - Doctorate

7.2.2 Analysis

Data sources include both written and spoken language (see Figure 7.1 and Figure 7.2). Spoken data from focus groups/interviews were transcribed orthographically/verbatim[4] and then further analysed. For specific data sources see Appendix P. Prior to the initial coding stages, all participant names were anonymised.

Given the qualitative nature of the data and the aim of identifying patterns of meaning across the respective datasets, a reflexive thematic analysis (TA) approach was adopted (Braun & Clarke, 2006, 2021b) (see Section 2.3.1.5.3). Reflexive TA is a theoretically independent research tool wherein the framing theory must be determined separately, as opposed to a theoretically

[4] For the Dutch data the quotes were translated into English while attempting to keep the unique character of the original quote.

informed research framework, or methodology (Braun & Clarke, 2021b, 2021c). This method is used for exploring patterned/across case-meaning (Braun & Clarke, 2021b, 2021c). TA is not a singular method with a sole set of procedures, but rather can be thought of a family (Fugard & Potts, 2020) or "spectrum of methods—from types that prioritise coding accuracy and reliability to reflexive approaches that emphasise the inescapable subjectivity of data interpretation" (Braun & Clarke, 2021b, p. 37). Under reflexive TA, themes (i.e., patterns of shared meaning built upon a central organising concept) are developed across cases from codes following a recursive six-phase process of: "familiarisation; coding; generating initial themes; reviewing and developing themes; refining, defining and naming themes; and writing up" (Braun & Clarke, 2021b, p. 39). This inductive approach is especially suited to answer questions related to people's experiences, views and perceptions and results in themes explaining these views.

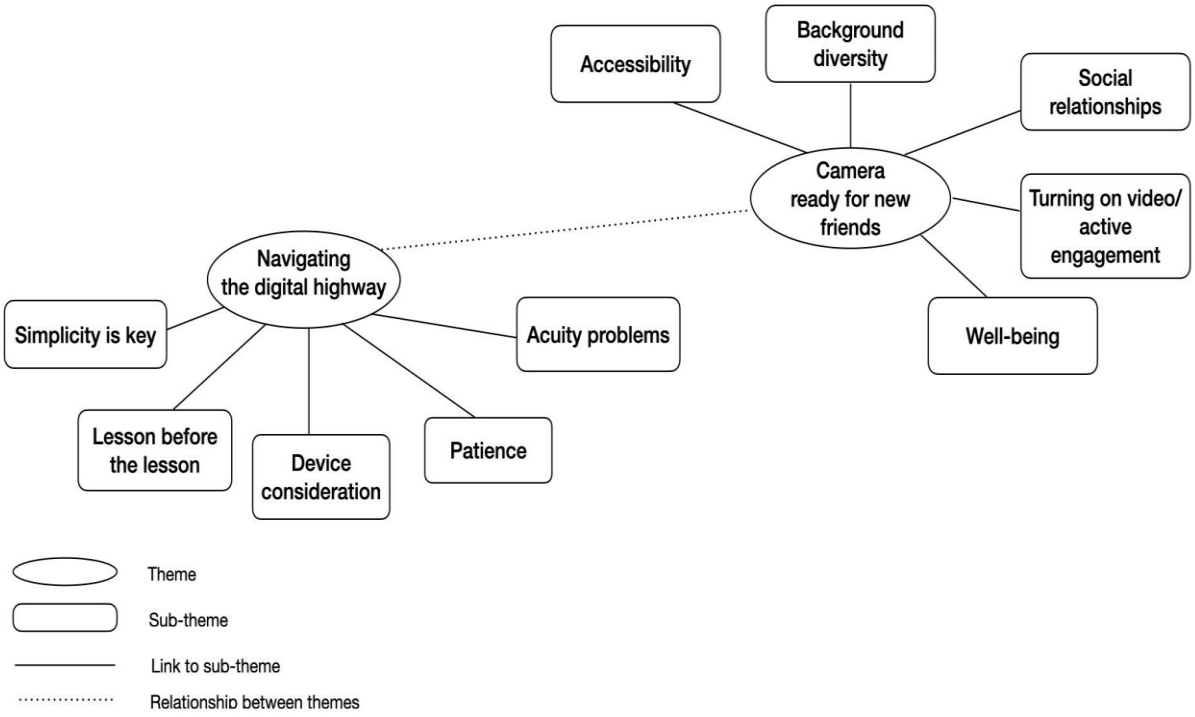
We adopted a critical realist epistemological stance in our analysis, recognising situated and interpreted realities influenced by social and cultural resources. Both authors independently immersed themselves in both the Dutch and Scottish data (for this purpose the Dutch data was translated to English). After this data familiarisation process, we conducted the analysis together (involving multiple rounds of coding prior to theme generation) including extensive discussions around not only the data itself but also reflexively considering field notes and impressions from the data collection process, establishing internal validity. While philosophically saturation may not be attainable, the fact that we present data from two different contexts and countries suggests a robust dataset and findings which may be transferrable. As the language courses took place over many weeks and months, and the authors were the primary researchers in their respective projects, we ensured

prolonged engagement. Further, the findings reported here are triangulated across numerous data sources and associated contexts (see Appendix P). We employed an inductive approach to coding, using both semantic (explicit or overt) and latent (implicit, underlying) coding frameworks, in order to inform two overarching themes: “Navigating the digital highway” and “Camera ready for new friends”.

7.3 Results and discussion

Figure 7.3 below guides the reader by depicting a thematic map of the themes and their sub-themes and order in which they are discussed in the two sections below. Results are presented as an analytic narrative, in accordance with the Reflective TA analytic method adopted (Braun & Clarke, 2021c).

Figure 7.3 Thematic map of RTA themes



7.3.1 Navigating the digital highway

When teaching language to seniors online there are various technological and logistical aspects that need to be kept in mind. This theme focuses on exactly these aspects in our data. In general, the technology needs to be kept as simple as possible. This then translates into more specific actions. First of all, regarding the platform through which the course is taught, it is important that this is a platform that can be used immediately. For example, Zoom needs to be downloaded first (and using it from the browser is quite complicated and requires multiple clicks) whereas Google Meet can be used instantly from the browser. When a conferencing platform is used that needs to be installed first, in whatever form, seniors want to be helped: *"we're not*

good with computers, we need a lesson before the actual lesson on the technological issues" (Hendrika, 85); "*New things happen so quickly. [A] bit of an intro from the start to the technology for everyone is the class so we are all on the same page would have been nice*" (Isaac, 67). In fact, this is something learners brought up unprompted in both datasets, a lesson that just focuses on how to do all the technological things associated with the course (e.g., accessing the meeting and chat box, (un)muting audio, etc.). This also holds true for the homework, as discussed below. Additionally, when selecting a platform, the type of device learners are most likely to use to access the course should also be taken into consideration. For example, amongst our senior learners, many did not have a computer with a keyboard or mouse but rather a touchscreen tablet.

In accordance with van der Ploeg et al. (submitted) our data showed that learners liked doing homework, even online. However, similar to the video conferencing platform, this needs to be made simple. One learner mentioned that "*I'm not skilful or experienced on the digital highway and I'm contaminated with a fear of devices. So everything took a long time.*" (Hedwig, 80). Indeed, all seniors reported their homework taking them a long time. The most commonly mentioned problems are switching between tabs, the use of applications, websites, and playing media. The problems reflect not only a discomfort with being expected to quickly execute actions/tasks on a computer (e.g., switching between tabs or applications), but also a lack of understanding of how to do so. This is evidenced in the two quotes below:

> "*I thought this was an awkward way of doing homework. Switching between the article and assignments took a lot of time, but I also wouldn't know how it can be more efficient. Maybe there are possibilities to put the two pieces of paper next to each other on the PC but I haven't mastered that technique." (Sien, 86)*

"It took me 20 minutes to understand… Then it was an 'aha' moment. Learnt something new on the PC." (Truus, 69)

Moreover, when selecting the materials for the homework, acuity problems need to be taken into consideration ("A very positive experience, surprised myself. Only negative is to do with the streaming/hearing problems to do with the software and possibly my equipment" – Matilda, 67). Overall, it is of vital importance that the teacher/researcher is patient while taking seniors on their technology-travels: "I think it is important that you are patient with us as this is all new and we are not proficient in all these computer things. You were [patient] and it made me feel more relaxed and confident" (Renske, 71).

Across both datasets, half of the seniors that dropped out of the classes, did so due to finding the technology too difficult to overcome.

What becomes apparent, however, is that, even in the face of technical challenges, many senior learners enjoyed the online learning opportunity: *"I feel I've achieved something, having been challenged"* (Penelope, 63). As phrased by Aad (75): *"the atmosphere during the lessons was great, I enjoy the online group lessons even though I didn't expect this"*. Jude (63) adds to this: "*I was surprised how effective online class was versus in person. Online is also very convenient and eliminates travel time*."

7.3.2 Camera ready for new friends

A fundamental element of language learning classes is the interaction between people, both among learners and with the teacher. Online learning being more accessible than in person classes was a common point our learners mentioned. All learners in the Dutch study who did not live in Groningen mentioned, without prompting, that they appreciated the online format because they would not have been able to join the classes in person. As Daisy

(64), from the Scottish study, describes: *"The plus side of online is you can get people from all over. We had someone in Israel, a few people in France, and I am sitting in Scotland".* In some cases where learners lived in more remote areas, the offer of online classes meant they were able to participate where otherwise it would not be possible to commute: *"It was great to have people from lots of different places. If it had been in Edinburgh, I wouldn't have been able to do it"* (Phoebe, 54). Not only does an online learning environment allow for more people to join, but this diversity of backgrounds was also appreciated by the senior learners: *"[I would like to stay in touch with my classmates] feel like we were starting to get to know each other - particularly lovely as we all have such different backgrounds"* (Scarlett, 57). The social relationships developed throughout the language learning classes transferred outside the classroom as well: *"Could you tell me Aaf's last name? I want to send her a request on Facebook"* (Cornelia, 67) " *The online classes were great, better than most. I enjoyed the company. We have organised a WhatsApp group to stay in touch. I always prefer in person, but I am ok with online."* (Hugo, 55).

Learners identified turning on the video stream during class as a fundamental element to not only the social success and feeling of cohesion in the online learning environment, but also the overall learning success: *"... one of the things that I really enjoyed about the class was that I felt I was in a group of people... I certainly felt more engaged in the class because of that ... The thing that really stood out to me ... the class was really in there; they were really showing up for it and I think keeping the camera on was part of that."* (Poppy, 56). All learners in the Dutch study consistently kept their cameras on as well.

Indeed, the classes were often described favourably as providing structure to an otherwise structureless week in lockdown as well as an activity that provides a sense of achievement and purpose: *"...it's been fantastic to have*

something to put a structure on the week because knowing what day of the week it was passed by, and it was very difficult… just to have something that would give a sense of achievement was an amazing help" (Avery, 65). Jaap (72) adds to this, "*the course was a gift in times of COVID because it provided structure to my week and it allowed me to talk to people".* The importance of having a sense of purpose is well documented in terms of well-being, which has even been connected to lower risk of chronic health conditions (Kim et al., 2020; Ryff, 2014).

7.3.3 General discussion

From our data we see that, in accordance with the literature, an online language course for seniors can be feasible and even enjoyable (Djoub, 2013; C. Ware et al., 2017). The suggestions put forward by Djoub (2013) are supported by our reflexive thematic analysis: courses must use simple programs that do not require speedy processing. More concretely, platforms should incorporate both lessons in real-time and homework and should be compatible with tablets. Interestingly, to our knowledge, no such platform exists currently, and we highly recommend this topic for future studies. Furthermore, Djoub (2013) mentions that speedy processing should be avoided due to potential working memory problems. In our studies, however, we found that expectations of quick computer tasks should be avoided because seniors often lack enough experience on "the digital highway" for these types of tasks.

Crucially, our data showed that language learning has a very important social component. Third-age language learning, even online, is a "social endeavour" (Duay & Bryan, 2008) which can provide crucial structure, purpose, and transitional friendships which can last beyond the scheduled course. This

is in line with Djoub (2013) who suggests online courses should still enable learners to interact with peers. Our results suggest that this can be accomplished by simply chatting during the lessons and, after the course, by creating WhatsApp groups and allowing seniors to add each other on social media platforms. Finally, Djoub (2013) states that teachers should encourage learners to reflect on their learning process and, more specifically, relating to technology. Based on our multi-site findings, we suggest starting this as early as possible, preferably before the first lesson by including a 'technology lesson' explaining the technology used and empowering seniors to use it.

Our results are consistently found across both datasets, which were designed and collected independently in different contexts both in terms of location and language. This suggests that our findings are generalisable within the Western European context both in online applied language learning/teaching contexts and in online third-age language learning research. Further, while the data reported here focuses on online language learning classes specifically, language learning can be considered a proxy for other online education. Indeed, all of the five concrete suggestions reported below, might apply to more general online learning of other topics. However, further research is needed to confirm this.

From the data in our multi-site, observational, bottom-up study we can offer concrete recommendations for online (language) courses for seniors:

- Plan a lesson before the course to explain the technology. Additionally, instruction videos can also help;
- Show patience in explaining the technological aspects of the course and repeat these;
- Use low-threshold video conferencing which can be accessed via laptop or, importantly, tablets;

- Incorporate the homework in one, easy to use, platform. The video conferencing could also be integrated here;
- Special attention needs to be paid to ensure a social infrastructure as this can include chatting during the lessons, and helping learners stay in contact by means of WhatsApp groups or other social media platforms.

7.4 How to teach languages to older learners: insights from focus groups and a survey (work in progress)

While still a progress, and thus not reported as a completed empirical chapter as part of this thesis, a further study was designed and initiated to further explore language learning across the lifespan, focus specifically on the key differences between younger and older learners from the perspective of learners themselves and teachers. Following an operationalising sequential design (i.e., qualitative priority with a quantitative follow-up; see Section 2.5.1.1), the opinions and experiences of third age language learners and language teachers of adult mixed age classes were first gathered via focus groups. An international online survey was then developed following the analysis of the focus group data.

As discussed in the beginning of this chapter, leisure activities such as lifelong learning are increasingly popular after retirement. While the use of technology in language learning classes with older adults was explored above, the question remains of what teaching methods are best suited to address the opportunities and challenges in mixed-age classes. This work in progress explores the opinions and experiences of language learners and teachers from mixed-age classes, which in conjunction with the study presented at the beginning of this chapter around the use of technology builds a bridge from the more empirical language learning chapters into how these findings can be best applied in real life practice. From these perceptions and experiences,

courses can be better designed to improve the learning experiences of adults of all ages in the mixed-age language learning classroom.

7.4.1 Methods

Language teachers and third-age language learners were recruited from local language schools, The Centre for Open Learning at the University of Edinburgh and Yakety Yak Language Café, to participate in focus groups. The focus groups were conducted at a café at a local arts and events venue, rather than a university building in an effort to minimise participants feeling too formal or clinical. Between 2 October 2019 and 19 November 2019, a total of 7 focus groups were conducted: 2 with learners and 3 with teachers (Table 7.2). Prior to this, two pilot focus groups were conducted to validate the interview schedule. One of the initial pilot focus groups was conducted with both participant types, so a mix of learners and teachers. While this was successful, the researchers deemed it more appropriate to separate learners and teachers in subsequent focus groups in order to encourage more open dialogue.

Table 7.2 Breakdown of conducted focus groups

	Focus group	
	Learners	**Teachers**
Pilot 1	—	3
Pilot 2	6	1
Focus group 1	4	—
Focus group 2	—	2
Focus group 3	2	—
Focus group 4	—	3
Focus group 5	—	1
Total	12	10

Based on the themes identified in the focus groups, an online questionnaire was developed in January 2020. As with the focus groups, the questionnaire was split into two version: learners and teachers, each with 20 questions. Demographic information such as age, gender, occupation, language background, and languages learnt or taught was gathered in the first 10 questions. Questions 11-20 covered topics and themes identified from the focus groups including:

- which aspects of language learning are perceived to be particularly difficult,
- motivations for learning the target language,
- perceived differences in younger and older learners in the classroom,
- if/how students may be treated differently by teachers due to their age,
- difficulties in mixed age classes,
- preferences around class structure,
- preferences around learning resources/teacher adaptation of materials,
- factors in language learning success, and
- the composition of an ideal class in terms of learners' ages and gender, teacher's age and gender, learner ability in the target language, and teacher's native speaker status.

These questions were matched where possible across the leaner and teacher version to allow for a direct comparison of responses by respondent type. For example, question 14 for the leaners is "*Do you feel that teachers*

treat older learners differently to younger leaners?" while in the teacher version the question is phrased, "*Do you feel that you need to treat older learners differently from younger learners?*". Ethics was approved by the University of Edinburgh PPLS Research Ethics Committee (ref: 318-1819/1).

7.4.2 Preliminary results and discussion

As of February 2022, there are 567 completed survey responses, comprised of 395 learner responses and 172 teacher responses (Figure 7.4). The survey remains open as we aim to collect additional teacher responses to have more comparable group sizes. Preliminary results support findings from the focus groups and suggest that older learners (defined as learners aged 60+) are generally regarded to have more realistic expectations around their progress while younger adult learners (defined as learners aged 18 to 30) are more confident speaking. This holds true for both learners and teachers (Figure 7.5). Younger learners and teachers found younger learners to be more confident at speaking, but middle-aged learners and older learners largely reported no difference. Interestingly, all learners and teachers were more likely to report that they did not perceive for there to be a difference between age groups in the social benefit from language learning classes. Further, teachers reported in their experience younger students worry less about making mistakes, which was not reflected to the same degree in the student responses. Younger students were relatively split on younger students (43%) and older students (38%) worrying less about mistakes. Middle-aged learners were split across the three responses options and older learners were more likely to report no difference (43%).

While this study is still in progress, the preliminary results suggest that there are some areas of diverging views around mixed age language learning

classes between learners and teachers as well as within learners between different age groups. Once completed, it will be interesting to explore if these patterns remain and if there are differences depending on the target language, country and cultural context in which the respondent lives, years in which the respondent has been learning/teaching the target language, and on the delivery of the language learning class (i.e., online or in-person). As evidenced by the preceding chapters within this thesis, there is a gap in knowledge around best practices in the teaching of languages in older age and hence a need for further translational research into applied, practical questions such as motivation/attitudes, logistics (e.g., use of online learning), the choice of mixed or separated age classes, etc., which this study aims to address.

Figure 7.4 World map of survey respondents

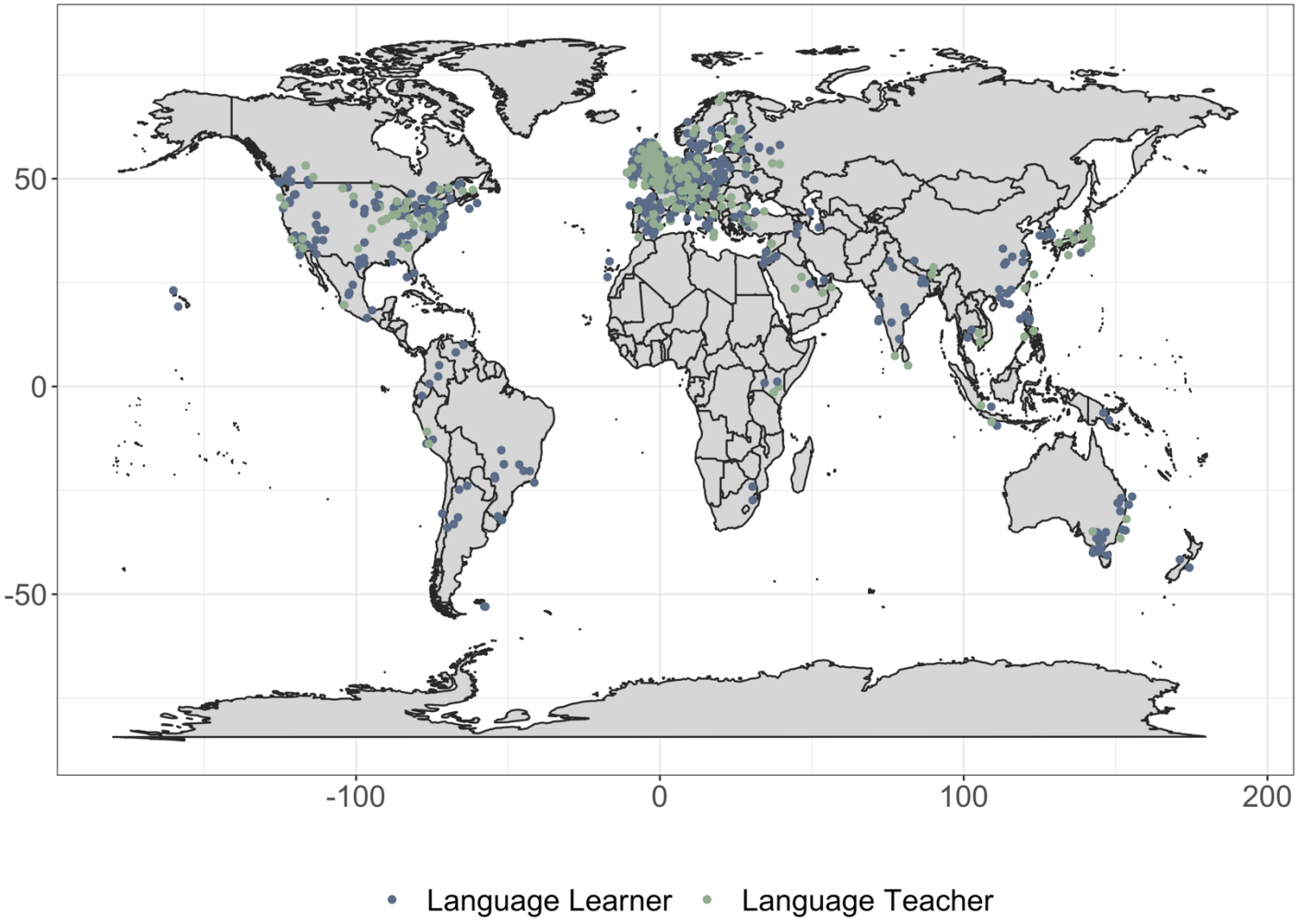
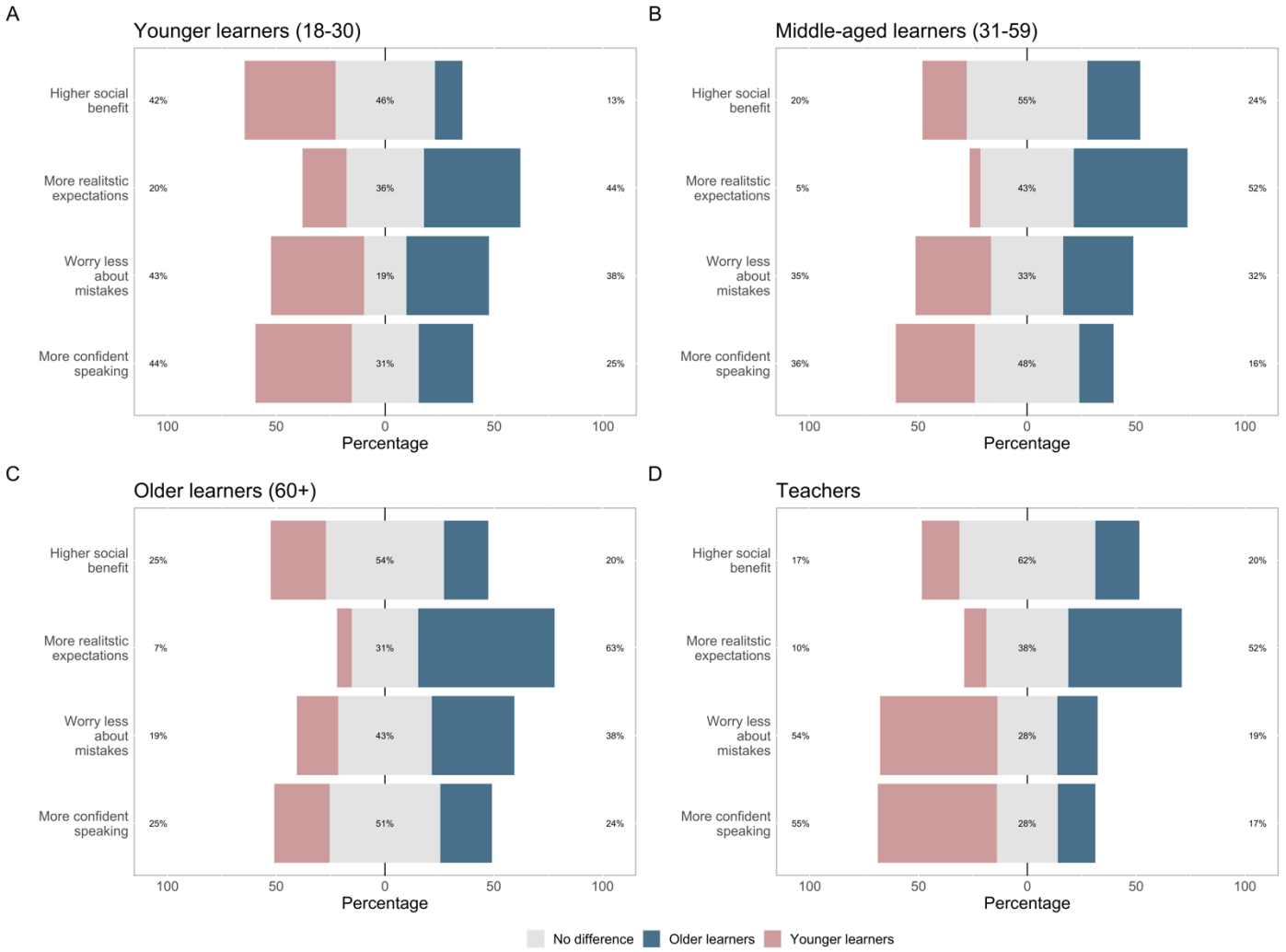


Figure 7.5 Perceptions of learner age differences in the mixed age classroom



Chapter 8 General Discussion

The aim of this thesis was to contribute novel empirical studies to the field of adult and third-age language learning and more generally the study of language and ageing. Taken together, Chapters 3 through 7 explore language across the lifespan through natural history (Chapter 3) and experimental approaches (Chapter 4, Chapter 5, and Chapter 6), concluding with practical applications (Chapter 7).

8.1 Natural history

Findings from Chapter 3, the study of autobiographical memory in conjunction with the preliminary results from the in-progress language changes in later life study, support a **dynamic nature of language across the lifespan** and the importance of considering individual differences. After establishing that language processing and interactions are not set in stone in childhood (e.g., are not static), the research focus moved to capitalising on this dynamic nature by asking, can adults across the lifespan learn a language in adulthood and what are the effects of these learning experiences cognitively and socially?

8.2 Cognitive effects of language learning

Language learning and training among adults across the lifespan was explored across the three studies presented in Chapters 4 to 6, which supported prior literature suggesting a cognitive effect of language learning among adults (e.g., Long et al., 2020; Meltzer et al., 2021; Pfenninger & Polz, 2018; Wong et al., 2019)**.** Crucially, these effects while overall positive, were not simple or uniform and seem to be moderated by certain key variables, in line

with the literature review conducted in Chapter 1. One of the factors which is likely to play a central role is the intensity and number of hours spent learning and practicing the language.

Collectively, the three language learning studies indicate a potential calibration between 4 and 6 hours a week in terms of the minimal **intensity and duration of language training** required to produce a measurable cognitive effect. Of the three studies presented, the most intense in terms of frequency (3 times a week) and duration (4 weeks) was the Spanish study. Accordingly, participants significantly improved from baseline on measures of inhibitory attention (ETD) and attentional switching (ETR), with moderate and large effect sizes, respectively. The second most intense course was the Scots study with the same frequency (3 times week) as the Spanish study, but a shorter duration of 3 weeks. Similar results were found with participants improving after the course from baseline measures of inhibitory attention, with a moderate effect size, and attentional switching, with a large effect size.

The Chinese study was the least intense in terms of frequency. Each block of classes was 2 times a week for 4 weeks, with 8 weeks across both courses. Notably, the findings were less robust with this intensity. After the first course participants did improve from baseline in measures of inhibitory attention and attentional switching, though the effect size was small in both cases. However, both courses taken together showed a significant improvement in attentional switching overall compared to baseline, with a large effect size. Similarly, after both courses (i.e., 32 hours total of language learning), participants significantly improved in the inhibitory attention measure, with a large effect size. After both classes at the lower intensity (i.e., 2 times a week for 8 weeks in total), the effect sizes were large, whereas at the increased intensity of 3 times a week in the Spanish and Scots studies, the

effect sizes were large after just 4 week and 3 weeks, respectively. Taken together, this suggests that lower frequency of classes across a longer duration of time may still produce cognitive effects, but more robust effects occur with higher frequency of classes and a moderate duration (4 weeks). Indeed, it is not surprising that a significant amount of exposure to language learning (i.e., practice) would be required to find short term measurable effects. Notably, these findings cannot be interpreted without the practicalities and feasibility of courses at such intensity. As discussed below in Section 8.3, the intensity and frequency of the courses was deemed palatable and feasible for participants as they knew it was only for a short period of time (e.g., for 3 or 4 weeks). Further, for many participants the extraordinary circumstances created by the CoViD-19 pandemic and subsequent lockdowns meant they had more time available than would be typically the case. Thus, while more intense and frequent language learning experiences may be better in terms of cognitive outcomes, this must be moderated by the practicalities of the duration of the course and accompanying feasibility for participants to dedicate their time to the course.

Furthermore, **time spent practicing the language** during the course is an interesting variable found to influence cognitive performance, which is worth future consideration. Time spent practicing the language significantly predicted both inhibitory attention and attentional switching, only the Spanish study. It was not a significant predictor of cognitive performance in either the Chinese study or Scots study. This finding indicates that the level of effort (as measured by time spent practicing) may be a differentiating factor in cognitive performance in higher intensity classes, such as in the Spanish study. This finding builds on Bak and colleagues (2016a) study which found that students

who the target language for at least 5 hours a week after completion of the course maintained their improvement in cognitive scores,

In the Scots study the status and prestige of the language and its linguistic distance to English was found to influence participant's willingness and comfort in speaking the language. As such, the learning in the classroom for some students was more passive, exposure based than in the other two studies. This is further supported by the fact that the most robust cognitive performance effects were found in the Scots Language Environment group, who more actively used the language. An area of future research to explore is if (a) time spent practicing is a reliable significant predictor only at higher intensities and (b) if the types of activities (i.e., more passive or more active practice of the language) participants engage with within and outwith the class have different effects.

Another avenue for future research is **attainment in the course and its relationship to cognitive outcomes.** While 2 of the 3 language learning studies within this thesis included quizzes at the end of the course, one of the limitations of these studies was the small number of students who completed the quizzes (or in the case of the Spanish study the data being lost by the teacher). Preliminary results, which should be interpreted with caution due to the reasons outlined above, suggest that attainment does not predict cognitive scores, which future studies should further investigate.

Moving to the question of age, across the language learning studies included in this thesis, participants ranged in age from 18 to 74 with an overall sample mean of 37.28 years old (*SD* = 14.4). While the main aim in recruitment for the studies presented was to include adults across the lifespan, the mean age of the overall participant pool is not directly comparable to the third age language learning literature which tends to include only older adults (see

Section 8.5 for a further discussion of limitations). However, as all three language learning studies included both younger, middle, and older adults the question of age as a continuous variable was able to be investigated. Overall, the older participants in each study performed worse on the cognitive measures of attention before the course (e.g., at baseline) compared to the younger adults, in line with previous studies using both the TEA and ANT (Bak et al., 2016a; Ishigami et al., 2016; Ishigami & Klein, 2011; Long et al., 2020; MacLeod et al., 2010). However, older adults did improve after the classes, showing similar patterns to younger participants. Furthermore, the background variables of gender and education level were non-significant predictors of cognitive performance across all three studies. This suggests that while all participants self-selected to participate in the courses, there was not a demographic variable (i.e., age, gender, education) which significantly differentiated the cognitive effects of the intensive language learning training.

Additionally, it is worth noting that dropouts across all the studies did not perform significantly worse at baseline measures of attention compared to the participants who completed the courses. This speaks against differences in cognitive performance playing a role in the decision to drop out of the courses. The primary reasons provided for dropping out of the courses were either (a) underestimating the intensity or unexpected work commitments or (b) technical issues in the initial online groups.

8.3 Social effects of language learning

The social aspect of language learning classes was seen to be beneficial to participants of various ages and backgrounds, in line with previous literature which would found language learning experiences to improve social well-being (Kliesch et al., 2018; Klimova et al., 2020; Pfenninger & Polz, 2018; Valis et al.,

2019). In particular, as shown in Figure 4.1, Figure 5.1, and Figure 6.1, online learning allows for a single class to include students from around the world, rather than only students who are able to commute to where the classes are being held. Therefore, online language learning classes in particular offer unique experiences to meet a diversity of people that you otherwise might not interact with or come across. This opportunity was highlighted as something of great value by many participants, as Rachel from the Chinese study describes "*I love the wide range of where people have come from, different walks of life etc. …the fact it's online has opened up who can take part so much more, different nationalities, different age groups, and I love that*." In particularly during the tumultuous time of lockdowns due to the CoViD-19 pandemic, the language classes were seen as a highly valuable resource. As Chloe from the Scots classes highlights, the classes were a good investment of time, not only due to the learning and language attainment, but also the social experience in an otherwise suddenly socially distanced world: "*You know, with the quarantine, it's, it's really a wise use of time, you know, learning and meeting people*."

Notably, however, many participants, particularly in the studies that were 3 times a week, highlighted that the intensity may not have been feasible, especially in person, during "normal life" with work and family commitments. Particularly relevant to future studies was the general agreement that the intensity was manageable because of the short duration (i.e., 3 times a week only for 3 week or 4 weeks). Even so, outwith the onset of the CoViD-19 pandemic, the intensity of the courses was primary reason for participants to drop out or withdraw from the study, be it underestimating the time commitment or unexpected work commitments arising. Intense and frequent

courses such as those presented here may be discouraging for recruitment and/or increase dropout rates over a longer duration of time.

In fact, the social element of the classes was identified by participants as a central factor in the desire to continuing to learn and practice the language. Many students described that in just 3 or 4 weeks the peer group formed a bond through the shared experience of the course. It is likely that the intensity of the courses, seeing the same group of people 2 or 3 times a week, encouraged social connections to form more quickly. The social aspect was not only a contributor to well-being, but also crucial to the likelihood of continue learning. In fact, some students reported that the social connections formed during the classes were also a contributing factor for their continued attendance during the study itself as they did not want to let down their peers or the teacher by missing a class. During the focus groups it was common for my final question asking for any other comments to be answered by students self-organising WhatsApp groups or Facebook groups to remain in touch after the study, if they had not done so already.

Furthermore, the qualitative data in conjunction with the quantitative cognitive data across all three language learning studies indicate a relevant interaction between intensity and enjoyment across all three studies. Namely, the intensity of the courses meant that participants did not forget what they had learned in the previous class and typically felt that as a result they attained more. Consequently, there was a common view that not only would the students like for the classes to continue as they enjoyed them, but also because they did not want to lose the progress that they felt they had made in learning the target language. For example, Poppy who participated in the Scots study shared *"Although the course was 3 x 3 weeks it was short (I would have liked to continue) but concentrated. This was beneficial as I didn't feel as though I'd*

forgotten anything from the previous class. Thoroughly enjoyable!". Therefore, as language attainment in more intensive classes is (or at least is perceived to be by students) higher, the students are thus more likely to continue learning, and in this way maintain the cognitive effects through continued practice. However, as discussed above, the intensity and frequency of the courses was feasible as participants were aware it was only for a short period of time.

Across all three studies, participants continued to learn the language, together, after the research study had completed. While there are no Scots classes offered at traditional language schools, in case of Spanish and Chinese it is comparatively easy to find resources for learning. In the Spanish and Chinese studies, the students continued to learn, in fact, within the same groups and with the same instructors as in the respective studies (in the Spanish study multiple classes hired the teacher as a private tutor and in the Chinese study peer groups continued to take courses at the Sheffield Confucius Institute). In the case of the Scots study, many participants self-organised to take the Open University course together as well, particularly from the Scots Language Environment group. Additionally, the friendships formed through these language learning experiences not only persisted through continued learning, but also outwith the classroom. It was an unexpected joy for me as a researcher when participants emailed me after the lockdown measures had eased with a picture of a group of them meeting up for lunch in person or a message of thanks again for the opportunity as they had made good friends from the experience of participating (which took place on multiple occasions).

Taken together, intensive classes may be a particularly effective introduction to learning an additional language for mixed age adult groups, which could be followed up by less intense and frequent classes in the longer term. While the intensity was a design element driven by the cognitive research

questions, it proved to be effective cognitively, socially, and in terms of feelings of attainment which in turn encouraged continued learning and engagement with the target language. The unintended additional variable of online delivery was generally perceived as opening doors for new social connections, particularly around the accessibility of classes, where students could attend from anywhere in the world, time zone and internet speeds permitting. Furthermore, as evidenced by Chapter 7, older adults did not perceive online learning to be exclusionary, though it does necessitate additional considerations (e.g., accessibility of the learning platform). Online learning should be seen not as a substitution for in-person classes, but a different genre with different considerations (e.g., smaller classes sizes, introduction to the learning interface, etc.).

8.4 Novel implementation and future considerations of the TEA

To my knowledge, the data collected in person as part of this thesis is among the first to employ the TEA in an **experimental group setting**. In a clinical application in which the measure was designed, group testing would not be appropriate. However, in research contexts group testing not only reduces the demand on resources (especially when there is only one researcher as was the case in the studies presented here), but also can be less demanding on participant time. During the in-person classes, I would ask for the students to arrive early to their first class and conduct the research session as a group. For the post-course research session, the students in attendance would stay after the class (and a coffee break) to participate. A by-product of this was that as the researcher, I could tangibly feel the social connection amongst the students in the second research session. Participants who were strangers in the

first research session were peers and friends by the second, often joking or even going for a coffee or to the pub after the study completed.

Online, however, group administration of the TEA was not feasible with varying internet speeds and technological set-ups. As described in the previous chapters, the transition to online research was not planned originally but rather made necessary by the unexpected and unpredictable CoViD-19 pandemics. It did include a few challenges, including discovering that Zoom suppressed background noises which meant the TEA could not be administered on the platform.

In terms of the specific subtests of the TEA employed throughout this thesis, the ET unsurprisingly produced ceiling effects across all studies. However, it is still a worthwhile measure to include as it is only 7 questions, though serves two important functions. First, as it is relatively straight forward, it can boost the confidence of participants who are wary of the task or getting questions incorrect. Second, after finding suitable platforms for the online research sessions, including the ET task proved to be a reliable means of testing participant internet connection or troubleshooting technical difficulties ahead of the more complex tasks with more tones.

Furthermore, future research including the TEA as a primary outcome measure, or indeed any cognitive tests generally speaking, should include baseline results as a stratum within the randomisation procedure if possible. While every effort was made in the study presented in Chapter 5 to prevent baseline differences between the randomised groups, the two groups did show a different, though statistically non-significant, pattern of baseline scores across the TEA and ANT.

8.5 Limitations and Strengths

The present research, however, is not without limitations, particularly regarding generalisability. A primary limitation is the fact that participants were recruited via convenience sampling and thus self-selected to participate. This self-selection bias limits the generalisability of the results found across this thesis. The move to online due to the CoViD-19 pandemic meant that the potential participant pool significantly widened from those able to commute to Edinburgh or Glasgow, increasing the geographical diversity of participants. However, unexpectedly opening a study to participants from various countries does not, in its own right, create generalisable findings. While a more representative sample than original envisaged was gathered, the recruited participants nonetheless largely lived in wealthy countries. Indeed, the online delivery of classes was limited in participant recruitment to those with the resources (e.g., computer, reliable internet), time, and interest in participating in such studies. Furthermore, although the pandemic and lockdowns resulted in many participants who, in their own words, would not have participated in similar studies "normally", the exceptional circumstances cannot be ignored. While the present work is limited by the self-selection bias of convenience sampling, the robust group sizes and number of participants included across the presented studies is a strength.

Additionally, the delineation and assessment of participants as multilingual by self-report is not entirely generalisable. It would have been ideal to assess knowledge of different languages more objectively, but given limited resources (e.g., myself as the only researcher collecting data, limited funding, etc.) minimising the testing battery time was prioritised.

Further limiting generalisability to the third-age language learning literature in particular is the age range of the recruited participants. As

described in Section 8.2, the mean age of the overall sample included in the language learning studies is 37.28 (*SD* = 14.4). As there are relatively low numbers of participants over 60 years of age in the present work, which is the typically the operationalisation of third age in the literature (see Chapter 1), results must be interpreted with caution as it relates to older age in particular. When relying on self-selection, it can be difficult to control variables such as the age of recruited participants. Given the limited resources available to conduct the present work and restricted funding timelines, an inclusion criterion of a particular age was deemed too strict. However, including participants across the lifespan within the same language learning classes is a unique aspect of the present work and one looked upon largely favourably by participants (see Section 8.3).

On the whole, the mixed-methods findings in the present work contribute to and support Pfenninger and Singleton's (2019) conclusion about language acquisition in the third age, here broadened to adults across the lifespan. The potential cognitive effects of language training and learning should not be seen as the only goal itself, but as a means of promoting well-being and social interaction across the lifespan.

8.6 Conclusions

The results of the studies presented within this thesis are timely given the Decade of Healthy Ageing has only just begun (World Health Assembly, 2020; World Health Organization, 2020a). WHO (2020a) outlines in their plan of action for the Decade that "healthy ageing requires life-long learning, enabling older people to do what they value, retain the ability to make decisions and preserve their purpose, identity and independence" (p. 7). The collective findings from this thesis work substantiates language learning across

the lifespan as an effective life-long learning activity contributing to healthy ageing by increasing the efficiency of attentional functions (particularly inhibitory attention and attentional switching), instilling confidence, providing purpose and structure (particularly during the CoViD-19 pandemic), promoting well-being, and broadening (often intergenerational) social connections among participants across the lifespan, all while having fun.

The diversity of research questions posed throughout this thesis required a diversity of methods to adequately address them. In order to do justice to this topic, group differences as well as individual experiences and perceptions must be taken into consideration, which necessitates a mixed methods investigation. It is only by using mixed methods in design and analysis that a wholistic picture is able to emerge, representative of both the group-level and participant-level effects and experiences.

Language and ageing are both complex phenomena, which require multifaceted methodological approaches in research. By adopting a mixed methods approach to questions of language and ageing, the studies presented here are able to not only successfully explore the research questions posed, but also result in clear avenues for future studies, as evidenced by the studies in progress reported in Sections 3.5 and 7.4. The first and last chapters, while answering several questions, open other lines of enquiry which will hopefully be addressed in the ongoing studies presented, as well as future work by other researchers.

References

Abutalebi, J., & Green, D. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics*, *20*(3), 242–275. https://doi.org/10.1016/j.jneuroling.2006.10.003

Afisi, O. T. (2013). Karl Popper's critical rationalism: Corroboration versus confirmation. *Philsophy Study*, *3*(6), 506–516. https://doi.org/10.17265/2159-5313/2013.06.007

Age UK. (2019). *Later Life in the United Kingdom 2019 Factsheet* (p. 27). Age UK. https://www.ageuk.org.uk/globalassets/age-uk/documents/reports-and-publications/later_life_uk_factsheet.pdf

Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health*, *86*(5), 726–728. https://doi.org/10.2105/AJPH.86.5.726

Alladi, S., Bak, T. H., Duggirala, V., Shailaja, M., Shukla, A. K., Chaudhuri, J. R., Kaul, S., & Surampudi, B. (2013). Bilingualism delays age at onset of dementia, independent of education and immigration status. *American Academy of Neurology*, *81*, 1–7.

Alladi, S., Bak, T. H., Mejala, S., Rahan, A., Chaudhuri, J. R., Mioshi, E., Krovvidi, R., Surampudi, B., Duggirala, V., & Kaul, S. (2016). Impact of

Bilingualism on Cognitive Outcome After Stroke. *Stroke*, *47*(1), 258–261.

Alladi, S., Bak, T. H., Shailaja, M., Gollahalli, D., Rajan, A., Surampudi, B., Hornberger, M., Duggirala, V., Chaudhuri, J. R., & Kaul, S. (2017). Bilingualism delays the onset of behavioral but not aphasic forms of frontotemporal dementia. *Neuropsychologia*, *99*, 207–212. https://doi.org/10.1016/j.neuropsychologia.2017.03.021

Allen, J. R. (2008). Why Learning To Write Chinese Is a Waste of Time: A Modest Proposal. *Foreign Language Annals*, *41*(2), 237–251. https://doi.org/10.1111/j.1944-9720.2008.tb03291.x

Alzheimer's Research UK. (2022, May 16). *Leading charity urges action as analysis shows dementia has been UK women's leading cause of death for a decade*. Alzheimersresearchuk.Org. https://www.alzheimersresearchuk.org/leading-charity-urges-action-as-analysis-shows-dementia-has-been-uk-womens-leading-cause-of-death-for-a-decade/

Alzheimer's Society. (2019, November). *What are the costs of dementia care in the UK?* Alzheimers.Org.Uk. https://www.alzheimers.org.uk/about-us/policy-and-influencing/dementia-scale-impact-numbers

Anderson, J. A. E., Hawrylewicz, K., & Grundy, J. G. (2020). Does bilingualism protect against dementia? A meta-analysis. *Psychonomic Bulletin & Review*, *27*(5), 952–965. https://doi.org/10.3758/s13423-020-01736-5

Antoniou, K., Grohmann, K. K., Kambanaros, M., & Katsos, N. (2016). The effect of childhood bilectalism and multilingualism on executive control. *Cognition*, *149*, 18–30. https://doi.org/10.1016/j.cognition.2015.12.002

Antoniou, M., Gunasekera, G. M., & Wong, P. C. M. (2013). Foreign language training as cognitive therapy for age-related cognitive decline: A hypothesis for future research. *Neuroscience & Biobehavioral Reviews*, *37*(10), 2689–2698. https://doi.org/10.1016/j.neubiorev.2013.09.004

Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using Zoom Videoconferencing for Qualitative Data Collection: Perceptions and Experiences of Researchers and Participants. *International Journal of Qualitative Methods*, *18*, 1–8. https://doi.org/10.1177/1609406919874596

Arons, A. M. M., Krabbe, P. F. M., van der Wilt, G. J., Olde Rikkert, M. G. M., & Adang, E. M. M. (2013). Visual analogue scales: Scale recalibration by patients with dementia and their proxies. *Quality of Life Research*, *22*(5), 979–986. https://doi.org/10.1007/s11136-012-0226-3

Aronson, J. (1995). A Pragmatic View of Thematic Analysis. *The Qualitative Report*. https://doi.org/10.46743/2160-3715/1995.2069

Bak, T. H. (2016). Cooking pasta in La Paz: Bilingualism, bias and the replication crisis. *Linguistic Approaches to Bilingualism*, *6*(5), 699–717. https://doi.org/10.1075/lab.16002.bak

Bak, T. H., & Alladi, S. (2016). Bilingualism, dementia and the tale of many variables: Why we need to move beyond the Western World. Commentary on Lawton et al. (2015) and Fuller-Thomson (2015). *Cortex*, *74*, 315–317. https://doi.org/10.1016/j.cortex.2015.04.025

Bak, T. H., Long, M. R., Vega-Mendoza, M., & Sorace, A. (2016a). Novelty, Challenge, and Practice: The Impact of Intensive Language Learning on Attentional Functions. *PLOS ONE*, *11*(4), e0153485. https://doi.org/10.1371/journal.pone.0153485

Bak, T. H., Long, M. R., Vega-Mendoza, M., & Sorace, A. (2016b). Novelty, Challenge, and Practice: The Impact of Intensive Language Learning on Attentional Functions. *PLOS ONE*, *11*(4), e0153485. https://doi.org/10.1371/journal.pone.0153485

Bak, T. H., Long, M., Vega-Mendoza, M., & Sorace, A. (2016c). Novelty, Challenge, and Practice: The Impact of Intensive Language Learning on Attentional Functions. *Plos One*, *11*(4), 1–11.

Bak, T. H., & Mehmedbegovic, D. (2017). *Healthy linguistic diet: The value of linguistic diversity and language learning across the lifespan*. https://doi.org/10.17863/CAM.9854

Bak, T. H., Nissan, J. J., Allerhand, M. M., & Deary, I. J. (2014). Does bilingualism influence cognitive aging?: Bilingualism and Aging. *Annals of Neurology*, *75*(6), 959–963. https://doi.org/10.1002/ana.24158

Bak, T. H., & Robertson, I. (2017). Biology enters the scene—A new perspective on bilingualism, cognition, and dementia. *Neurobiology of Aging*, *50*, iii–iv. https://doi.org/10.1016/j.neurobiolaging.2016.10.020

Bak, T. H., Vega-Mendoza, M., & Sorace, A. (2014). Never too late? An advantage on tests of auditory attention extends to late bilinguals. *Frontiers in Psychology*, *5*(485), 1–6.

Barnette, J. J. (2000). Effects of Stem and Likert Response Option Reversals on Survey Internal Consistency: If You Feel the Need, There is a Better Alternative to Using those Negatively Worded Stems. *Educational and Psychological Measurement*, *60*(3), 361–370. https://doi.org/10.1177/00131640021970592

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

Bartlett, D., & Payne, S. (1997). Grounded Theory—Its Basis, Rationale and Procedures. In G. W. McKenzie, J. Powell, & R. Usher (Eds.), *Understanding social research: Perspectives on methodology and practice* (pp. 167–189). Taylor & Francis e-Library. http://www.myilibrary.com?id=14453

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

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418. https://doi.org/10.1098/rstl.1763.0053

BBC. (2019, December 5). *'Gie it a shot'—OU offers free Scots language course*. Bbc.Co.Uk. https://www.bbc.co.uk/news/uk-scotland-50662854

BBC Two. (2018, November). *Can learning a new language boost your brain?* Bbc.Co.Uk. https://www.bbc.co.uk/programmes/articles/2NsHY4MhL95FgclByxS1XRl/can-learning-a-new-language-boost-your-brain

Beauvillain, C., & Grainger, J. (1987). Accessing interlexical homographs: Some limitations of a language-selective access. *Journal of Memory and*

Language, *26*(6), 658–672. https://doi.org/10.1016/0749-596X(87)90108-2

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507.

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Berggren, R., Nilsson, J., Brehmer, Y., Schmiedek, F., & Lövdén, M. (2020). Foreign language learning in older age does not improve memory or intelligence: Evidence from a randomized controlled study. *Psychology and Aging*, *35*(2), 212–219. https://doi.org/10.1037/pag0000439

Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G. (2003). Comparing the effects of text size and format on the readibility of computer-displayed Times New Roman and Arial text. *International Journal of Human-Computer Studies*, *59*(6), 823–835. https://doi.org/10.1016/S1071-5819(03)00121-6

Bhaskar, R. (1997). On the Ontological Status of Ideas. *Journal for the Theory of Social Behaviour*, *27*(2–3), 139–147. https://doi.org/10.1111/1468-5914.00031

Bhide, A., Shah, P. S., & Acharya, G. (2018). A simplified guide to randomized controlled trials. *Acta Obstetricia et Gynecologica Scandinavica*, *97*(4), 380–387. https://doi.org/10.1111/aogs.13309

Bialystok, E. (1999). Cognitive Complexity and Attentional Control in the Bilingual Mind. *Child Development*, *70*(3), 636–644. https://doi.org/10.1111/1467-8624.00046

Bialystok, E. (2010). Global–local and trail-making tasks by monolingual and bilingual children: Beyond inhibition. *Developmental Psychology*, *46*(1), 93–105. https://doi.org/10.1037/a0015466

Bialystok, E. (2011). Coordination of executive functions in monolingual and bilingual children. *Journal of Experimental Child Psychology*, *110*(3), 461–468. https://doi.org/10.1016/j.jecp.2011.05.005

Bialystok, E., Anderson, J. A. E., & Grundy, J. G. (2021). Interpreting cognitive decline in the face of cognitive reserve: Does bilingualism affect cognitive aging? *Linguistic Approaches to Bilingualism*, *11*(4), 484–504. https://doi.org/10.1075/lab.18040.bia

Bialystok, E., Barac, R., Blaye, A., & Poulin-Dubois, D. (2010). Word Mapping and Executive Functioning in Young Monolingual and Bilingual

Children. *Journal of Cognition and Development*, *11*(4), 485–508. https://doi.org/10.1080/15248372.2010.516420

Bialystok, E., Craik, F. I. M., & Freedman, M. (2007). Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia*, *45*(2), 459–464. https://doi.org/10.1016/j.neuropsychologia.2006.10.009

Bialystok, E., Craik, F. I. M., Klein, R., & Viswanathan, M. (2004). Bilingualism, Aging, and Cognitive Control: Evidence From the Simon Task. *Psychology and Aging*, *19*(2), 290–303. https://doi.org/10.1037/0882-7974.19.2.290

Bialystok, E., & Viswanathan, M. (2009). Components of executive control with advantages for bilingual children in two cultures. *Cognition*, *112*(3), 494–500. https://doi.org/10.1016/j.cognition.2009.06.014

Biau, D. J., Jolles, B. M., & Porcher, R. (2010). P Value and the Theory of Hypothesis Testing: An Explanation for New Researchers. *Clinical Orthopaedics & Related Research*, *468*(3), 885–892. https://doi.org/10.1007/s11999-009-1164-4

BMJ. (2022). Global, regional, and national burden of diseases and injuries for adults 70 years and older: Systematic analysis for the Global Burden of Disease 2019 Study. *BMJ*, e068208. https://doi.org/10.1136/bmj-2021-068208

Bonfiglio, T. P. (2002). The legitimation of accent. In M. Heller & R. J. Watts (Eds.), *Race and the Rise of Standard American:* (Vol. 7, pp. 9–75). DE GRUYTER MOUTON. https://doi.org/10.1515/9783110851991

Bosisio, N. (2019). Language Learning in the Third Age. *Geopolitical, Social Security and Freedom Journal*, *2*(1), 21–36. https://doi.org/10.2478/gssfj-2019-0003

Boulton-Lewis, G. M., & Tam, M. (2018). Issues in teaching and learning for older adults in Hong Kong and Australia. *Educational Gerontology*, *44*(10), 639–647. https://doi.org/10.1080/03601277.2018.1521902

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage Publications.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. SAGE.

Braun, V., & Clarke, V. (2021a). *Thematic analysis: A practical guide to understanding and doing* (1st ed.). SAGE Publications.

Braun, V., & Clarke, V. (2021b). Can I use TA? Should I use TA? Should I *not* use TA? Comparing reflexive thematic analysis and other pattern-

based qualitative analytic approaches. *Counselling and Psychotherapy Research*, *21*(1), 37–47. https://doi.org/10.1002/capr.12360

Braun, V., & Clarke, V. (2021c). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Brewer, J., & Hunter, A. (1993). *Multimethod research: A synthesis of styles* (Nachdr.). Sage.

Brewer, W. F. (1996). What is recollective memory? In D. C. Rubin (Ed.), *Remembering our Past: Studies in Autobiographical Memory* (1st ed., pp. 19–66). Cambridge University Press. https://doi.org/10.1017/CBO9780511527913.002

Brislin, R. W. (1970). Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Brooks, L. (2020, January 2). *Duolingo sparks Gaelic boom as young Scots shrug off 'cringe' factor*. The Guardian. https://www.theguardian.com/education/2020/jan/02/duolingo-sparks-gaelic-boom-as-young-scots-shrug-off-cringe-factor

Brooks, L., & Hern, A. (2020, August 26). *Shock an aw: US teenager wrote huge slice of Scots Wikipedia* [News]. The Guardian.

https://www.theguardian.com/uk-news/2020/aug/26/shock-an-aw-us-teenager-wrote-huge-slice-of-scots-wikipedia

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research*, *6*(1), 97–113. https://doi.org/10.1177/1468794106058877

Bubbico, G., Chiacchiaretta, P., Parenti, M., di Marco, M., Panara, V., Sepede, G., Ferretti, A., & Perrucci, M. G. (2019). Effects of Second Language Learning on the Plastic Aging Brain: Functional Connectivity, Cognitive Decline, and Reorganization. *Frontiers in Neuroscience*, *13*, 423. https://doi.org/10.3389/fnins.2019.00423

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, *56*(3), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

Cacioppo, J. T., Fowler, J. H., & Christakis, N. A. (2009). Alone in the crowd: The structure and spread of loneliness in a large social network. *Journal of Personality and Social Psychology*, *97*(6), 977–991. https://doi.org/10.1037/a0016076

Cady, E. T., Harris, R. J., & Knappenberger, J. B. (2008). Using music to cue autobiographical memories of different lifetime periods. *Psychology of Music*, *36*(2), 157–177. https://doi.org/10.1177/0305735607085010

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Care Connect, & Age UK. (2018). *Loneliness and isolation – understanding the difference and why it matters*. Care Connect, Age UK. https://www.ageuk.org.uk/globalassets/age-uk/documents/reports-and-publications/reports-and-briefings/loneliness/rb_feb2018_180208_careconnect_ageuk_loneliness_research_article_isolation.pdf

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, *19*(9), 1141–1164. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F

Carr, D., & Bodnar-Deren, S. (2009). Gender, Aging and Widowhood. In P. Uhlenberg (Ed.), *International Handbook of Population Aging* (pp. 705–728). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8356-3_32

Cattan, M., White, M., Bond, J., & Learmouth, A. (2005). Preventing social isolation and loneliness among older people: A systematic review of health promotion interventions. *Ageing and Society*, *25*(01), 41–67. https://doi.org/10.1017/S0144686X04002594

Center for Open Learning. (n.d.). *Languages*. https://www.ed.ac.uk/studying/short-courses/subjects/languages-courses

Central Intelligence Agency. (n.d.). *The World Fact Book: Languages* [Governmental]. Cia.Gov. https://www.cia.gov/the-world-factbook/field/languages/

Chan, J. C. (1991). Response-Order Effects in Likert-Type Scales. *Educational and Psychological Measurement*, *51*(3), 531–540. https://doi.org/10.1177/0013164491513002

Charmaz, K. (1990). 'Discovering' chronic illness: Using grounded theory. *Social Science & Medicine*, *30*(11), 1161–1172. https://doi.org/10.1016/0277-9536(90)90256-R

Charmaz, K. (2015). Grounded theory. In J. A. Smith (Ed.), *Qualitative psychology: A practical guide to research methods* (3rd edition, pp. 53–85). SAGE.

Chen, S.-Y., Feng, Z., & Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *Journal of Thoracic Disease*, *9*(6), 1725–1729. https://doi.org/10.21037/jtd.2017.05.34

Christ, T. (1970). A Thematic Analysis of the American Business Creed. *Social Forces*, *49*(2), 239. https://doi.org/10.2307/2576523

Christensen, R. H. B. (2019). *ordinal—Regression Models for Ordinal Data* (R package version 2019.12-10). https://CRAN.R-project.org/package=ordinal

Chu, S., & Downes, J. J. (2002). Proust nose best: Odors are better cues of autobiographical memory. *Memory & Cognition*, *30*(4), 511–518. https://doi.org/10.3758/BF03194952

Chyung, S. Y. Y., Swanson, I., Roberts, K., & Hankinson, A. (2018). Evidence-Based Survey Design: The Use of Continuous Rating Scales in Surveys. *Performance Improvement*, *57*(5), 38–48. https://doi.org/10.1002/pfi.21763

Clarke, A. E. (2003). Situational Analyses: Grounded Theory Mapping After the Postmodern Turn. *Symbolic Interaction*, *26*(4), 553–576. https://doi.org/10.1525/si.2003.26.4.553

Clarke, V., & Braun, V. (2016). Thematic Analysis. In E. Lyons & A. Coyle (Eds.), *Analysing qualitative data in psychology* (2nd edition, pp. 84–103). SAGE.

Coffey, A. J., & Atkinson, P. A. (1996). *Making sense of qualitative data: Complementary research strategies*. Sage Publications.

Colomé, À. (2001). Lexical Activation in Bilinguals' Speech Production: Language-Specific or Language-Independent? *Journal of Memory and Language*, *45*(4), 721–736. https://doi.org/10.1006/jmla.2001.2793

Costa, A. (2005). Lexical Access in Bilingual Production. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 308–325). Oxford University Press.

Costa, A., Calabria, M., Marne, P., Hernández, M., Juncadella, M., Gascón-Bayarri, J., Lleó, A., Ortiz-Gil, J., Ugas, L., Blesa, R., & Reñé, R. (2012). On the parallel deterioration of lexico-semantic processes in the bilinguals' two languages: Evidence from Alzheimer's disease. *Neuropsychologia*, *50*(5), 740–753. https://doi.org/10.1016/j.neuropsychologia.2012.01.008

Costa, A., Hernández, M., Costa-Faidella, J., & Sebastián-Gallés, N. (2009). On the bilingual advantage in conflict processing: Now you see it, now you don't. *Cognition*, *113*(2), 135–149. https://doi.org/10.1016/j.cognition.2009.08.001

Costa, A., Hernández, M., & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, *106*(1), 59–86. https://doi.org/10.1016/j.cognition.2006.12.013

Costa, J. (2015). Can Schools Dispense with Standard Language? Some Unintended Consequences of Introducing Scots in a Scottish Primary School: Can Schools Dispense with Standard Language. *Journal of Linguistic Anthropology*, *25*(1), 25–42. https://doi.org/10.1111/jola.12069

Cox, J. G. (2017). EXPLICIT INSTRUCTION, BILINGUALISM, AND THE OLDER ADULT LEARNER. *Studies in Second Language Acquisition*, *39*(1), 29–58. https://doi.org/10.1017/S0272263115000364

Cox, J. G., & Zlupko, G. M. (2019). Individual differences in language experience and bilingual autobiographical memory. *International Journal of Bilingualism*, *23*(5), 1180–1194. https://doi.org/10.1177/1367006917728394

Cox, S. R., Bak, T. H., Allerhand, M., Redmond, P., Starr, J. M., Deary, I. J., & MacPherson, S. E. (2016). Bilingualism, social cognition and executive functions: A tale of chickens and eggs. *Neuropsychologia*, *91*, 299–306.

Coyle, A. (2007). Introduction to Qualitative Psychological Research. In E. Lyons & A. Coyle, *Analysing Qualitative Data in Psychology* (pp. 9–30). SAGE Publications, Ltd. https://doi.org/10.4135/9781446207536.d7

Coyle, A. (2016a). Discourse Analysis. In E. Lyons & A. Coyle (Eds.), *Analysing qualitative data in psychology* (2nd edition, pp. 160–181). SAGE.

Coyle, A. (2016b). Introduction to Qualitative Psychological Research. In E. Lyons & A. Coyle (Eds.), *Analysing qualitative data in psychology* (2nd edition, pp. 9–30). SAGE.

Creath, R. (2021). Logical Empiricism. *The Stanford Encyclopedia of Philosophy*, *Winter 2021*. https://plato.stanford.edu/archives/win2021/entries/logical-empiricism/

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed). SAGE Publications.

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (Third Edition (International Student Edition)). SAGE Publications, Inc. https://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9781483347011

Crovitz, H. F., & Quina-Holland, K. (1976). Proportion of episodic memories from early childhood by years of age. *Bulletin of the Psychonomic Society*, *7*(1), 61–62. https://doi.org/10.3758/BF03337122

Crovitz, H. F., & Schiffman, H. (1974). Frequency of episodic memories as a function of their age. *Bulletin of the Psychonomic Society*, *4*(5), 517–518. https://doi.org/10.3758/BF03334277

Curtin, F., & Schulz, P. (1998). Multiple correlations and bonferroni's correction. *Biological Psychiatry*, *44*(8), 775–777. https://doi.org/10.1016/S0006-3223(98)00043-2

Darcy, N. T. (1963). Bilingualism and the Measurement of Intelligence: Review of a Decade of Research. *The Journal of Genetic Psychology*, *103*(2), 259–282. https://doi.org/10.1080/00221325.1963.10532521

de Bot, K. (2010). Cognitive Processing in Bilinguals: From Static to Dynamic Models. In R. B. Kaplan (Ed.), *The Oxford Handbook of Applied Linguistics* (2nd ed., pp. 335–348). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195384253.013.0023

de Bot, K. (2012). Rethinking multilingual processing: From a static to a dynamic approach. In J. Cabrelli Amaro, S. Flynn, & J. Rothman (Eds.), *Third language acquisition in adulthood* (Vol. 46, pp. 79–95). John Benjamins Pub. Co.

de Bot, K., & Clyne, M. (1989). Language Reversion Revisited. *Studies in Second Language Acquisition*, *11*(2), 167–177. https://doi.org/10.1017/S0272263100000590

de Bot, K., & Weltens, B. (1995). Foreign Language Attrition. *Annual Review of Applied Linguistics*, *15*, 151–164. https://doi.org/10.1017/S026719050000266X

de Bruin, A., Treccani, B., & Della Sala, S. (2015). Cognitive Advantage in Bilingualism: An Example of Publication Bias? *Psychological Science*, *26*(1), 99–107. https://doi.org/10.1177/0956797614557866

de Groot, A. M. B., Delmaar, P., & Lupker, S. J. (2000). The Processing of Interlexical Homographs in Translation Recognition and Lexical Decision: Support for Non-Selective Access to Bilingual Memory. *The Quarterly Journal of Experimental Psychology Section A*, *53*(2), 397–428. https://doi.org/10.1080/713755891

de Jong-Gierveld, J., & Kamphuls, F. (1985). The Development of a Rasch-Type Loneliness Scale. *Applied Psychological Measurement*, *9*(3), 289–299. https://doi.org/10.1177/014662168500900307

de Jong-Gierveld, J., Tilburg, T. G. van, & Dykstra, P. A. (2018). New Ways of Theorizing and Conducting Research in the Field of Loneliness and Social Isolation. In A. L. Vangelisti & D. Perlman (Eds.), *The Cambridge Handbook of Personal Relationships* (2nd ed., pp. 391–404). Cambridge University Press. https://doi.org/10.1017/9781316417867.031

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Declerck, M., Özbakar, E., & Kirk, N. W. (2021). Is there proactive inhibitory control during bilingual and bidialectal language production? *PLOS ONE*, *16*(9), e0257355. https://doi.org/10.1371/journal.pone.0257355

Denzin, N. K. (1970). *The Research Act: A Theoretical Introduction to Sociological Methods* (1st ed.). Aldine. https://doi.org/10.4324/9781315134543

Denzin, N. K. (1989). *The research act: A theoretical introduction to sociological methods* (3rd ed). Prentice Hall.

Denzin, N. K., & Lincoln, Y. S. (1994). Introduction: Entering the Field of Qualitative Research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 1–17). Sage Publications.

Department of Work and Pensions. (2011). *Number of Future Centenarians by Age Group*. Deparment of Work and Pensions. https://www.gov.uk/government/statistics/future-centenarians-by-age-group

Derksen, M. (2019). Putting Popper to work. *Theory & Psychology*, *29*(4), 449–465. https://doi.org/10.1177/0959354319838343

Dew, J. E. (2005, August). *Language is primary, script is secondary: The importance of gaining a strong foundation in the language before devoting major efforts to character recognition.* International and

Interdisciplinary Conference, University of Mainz, Germersheim, Germany.

Diamond, M. C., Krech, D., & Rosenzweig, M. R. (1964). The effects of an enriched environment on the histology of the rat cerebral cortex. *The Journal of Comparative Neurology*, *123*(1), 111–119. https://doi.org/10.1002/cne.901230110

Dijkstra, T., Van Heuven, W. J. B., & Grainger, J. (1998). Simulating Cross-Language Competition with the Bilingual Interactive Activation Model. *Psychologica Belgica*, *38*(3–4), 177. https://doi.org/10.5334/pb.933

Djoub, Z. (2013). ICT education and motivating elderly people. *Ariadna*, *1*, 88–92. https://doi.org/10.6035/Ariadna.2013.1.15

Duay, D. L., & Bryan, V. C. (2008). Learning in Later Life: What Seniors Want in a Learning Experience. *Educational Gerontology*, *34*(12), 1070–1086. https://doi.org/10.1080/03601270802290177

Dubin, R. (1969). *Theory building*. The Free Press.

Dwan, K., Li, T., Altman, D. G., & Elbourne, D. (2019). CONSORT 2010 statement: Extension to randomised crossover trials. *BMJ*, l4378. https://doi.org/10.1136/bmj.l4378

Edwards, A. D. (1976). *Language in culture and class: The sociology of language and education*. Heinemann Educational.

El Haj, M., Kapogiannis, D., & Antoine, P. (2020). The picture of the past: Pictures to cue autobiographical memory in Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, *42*(9), 914–923. https://doi.org/10.1080/13803395.2020.1825636

Emery, C. (2022a, March 22). *The 27 Most Spoken Languages in the World (2021)*. Langoly: The Ultimate Language Resource. https://www.langoly.com/most-spoken-languages/

Emery, C. (2022b, April 11). *34 of the Most Spoken Languages in Europe (2022): Facts and Figures*. Langoly: The Ultimate Language Resource. https://www.langoly.com/most-spoken-languages-in-europe/

E-Prime (2.0). (2012). Psychology Software Tools, Inc. https://support.pstnet.com

Esposito, A. G., & Baker-Ward, L. (2016). Immigration, language proficiency, and autobiographical memories: Lifespan distribution and second-language access. *Memory*, *24*(7), 949–960. https://doi.org/10.1080/09658211.2015.1061010

Estanga, A., Ecay-Torres, M., Ibañez, A., Izagirre, A., Villanua, J., Garcia-Sebastian, M., Iglesias Gaspar, M. T., Otaegui-Arrazola, A., Iriondo, A., Clerigue, M., & Martinez-Lage, P. (2017). Beneficial effect of bilingualism on Alzheimer's disease CSF biomarkers and cognition.

Neurobiology of Aging, *50*, 144–151. https://doi.org/10.1016/j.neurobiolaging.2016.10.013

Ethnologue: Languages of the World. (2022). In D. M. Eberhard, G. F. Simons, & C. D. Fennig (Eds.), *Ethnologue: Languages of the World.* (25th ed.). SIL International. https://www.ethnologue.com/

Evans, A. S., & Preston, A. S. (2011). Test of Everyday Attention. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (pp. 2491–2492). Springer New York. https://doi.org/10.1007/978-0-387-79948-3_1298

Evans, L. (2011). *Endangered languages: The full list*. The Guardian. https://www.theguardian.com/news/datablog/2011/apr/15/language-extinct-endangered

Evans, N. (2018). Did language evolve in multilingual settings? *Biology & Philosophy*, *32*(6), 905–933. https://doi.org/10.1007/s10539-018-9609-3

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the Efficiency and Independence of Attentional Networks. *Journal of Cognitive Neuroscience*, *14*(3), 340–347. https://doi.org/10.1162/089892902317361886

Fernandez-Duque, D., & Black, S. E. (2006). Attentional networks in normal aging and Alzheimer's disease. *Neuropsychology*, *20*(2), 133–143. https://doi.org/10.1037/0894-4105.20.2.133

Fetters, M. D., & Molina-Azorin, J. F. (2017). The *Journal of Mixed Methods Research* Starts a New Decade: Principles for Bringing in the New and Divesting of the Old Language of the Field. *Journal of Mixed Methods Research*, *11*(1), 3–10. https://doi.org/10.1177/1558689816682092

Fidler, F., Singleton Thorn, F., Barnett, A., Kambouris, S., & Kruger, A. (2018). The Epistemic Importance of Establishing the Absence of an Effect. *Advances in Methods and Practices in Psychological Science*, *1*(2), 237–244. https://doi.org/10.1177/2515245918770407

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics: And sex and drugs and rock 'n' roll* (4th edition). Sage.

Finniss, D. G., Kaptchuk, T. J., Miller, F., & Benedetti, F. (2010). Biological, clinical, and ethical advances of placebo effects. *The Lancet*, *375*(9715), 686–695. https://doi.org/10.1016/S0140-6736(09)61706-2

Fisher, R. A. (1921). On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample. *Metron*, *1*, 3–32.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

Fisher, R. A. (1950). *Statistical Methods for Research Workers* (11th edition). Oliver & Boyd.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver & Boyd.

Fishman, J. A. (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*. Multilingual Matters.

Flynn, C. J. (2020). *Adult Minority Language Learning*. Multilingual Matters. https://doi.org/10.21832/FLYNN7031

Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J. P. B., & Price, G. J. (2022). Understanding the Differences Between Bayesian and Frequentist Statistics. *International Journal of Radiation Oncology*Biology*Physics*, *112*(5), 1076–1082. https://doi.org/10.1016/j.ijrobp.2021.12.011

Frank, A. W. (2010). *Letting stories breathe: A socio-narratology*. University of Chicago Press. http://site.ebrary.com/id/10431295

Franke, M. (2021). *An Introduction to Data Analysis*. https://michael-franke.github.io/intro-data-analysis/index.html

Fratiglioni, L., Wang, H.-X., Ericsson, K., Maytan, M., & Winblad, B. (2000). Influence of social network on occurrence of dementia: A community-based longitudinal study. *The Lancet*, *355*(9212), 1315–1319. https://doi.org/10.1016/S0140-6736(00)02113-9

Freyd, M. (1923). The Graphic Rating Scale. *Journal of Educational Psychology*, *14*(2), 83–102. https://doi.org/10.1037/h0074329

Friedman, H., Herskovitz, P. J., & Pollack, S. (1993). Biasing effects of scale-checking styles on response to a Likert scale. *Section on Survey Research Methods: Volume 2*, *2792–795*, 792–794. http://www.amstat.org/sections/srms/proceedings

Fugard, A., & Potts, H. W. W. (2020). Thematic Analysis. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. A. Williams (Eds.), *SAGE Research Methods Foundations*. SAGE Publications Ltd. https://doi.org/10.4135/9781526421036858333

Funke, F. (2016). A Web Experiment Showing Negative Effects of Slider Scales Compared to Visual Analogue Scales and Radio Button Scales. *Social Science Computer Review*, *34*(2), 244–254. https://doi.org/10.1177/0894439315575477

Funke, F., & Reips, U.-D. (2012). Why Semantic Differentials in Web-Based Research Should Be Made from Visual Analogue Scales and Not from 5-Point Scales. *Field Methods*, *24*(3), 310–327. https://doi.org/10.1177/1525822X12444061

Galton, F. (1879a). Psychometric Experiments. *Brain*, *2*, 148–162.

Galton, F. (1879b). Psychometric Facts. *The Nineteenth Century*, *5*, 425–433.

Garcia, V. R. M. (2017). Learning English in the elderly: An analysis of motivational factors and language learning strategies. *BELT - Brazilian English Language Teaching Journal*, *8*(2), 234. https://doi.org/10.15448/2178-3640.2017.2.28264

Gardiner, K. (2016, January 26). *Is it the end of the Scottish Cringe?* Bbc.Com. https://www.bbc.com/culture/article/20160114-is-it-the-end-of-the-scottish-cringe

Gibbs, G. (2007). Thematic Coding and Categorizing. In *Analyzing Qualitative Data* (pp. 38–56). SAGE Publications, Ltd. https://doi.org/10.4135/9781849208574

Giddings, L. S. (2006). Mixed-methods research: Positivism dressed in drag? *Journal of Research in Nursing*, *11*(3), 195–203. https://doi.org/10.1177/1744987106064635

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gilljam, M., & Granberg, D. (1993). Should We Take Don't Know for an Answer? *The Public Opinion Quarterly*, *57*(3), 348–357. http://www.jstor.org/stable/2749095

Gillund, G. (2012). Episodic Memory. In *Encyclopedia of Human Behavior* (pp. 68–72). Elsevier. https://doi.org/10.1016/B978-0-12-375000-6.00152-X

Glaser, B. G. (1992). *Basics of grounded theory analysis: Emergence vs. forcing*. Sociology Press.

Glaser, B. G., & Strauss, A. L. (1965). *Awareness of dying*. Aldine Publishing Company.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research* (New York). Aldine Transaction.

Godfrey-Smith, P. (2009). *Theory and reality: An introduction to the philosophy of science*. University of Chicago Press.

Gómez, R. (2008). *Aplicabilidad del test de la afasia para bilingües de Michel Paradis a la población catalano/castellano parlante* [Universitat de Barcelona]. http://www.tdx.cat/TDX-0128109-111314/

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(2), 67–81. https://doi.org/10.1017/S1366728998000133

Green, D. W., & Abutalebi, J. (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology*, *25*(5), 515–530. https://doi.org/10.1080/20445911.2013.796377

Green, J., & Thorogood, N. (2014). *Qualitative methods for health research*. http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9781446296158

Greenberg, D. L., Ogar, J. M., Viskontas, I. V., Gorno Tempini, M. L., Miller, B., & Knowlton, B. J. (2011). Multimodal cuing of autobiographical memory in semantic dementia. *Neuropsychology*, *25*(1), 98–104. https://doi.org/10.1037/a0021005

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a Conceptual Framework for Mixed-Method Evaluation Designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255–274. https://doi.org/10.3102/01623737011003255

Greene, J. D. W., Hodges, J. R., & Baddeley, A. D. (1995). Autobiographical memory and executive function in early dementia of Alzheimer type. *Neuropsychologia*, *33*(12), 1647–1670. https://doi.org/10.1016/0028-3932(95)00046-1

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Grosjean, F. (1984). *Life with two languages: An introduction to bilingualism* (11. print). Harvard Univ. Press.

Grossman, J., & Mackenzie, F. J. (2005). The Randomized Controlled Trial: Gold standard, or merely standard? *Perspectives in Biology and Medicine*, *48*(4), 516–534. https://doi.org/10.1353/pbm.2005.0092

Grundy, J. G. (2020). The effects of bilingualism on executive functions: An updated quantitative analysis. *Journal of Cultural Cognitive Science*, *4*(2), 177–199. https://doi.org/10.1007/s41809-020-00062-5

Guardian Carers—Who are we. (n.d.). Guardian.Carers.Co.Uk. https://guardiancarers.co.uk/who-we-are

Guba, E. G., & Lincoln, Y. S. (1994). Competing Paradigms in Qualitative Research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 105–117). Sage Publications.

Guest, G., MacQueen, K. M., & Namey, E. E. (2012). *Applied Thematic Analysis*. SAGE Publications, Inc. https://doi.org/10.4135/9781483384436

Hackett, R. A., Steptoe, A., Cadar, D., & Fancourt, D. (2019). Social engagement before and after dementia diagnosis in the English Longitudinal Study of Ageing. *PLOS ONE*, *14*(8), e0220195. https://doi.org/10.1371/journal.pone.0220195

Haig, B. D. (2003). What Is a Spurious Correlation? *Understanding Statistics*, *2*(2), 125–132. https://doi.org/10.1207/S15328031US0202_03

Hall, J., O'Carroll, R. E., & Frith, C. D. (2010). Neuropsychology. In *Companion to Psychiatric Studies* (8th ed., pp. 121–140). Elsevier. https://doi.org/10.1016/B978-0-7020-3137-3.00007-3

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG: An International Journal of Obstetrics & Gynaecology*, *125*(13), 1716–1716. https://doi.org/10.1111/1471-0528.15199

Hawthorne, G. (2006). Measuring Social Isolation in Older Adults: Development and Initial Validation of the Friendship Scale. *Social Indicators Research*, *77*(3), 521–548. https://doi.org/10.1007/s11205-005-7746-y

Hayakawa, S., & Keysar, B. (2018). Using a foreign language reduces mental imagery. *Cognition*, *173*, 8–15. https://doi.org/10.1016/j.cognition.2017.12.010

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*, *18*, 98–99.

Heller, G. Z., Manuguerra, M., & Chow, R. (2016). How to analyze the Visual Analogue Scale: Myths, truths and clinical relevance. *Scandinavian Journal of Pain*, *13*(1), 67–75. https://doi.org/10.1016/j.sjpain.2016.06.012

Henwood, K., & Pidgeon, N. (1994). Beyond the qualitative paradigm: A framework for introducing diversity within qualitative psychology. *Journal of Community & Applied Social Psychology*, *4*(4), 225–238. https://doi.org/10.1002/casp.2450040403

Hermans, S. (2012). *Attrition and Language Reversion in Dutch-English Migrants in Australia: Working Memory Capacity and Inhibitory Functioning across Languages and Age Groups* [Master's]. Utrecht Univeristy.

Hesse-Biber, S. (2015). Mixed Methods Research: The "Thing-ness" Problem. *Qualitative Health Research*, *25*(6), 775–788. https://doi.org/10.1177/1049732315580558

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. http://www.jstor.org/stable/4615733

Holt, N. J., Bremner, A., Sutherland, E., Vliek, M. L. W., Passer, M. W., & Smith, R. E. (2019). *Psychology: The science of mind and behaviour* (Fourth edition). McGraw-Hill Education.

Hox, J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel Analysis: Techniques and Application* (3rd ed.). Routledge. https://doi.org/10.4324/9780203852279

Hyltenstam, K., & Stroud, C. (1993). Second language regression in Alzheimer's dementia. In K. Hyltenstam & Å. Viberg (Eds.), *Progression & regression in language: Sociocultural, neuropsychological, & linguistic perspectives* (pp. 222–243). Cambridge University Press.

Hyslop, F., & Allan, A. (2015, September 5). *Scots language policy: English version*. Gov.Scot. https://www.gov.scot/binaries/content/documents/govscot/publications/factsheet/2015/09/scots-language-policy-english/documents/scots-language-policy-english-pdf/scots-language-policy-english-pdf/govscot%3Adocument/scots%2Blanguage%2Bpolicy%2B-%2Benglish.pdf

Ishigami, Y., Eskes, G. A., Tyndall, A. V., Longman, R. S., Drogos, L. L., & Poulin, M. J. (2016). The Attention Network Test-Interaction (ANT-I): Reliability and validity in healthy older adults. *Experimental Brain Research*, *234*(3), 815–827. https://doi.org/10.1007/s00221-015-4493-4

Ishigami, Y., & Klein, R. M. (2010). Repeated measurement of the components of attention using two versions of the Attention Network Test (ANT): Stability, isolability, robustness, and reliability. *Journal of Neuroscience Methods*, *190*(1), 117–128. https://doi.org/10.1016/j.jneumeth.2010.04.019

Ishigami, Y., & Klein, R. M. (2011). Repeated Measurement of the Components of Attention of Older Adults using the Two Versions of the Attention Network Test: Stability, Isolability, Robustness, and Reliability. *Frontiers in Aging Neuroscience*, *3*. https://doi.org/10.3389/fnagi.2011.00017

Jaeger, M. E., & Rosnow, R. L. (1988). Contextualism and its implications for psychological inquiry. *British Journal of Psychology*, *79*(1), 63–75. https://doi.org/10.1111/j.2044-8295.1988.tb02273.x

Janssen, S. M. J., Chessa, A. G., & Murre, J. M. J. (2006). Memory for time: How people date events. *Memory & Cognition*, *34*(1), 138–147. https://doi.org/10.3758/BF03193393

Janssen, S. M. J., Rubin, D. C., & St. Jacques, P. L. (2011). The temporal distribution of autobiographical memory: Changes in reliving and vividness over the life span do not explain the reminiscence bump. *Memory & Cognition*, *39*(1), 1–11. https://doi.org/10.3758/s13421-010-0003-x

Jha, A. P., Krompinger, J., & Baime, M. J. (2007). Mindfulness training modifies subsystems of attention. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(2), 109–119. https://doi.org/10.3758/CABN.7.2.109

Jin, L. (2010). Remembering Simplified Hanzi 1: How Not to Forget the Meaning and Writing of Chinese Characters. by HEISIG, JAMES, &

TIMOTHY W. RICHARDSON. *The Modern Language Journal*, *94*(4), 684–685. https://doi.org/10.1111/j.1540-4781.2010.01114.x

Joffe, H. (2012). Thematic Analysis. In D. Harper & A. R. Thompson (Eds.), *Qualitative research methods in mental health and psychotherapy: A guide for students and practitioners* (pp. 209–223). John Wiley & Sons.

Johnston Jr, P. A. (2007). Scottish English and Scots. In D. Britain (Ed.), *Language in the British Isles* (pp. 105–121). Cambridge University Press.

Jorden, E. H., & Walton, R. A. (1987). Truly foreign languages: Instructional challenges. *Annals of American Academy of Political and Social Science*, *490*, 110–149.

Jylhä, M. (2004). Old Age and Loneliness: Cross-sectional and Longitudinal Analyses in the Tampere Longitudinal Study on Aging. *Canadian Journal on Aging / La Revue Canadienne Du Vieillissement*, *23*(2), 157–168. https://doi.org/10.1353/cja.2004.0023

Kalfus, R. (1977). A New Audience for Foreign Language Instruction. *Bulletin of the Association of Departments of Foreign Languages*, *9*(2), 49–50.

Kassambara, A. (2021). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests.* (R package version 0.7.0.). https://CRAN.R-project.org/package=rstatix

Kay, B. (1986). *Scots: The mither tongue*. Mainstream.

Keijzer, M. (2011). Language reversion versus general cognitive decline: Towards a new taxonomy of language change in elderly bilingual immigrants. In M. S. Schmid & W. Lowie (Eds.), *Studies in Bilingualism* (Vol. 43, pp. 221–232). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.43.15kei

Kelman, J. (1994). *Elitist Slurs Are Racism By Another Name*. https://web.archive.org/web/20130520120928/http://www.citystrolls.com/z-temp/z-pages/kelman.htm

Kidder, L. H., & Fine, M. (1987). Qualitative and quantitative methods: When stories converge. *New Directions for Program Evaluation*, *1987*(35), 57–75. https://doi.org/10.1002/ev.1459

King, N. (1998). Template Analysis. In G. Symon & C. Cassell (Eds.), *Qualitative methods and analysis in organizational research: A practical guide* (pp. 118–134). SAGE.

Kinsella, K., & Wan, H. (2009). *An Aging World: 2008* (P95/09-1; p. 196). U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/publications/2009/demo/p95-09-1.pdf

Kirk, N. W., Declerck, M., Kemp, R. J., & Kempe, V. (2022). Language control in regional dialect speakers – monolingual by name, bilingual by nature?

Bilingualism: Language and Cognition, *25*(3), 511–520. https://doi.org/10.1017/S1366728921000973

Kirk, N. W., Fiala, L., Scott-Brown, K. C., & Kempe, V. (2014). No evidence for reduced Simon cost in elderly bilinguals and bidialectals. *Journal of Cognitive Psychology*, *26*(6), 640–648. https://doi.org/10.1080/20445911.2014.929580

Kirk, N. W., Kempe, V., Scott-Brown, K. C., Philipp, A., & Declerck, M. (2018). Can monolinguals be like bilinguals? Evidence from dialect switching. *Cognition*, *170*, 164–178. https://doi.org/10.1016/j.cognition.2017.10.001

Kiyak, H. A., Teri, L., & Borson, S. (1994). Physical and Functional Health Assessment in Normal Aging and in Alzheimer's Disease: Self-Reports vs Family Reports. *The Gerontologist*, *34*(3), 324–331. https://doi.org/10.1093/geront/34.3.324

Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, *125*(5), 524–543. https://doi.org/10.1037/0033-2909.125.5.524

Kliesch, M., Giroud, N., & Meyer, M. (2021). EEG Resting-State and Event-Related Potentials as Markers of Learning Success in Older Adults Following Second Language Training: A Pilot Study. *Brain Plasticity*, *7*(2), 143–162. https://doi.org/10.3233/BPL-200117

Kliesch, M., Giroud, N., Pfenninger, S. E., & Meyer, M. (2018). 4. Research on Second Language Acquisition in Old Adulthood: What We Have and What We Need. In D. Gabryś-Barker (Ed.), *Third Age Learners of Foreign Languages* (pp. 48–75). Multilingual Matters. https://doi.org/10.21832/9781783099412-006

Kliesch, M., & Pfenninger, S. E. (2021). Cognitive and Socioaffective Predictors of L2 Microdevelopment in Late Adulthood: A Longitudinal Intervention Study. *The Modern Language Journal*, *105*(1), 237–266. https://doi.org/10.1111/modl.12696

Kliesch, M., Pfenninger, S. E., Wieling, M., Stark, E., & Meyer, M. (2021). Cognitive Benefits of Learning Additional Languages in Old Adulthood? Insights from an Intensive Longitudinal Intervention Study. *Applied Linguistics*, amab077. https://doi.org/10.1093/applin/amab077

Klimova, B., Slaninova, G., Prazak, P., Kacetl, J., & Valis, M. (2020). Enhancing Cognitive Performance of Healthy Czech Seniors through Non-Native Language Learning—A Mixed-Methods Pilot Study. *Brain Sciences*, *10*(9), 573. https://doi.org/10.3390/brainsci10090573

Knudson, D. V., & Lindsey, C. (2014). Type I and Type II Errors in Correlations of Various Sample Sizes. *Comprehensive Psychology*, *3*, 03.CP.3.1. https://doi.org/10.2466/03.CP.3.1

Kovács, Á. M., & Mehler, J. (2009). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences*, *106*(16), 6556–6560. https://doi.org/10.1073/pnas.0811323106

Kroll, J. F., Bobb, S. C., & Hoshino, N. (2014). Two Languages in Mind: Bilingualism as a Tool to Investigate Language, Cognition, and the Brain. *Current Directions in Psychological Science*, *23*(3), 159–163. https://doi.org/10.1177/0963721414528511

Kroll, J. F., Bobb, S. C., Misra, M., & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica*, *128*(3), 416–430. https://doi.org/10.1016/j.actpsy.2008.02.001

Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition*, *13*(3), 373–381. https://doi.org/10.1017/S136672891000009X

Krueger, R. A., & Casey, M. A. (2015). *Focus groups: A practical guide for applied research* (5th edition). SAGE.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics.

Labov, W. (1997). The Social Stratification of (r) in New York City Department Stores. In N. Coupland & A. Jaworski (Eds.), *Sociolinguistics* (pp. 168–178). Macmillan Education UK. https://doi.org/10.1007/978-1-349-25582-5_14

Lakens, D. (2022). *Improving Your Statistical Inferences*. https://lakens.github.io/statistical_inferences/index.html

Languages—Gov.scot. (n.d.). Gov.Scot. https://www.gov.scot/policies/languages/

Larsen, S. F., Schrauf, R. W., Fromholt, P., & Rubin, D. C. (2002). Inner speech and bilingual autobiographical memory: A Polish-Danish cross-cultural study. *Memory*, *10*(1), 45–54. https://doi.org/10.1080/09658210143000218

Lawrence-Jones, C. (2022, March 8). *The Scottish island voted one of the most beautiful in the world (but it is 11 hours from London)*. MyLondon. https://www.mylondon.news/news/uk-world-news/scottish-island-voted-one-most-23324497

Lee, G., Yip, C. C. K., Yu, E. C. S., & Man, D. W. K. (2013). Evaluation of a computer-assisted errorless learning-based memory training program for patients with early Alzheimer’s disease in Hong Kong: A pilot study. *Clinical Interventions in Aging*, 623. https://doi.org/10.2147/CIA.S45726

Lee, R. M., & Robbins, S. B. (1995). Measuring belongingness: The Social Connectedness and the Social Assurance scales. *Journal of Counseling Psychology*, *42*(2), 232–241. https://doi.org/10.1037/0022-0167.42.2.232

Leech, N. L., & Onwuegbuzie, A. J. (2009). A typology of mixed methods research designs. *Quality & Quantity*, *43*(2), 265–275. https://doi.org/10.1007/s11135-007-9105-3

Leeden, R. van der, Meijer, E., & Busing, F. M. T. A. (2008). Resampling Multilevel Models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of Multilevel Analysis* (pp. 401–433). Springer New York. https://doi.org/10.1007/978-0-387-73186-5_11

Levitt, H. M., Motulsky, S. L., Wertz, F. J., Morrow, S. L., & Ponterotto, J. G. (2017). Recommendations for designing and reviewing qualitative research in psychology: Promoting methodological integrity. *Qualitative Psychology*, *4*(1), 2–22. https://doi.org/10.1037/qup0000082

Lewis, P., & Simons, G. (2010). Assessing endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique*, *55*, 103–120.

Li, P., & Ren, Y. (1986). *汉字教学简述——对外汉语教学发展史之一章 (A brief introduction to the teaching of Chinese characters——A chapter of the history of teaching Chinese as a foreign language)*. 307–314.

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. https://doi.org/10.1016/j.jesp.2018.08.009

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 5–55.

Lingo Flamingo—Our story. (n.d.). Lingoflamingo.Co.Uk. https://www.lingoflamingo.co.uk/our-story

Liu, I. (1983). The learning of characters: A conceptual learning approach. *Journal of the Chinese Language Teachers Association*, *18*, 65–76.

Liu, M. (2017). Labelling and Direction of Slider Questions: Results from web survey experiments. *International Journal of Market Research*, *59*(5).

Liu, M., & Conrad, F. G. (2019). Where Should I Start? On Default Values for Slider Questions in Web Surveys. *Social Science Computer Review*, *37*(2), 248–269. https://doi.org/10.1177/0894439318755336

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S. G., Dias, A., Fox, N., Gitlin, L. N., Howard, R., Kales, H. C., Kivimäki, M.,

Larson, E. B., Ogunniyi, A., … Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet*, *396*(10248), 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6

Long, M. R., Vega-Mendoza, M., Rohde, H., Sorace, A., & Bak, T. H. (2020). Understudied factors contributing to variability in cognitive performance related to language learning. *Bilingualism: Language and Cognition*, *23*(4), 801–811. https://doi.org/10.1017/S1366728919000749

Lost for Words. (2021, July 5). BBC Radio 4. https://www.bbc.co.uk/programmes/m000xf0l

Loy, A., Steele, S., & Korobova, J. (2022). *lmeresampler: Bootstrap Methods for Nested Linear Mixed-Effects Models* (R package version 0.2.2). https://github.com/aloy/lmeresampler

Lüdecke, D. (2021). *sjPlot: Data Visualization for Statistics in Social Science* (R package version 2.8.10). https://CRAN.R-project.org/package=sjPlot

Luengo-Fernandez, R., Leal, J., & Gray, A. (2015). UK research spend in 2008 and 2012: Comparing stroke, cancer, coronary heart disease and dementia. *BMJ Open*, *5*(4), e006648–e006648. https://doi.org/10.1136/bmjopen-2014-006648

Lundquist, B., & Vangsnes, Ø. A. (2018). Language Separation in Bidialectal Speakers: Evidence From Eye Tracking. *Frontiers in Psychology*, *9*, 1394. https://doi.org/10.3389/fpsyg.2018.01394

Ma, Y. (2018). 论面向非汉字文化圈外国成年人的汉字教学原则 (On the principles of teaching Chinese characters for foreign adults in non-Chinese culture circle). *国际汉语教学研究 (Journal of International Chinese Teaching)*, *2*, 18–25.

Mackey, W. F. (1989). Determining the Status and Function of Languages in Multinational Societies. In U. Ammon (Ed.), *Status and Function of Languages and Language Varieties* (pp. 3–21). DE GRUYTER. https://doi.org/10.1515/9783110860252.3

MacLeod, J. W., Lawrence, M. A., McConnell, M. M., Eskes, G. A., Klein, R. M., & Shore, D. I. (2010). Appraising the ANT: Psychometric and theoretical considerations of the Attention Network Test. *Neuropsychology*, *24*(5), 637–651. https://doi.org/10.1037/a0019803

Madill, A., Jordan, A., & Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *British Journal of Psychology*, *91*(1), 1–20. https://doi.org/10.1348/000712600161646

Maeda, H. (2015). Response option configuration of online administered Likert scales. *International Journal of Social Research Methodology*, *18*(1), 15–26. https://doi.org/10.1080/13645579.2014.885159

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Marian, V., & Kaushanskaya, M. (2004). Self-construal and emotion in bicultural bilinguals. *Journal of Memory and Language*, *51*(2), 190–201. https://doi.org/10.1016/j.jml.2004.04.003

Marian, V., & Kaushanskaya, M. (2008). Words, feelings, and bilingualism: Cross-linguistic differences in emotionality of autobiographical memories. *The Mental Lexicon*, *3*(1), 72–91. https://doi.org/10.1075/ml.3.1.06mar

Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General*, *129*(3), 361–368. https://doi.org/10.1037/0096-3445.129.3.361

Marian, V., & Spivey, M. (2003). Bilingual and monolingual processing of competing lexical items. *Applied Psycholinguistics*, *24*(2), 173–193. https://doi.org/10.1017/S0142716403000092

Marian, V., Spivey, M., & Hirsch, J. (2003). Shared and separate systems in bilingual language processing: Converging evidence from eyetracking and brain imaging. *Brain and Language*, *86*(1), 70–82. https://doi.org/10.1016/S0093-934X(02)00535-7

Marsh, B. U., Kanaya, T., & Pezdek, K. (2015). The language dependent recall effect influences the number of items recalled in autobiographical memory reports. *Journal of Cognitive Psychology*, *27*(7), 829–843. https://doi.org/10.1080/20445911.2015.1046876

Martin, G. N., Carlson, N. R., & Buskist, W. (2010). *Psychology* (4th ed). Allyn and Bacon.

Martin-Rhee, M. M., & Bialystok, E. (2008). The development of two types of inhibitory control in monolingual and bilingual children. *Bilingualism: Language and Cognition*, *11*(1), 81–93. https://doi.org/10.1017/S1366728907003227

Matheson, C., & Matheson, D. (2000). Languages of Scotland: Culture and the classroom. *Comparative Education*, *36*(2), 211–221. https://doi.org/10.1080/713656603

Matsumoto, A., & Stanny, C. (2006). Language-dependent access to autobiographical memory in Japanese-English bilinguals and US monolinguals. *Memory*, *14*(3), 378–390. https://doi.org/10.1080/09658210500365763

Mattschey, J. (2020). *Repeating the same mistakes: A century of research interested in the effects of bilingualism on non-linguistic cognitive functions* [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/xe5k2

Maxwell, J. A. (2012). *A realist approach for qualitative research*. SAGE Publications.

McCLure, J. D. (1979). Scots: Its range of uses. In A. J. Aitken & T. McArthur (Eds.), *Languages of Scotland* (pp. 26–48). W. and R. Chambers.

McCLure, J. D. (2000). *Language, poetry, and nationhood: Scots as a poetic language from 1878 to the present*. Tuckwell Press.

McRae, K., & Jones, M. N. (2013). Semantic Memory. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 206–216). Oxford University Press.

Meehl, P. E. (1990). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Meeter, M., Murre, J. M. J., & Janssen, S. M. J. (2005). Remembering the news: Modeling retention data from a study with 14,000 participants. *Memory & Cognition*, *33*(5), 793–810. https://doi.org/10.3758/BF03193075

Meltzer, J. A., Kates Rose, M., Le, A. Y., Spencer, K. A., Goldstein, L., Gubanova, A., Lai, A. C., Yossofzai, M., Armstrong, S. E. M., & Bialystok, E. (2021). Improvement in executive function for older adults through smartphone apps: A randomized clinical trial comparing language learning and brain training. *Aging, Neuropsychology, and Cognition*, 1–22. https://doi.org/10.1080/13825585.2021.1991262

Merton, R. K., Fiske, M., & Kendall, P. L. (1956). *The focused interview; a manual of problems and procedures*. Free Press.

Merton, R. K., & Kendall, P. L. (1946). The Focused Interview. *American Journal of Sociology*, *51*(6), 541–557. https://doi.org/10.1086/219886

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Millar, R. M. (2006). 'Burying Alive': Unfocussed Governmental Language Policy and Scots. *Language Policy*, *5*(1), 63–86. https://doi.org/10.1007/s10993-005-5626-6

Morey, R., & Lakens, D. (2016). *Why Most Of Psychology Is Statistically Unfalsifiable*. https://doi.org/10.5281/ZENODO.838685

Morgan, D. (1997). *Focus Groups as Qualitative Research*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984287

Morgan, D. L. (1998). Practical Strategies for Combining Qualitative and Quantitative Methods: Applications to Health Research. *Qualitative Health Research*, *8*(3), 362–376. https://doi.org/10.1177/104973239800800307

Morgan, D. L. (2014). *Integrating Qualitative and Quantitative Methods: A Pragmatic Approach*. SAGE Publications, Inc. https://doi.org/10.4135/9781544304533

Morgan, D. L., & Hoffman, K. (2021). Searching for qualitatively driven mixed methods research: A citation analysis. *Quality & Quantity*, *55*(2), 731–740. https://doi.org/10.1007/s11135-020-01025-2

Morse, J. M. (Ed.). (1992). *Qualitative health research*. SAGE Publications.

Mortensen, L., Berntsen, D., & Bohn, O.-S. (2015). Retrieval of bilingual autobiographical memories: Effects of cue language and cue imageability. *Memory*, *23*(2), 138–156. https://doi.org/10.1080/09658211.2013.873809

Mosley, M. (Director). (2018, November 28). Trust me, I'm a Doctor (No. 1). In *Trust me, I'm a Doctor*. BBC Two. https://www.bbc.co.uk/programmes/b0bjj24p

Nadin, S., & Cassell, C. (2004). Using Data Matrices. In C. Cassell & G. Symon, *Essential Guide to Qualitative Methods in Organizational Research* (pp.

271–287). SAGE Publications Ltd. https://doi.org/10.4135/9781446280119.n22

National Academies of Sciences, Engineering, and Medicine. (2020). *Social isolation and loneliness in older adults: Opportunitiies for the health care system*. the National Academies Press. https://doi.org/10.17226/25663.

Neroni, M. A., Gamboz, N., de Vito, S., & Brandimonte, M. A. (2016). Effects of self-generated versus experimenter-provided cues on the representation of future events. *Quarterly Journal of Experimental Psychology*, *69*(9), 1799–1811. https://doi.org/10.1080/17470218.2015.1100205

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694–706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Nicholls, M. E. R., Orr, C. A., Okubo, M., & Loftus, A. (2006). Satisfaction Guaranteed: The Effect of Spatial Biases on Responses to Likert Scales. *Psychological Science*, *17*(12), 1027–1028. https://doi.org/10.1111/j.1467-9280.2006.01822.x

Nilsson, J., Berggren, R., Garzón, B., Lebedev, A. V., & Lövdén, M. (2021). Second Language Learning in Older Adults: Effects on Brain Structure and Predictors of Learning Success. *Frontiers in Aging Neuroscience*, *13*, 666851. https://doi.org/10.3389/fnagi.2021.666851

Noble, W. S. (2009). How does multiple testing correction work? *Nature Biotechnology*, *27*(12), 1135–1137. https://doi.org/10.1038/nbt1209-1135

Nosowitz, D. (2018, September 28). How the English Failed to Stamp Out the Scots Language. *Atlas Obscura*. https://www.atlasobscura.com/articles/scots-language

NVivo (Version 12). (2022). QSR International Pty Ltd. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Onwuegbuzie, A. J. (2012). Introduction: Putting the *MIXED* back into quantitative and qualitative research in educational research and beyond: Moving toward the *radical middle*. *International Journal of Multiple Research Approaches*, *6*(3), 192–219. https://doi.org/10.5172/mra.2012.6.3.192

Onwuegbuzie, A. J., & Johnson, B. R. (2006). The Validity Issue in Mixed Research. *Research in Schools*, *13*(1), 48–63.

http://videolectures.net/site/normal_dl/tag=48066/MixedMethodsaandValidity.RITS.pdf

Ooi, S. H., Goh, W. D., Sorace, A., & Bak, T. H. (2018). From Bilingualism to Bilingualisms: Bilingual experience in Edinburgh and Singapore affects attentional control differently. *Bilingualism: Language and Cognition*, *21*(4), 867–879. https://doi.org/10.1017/S1366728918000020

Osborne, C., Zhang, Q., & Zhang, G. X. (2020). Which is more effective in introducing Chinese characters? An investigative study of four methods used to teach CFL beginners. *The Language Learning Journal*, *48*(4), 385–401. https://doi.org/10.1080/09571736.2017.1393838

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*, 232–258.

Paap, K. R., Johnson, H. A., & Sawi, O. (2014). Are bilingual advantages dependent upon specific tasks or specific bilingual experiences? *Journal of Cognitive Psychology*, *26*(6), 615–639. https://doi.org/10.1080/20445911.2014.944914

Paap, K. R., Johnson, H. A., & Sawi, O. (2015). Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. *Cortex*, *69*, 265–278. https://doi.org/10.1016/j.cortex.2015.04.014

Paap, K. R., Myuz, H. A., Anders, R. T., Bockelman, M. F., Mikulinsky, R., & Sawi, O. M. (2017). No compelling evidence for a bilingual advantage in switching or that frequent language switching reduces switch cost. *Journal of Cognitive Psychology*, *29*(2), 89–112. https://doi.org/10.1080/20445911.2016.1248436

Packard, J. L. (1990). Effects of Time Lag in the Introduction of Characters into the Chinese Language Curriculum. *The Modern Language Journal*, *74*(2), 167–175. https://doi.org/10.1111/j.1540-4781.1990.tb02562.x

Paplikar, A., Mekala, S., Bak, T. H., Dharamkar, S., Alladi, S., & Kaul, S. (2019). Bilingualism and the severity of poststroke aphasia. *Aphasiology*, *33*(1), 58–72. https://doi.org/10.1080/02687038.2017.1423272

Paradis, M. (2004). *A neurolinguistic theory of bilingualism*. J. Benjamins Pub.

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (Fourth edition). SAGE Publications, Inc.

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Elsevier. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Payne, S. (2016). Grounded Theory. In E. Lyons & A. Coyle (Eds.), *Analysing qualitative data in psychology* (2nd edition, pp. 119–146). SAGE.

Peal, E., & Lambert, W. E. (1962). The relation of bilingualism to intelligence. *Psychological Monographs: General and Applied*, *76*(27), 1–23. https://doi.org/10.1037/h0093840

Pedersen, T. L. (2020). *patchwork: The Composer of Plots* (R package version 1.1.1). https://CRAN.R-project.org/package=patchwork

Perlman, D., & Landolt, M. A. (1999). Examination of Loneliness in Children–Adolescents and in Adults: Two Solitudes or Unified Enterprise? In K. J. Rotenberg & S. Hymel (Eds.), *Loneliness in Childhood and Adolescence* (1st ed., pp. 325–347). Cambridge University Press. https://doi.org/10.1017/CBO9780511551888.016

Perlman, D., & Peplau, L. A. (1981). Toward a Social Psychology of Loneliness. In S. Duck & R. Gilmour (Eds.), *Personal relationships. 3: Personal relationships in disorder* (Vol. 3, pp. 31–56). Academic Press.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, *316*(7139), 1236–1238. https://doi.org/10.1136/bmj.316.7139.1236

Pfenninger, S. E., & Polz, S. (2018). Foreign language learning in the third age: A pilot feasibility study on cognitive, socio-affective and linguistic drivers and benefits in relation to previous bilingualism of the learner. *Journal of the European Second Language Association*, *2*(1), 1. https://doi.org/10.22599/jesla.36

Pfenninger, S. E., & Singleton, D. (2019). A critical review of research relating to the learning, use and effects of additional and multiple languages in later life. *Language Teaching*, *52*(4), 419–449. https://doi.org/10.1017/S0261444819000235

Philip, B. K. (1990). Parametric Statistics for Evaluation of the Visual Analog Scale: *Anesthesia & Analgesia*, *71*(6), 710. https://doi.org/10.1213/00000539-199012000-00027

Pidgeon, N., & Henwood, K. (1997). Using grounded theory in psychological research. In Nicky Hayes (Ed.), *Doing Qualitative Analysis In Psychology* (0 ed., pp. 245–273). Psychology Press. https://doi.org/10.4324/9781315804552

Pilgrim, D. (2014). Some implications of critical realism for mental health research. *Social Theory & Health*, *12*(1), 1–21. https://doi.org/10.1057/sth.2013.17

Platz, F., Kopiez, R., Hasselhorn, J., & Wolf, A. (2015). The impact of song-specific age and affective qualities of popular songs on music-evoked autobiographical memories (MEAMs). *Musicae Scientiae*, *19*(4), 327–349. https://doi.org/10.1177/1029864915597567

Pliatsikas, C. (2020). Understanding structural plasticity in the bilingual brain: The Dynamic Restructuring Model. *Bilingualism: Language and*

Cognition, *23*(2), 459–471. https://doi.org/10.1017/S1366728919000130

Polit, D. F., & Beck, C. T. (2010). Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies*, *47*(11), 1451–1458. https://doi.org/10.1016/j.ijnurstu.2010.06.004

Popper, K. R. (1959). *The logic of scientific discovery*. Routledge. http://sfx.urv.cat/urv?url_ver=Z39.88-2004&url_ctx_fmt=info:ofi/fmt:kev:mtx:ctx&ctx_enc=info:ofi/enc:UTF-8&ctx_ver=Z39.88-2004&rfr_id=info:sid/sfxit.com:azbook&sfx.ignore_date_threshold=1&rft.object_id=1000000000009038

Popper, K. R. (1969). *Conjectures and refutations: The growth of scientific knowledge*. Routledge and Kegan Paul.

Posner, M. I., & Petersen, S. E. (1990). The Attention System of the Human Brain. *Annual Review of Neuroscience*, *13*(1), 25–42. https://doi.org/10.1146/annurev.ne.13.030190.000325

Pot, A., Porkert, J., & Keijzer, M. (2019). The Bidirectional in Bilingual: Cognitive, Social and Linguistic Effects of and on Third-Age Language Learning. *Behavioral Sciences*, *9*(9), 98. https://doi.org/10.3390/bs9090098

Potter, J. (2004). Discourse Analysis. In M. Hardy & A. Bryman, *Handbook of Data Analysis* (pp. 607–624). SAGE Publications, Ltd. https://doi.org/10.4135/9781848608184.n27

Potter, J., & Wetherell, M. (1987). *Discourse and social psychology: Beyond attitudes and behaviour*. Sage Publications.

Prior, A., & Macwhinney, B. (2010). A bilingual advantage in task switching. *Bilingualism: Language and Cognition*, *13*(2), 253–262. https://doi.org/10.1017/S1366728909990526

Purdie, N., & Boulton-Lewis, G. (2003). THE LEARNING NEEDS OF OLDER ADULTS. *Educational Gerontology*, *29*(2), 129–149. https://doi.org/10.1080/713844281

Qualtrics. (n.d.). Qualtrics. https://www.qualtrics.com

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rasmussen, A. S., & Berntsen, D. (2009). Emotional valence and the functions of autobiographical memories: Positive and negative memories serve different functions. *Memory & Cognition*, *37*(4), 477–492. https://doi.org/10.3758/MC.37.4.477

Raven, J. C., & Foulds, G. A. (1962). *Advanced Progressive Matrices*. H.K. Lewis & Co. Ltd.

Raven, J. C., Raven, J., & Court, J. H. (1998). Section 4: Advanced Progressive Matrices. In *Manual for Raven's progressive matrices and vocabulary scales*. Harcourt Assessment, Inc.

Reeve, A., Simcox, E., & Turnbull, D. (2014). Ageing and Parkinson's disease: Why is advancing age the biggest risk factor? *Ageing Research Reviews*, *14*, 19–30. https://doi.org/10.1016/j.arr.2014.01.004

Reicher, S. (2000). Against methodolatry: Some comments on Elliott, Fischer, and Rennie. *British Journal of Clinical Psychology*, *39*(1), 1–6. https://doi.org/10.1348/014466500163031

Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, *40*(3), 699–704. https://doi.org/10.3758/BRM.40.3.699

Reitz, C., Brayne, C., & Mayeux, R. (2011). Epidemiology of Alzheimer disease. *Nature Reviews Neurology*, *7*(3), 137–152. https://doi.org/10.1038/nrneurol.2011.2

Rennie, D. L. (2000). Grounded Theory Methodology as Methodical Hermeneutics: Reconciling Realism and Relativism. *Theory & Psychology*, *10*(4), 481–502. https://doi.org/10.1177/0959354300104003

Richardson, T. W. (1998). *James W. Heisig's System for Remembering Kanji: An Examination of Relevant Theory and Research – and a 1,000-Character Adaptation for Chinese* [Doctor of Philosophy]. The University of Texas at Austin.

Ricoeur, P. (1996). On Interpretation. In R. Kearney & M. Rainwater (Eds.), *The Continental philosophy reader* (pp. 138–155). Routledge.

Riessman, C. K. (2008). *Narrative methods for the human sciences*. Sage Publications.

Ritchie, J., & Spencer, L. (1994). Qualitative data analysis for applied policy research. In A. Bryman & R. G. Burgess (Eds.), *Analyzing qualitative data* (pp. 173–194). Routledge.

Rivas, C. (2018). Finding themes in qualiative data. In C. Seale (Ed.), *Researching society and culture* (4th edition, pp. 429–453). Sage.

Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1994). *The Test of Everyday Attention (TEA)*. Pearson Assessment.

Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1996). The structure of normal human attention: The Test of Everyday Attention. *Journal of the International Neuropsychological Society*, *2*(6), 525–534. https://doi.org/10.1017/S1355617700001697

Rosenzweig, M. R., Bennett, E. L., & Krech, D. (1964). Cerebral effects of environmental complexity and training among adult rats. *Journal of*

Comparative and Physiological Psychology, *57*(3), 438–439. https://doi.org/10.1037/h0046387

Rosenzweig, M. R., Krech, D., Bennett, E. L., & Diamond, M. C. (1962). Effects of environmental complexity and training on brain chemistry and anatomy: A replication and extension. *Journal of Comparative and Physiological Psychology*, *55*(4), 429–437. https://doi.org/10.1037/h0041137

Roster, C. A., Lucianetti, L., & Albaum, G. (2015). Exploring Slider vs. Categorical Response Formats in Web-Based Surveys. *Journal of Research Practice*, *11*(1), Article D1. https://jrp.icaap.org/index.php/jrp/article/view/509/413

Rubin, D. C. (1980). 51 properties of 125 words: A unit analysis of verbal behavior. *Journal of Verbal Learning and Verbal Behavior*, *19*(6), 736–755. https://doi.org/10.1016/S0022-5371(80)90415-6

Rubin, D. C. (1998). Beginnings of a theory of autobiographical remembering. In C. P. Thompson, D. J. Herrmann, D. Bruce, J. D. Read, D. G. Payne, & M. P. Toglia (Eds.), *Autobiographical memory: Theoretical and applied perspectives* (pp. 44–67). Lawrence Erlbaum Associates.

Rubin, D. C., Groth, E., & Goldsmith, D. J. (1984). Olfactory Cuing of Autobiographical Memory. *The American Journal of Psychology*, *97*(4), 493. https://doi.org/10.2307/1422158

Rubin, D. C., Schrauf, R. W., & Greenberg, D. L. (2003). Belief and recollection of autobiographical memories. *Memory & Cognition*, *31*(6), 887–901. https://doi.org/10.3758/BF03196443

Rubin, D. C., & Schulkind, M. D. (1997). The distribution of autobiographical memories across the lifespan. *Memory & Cognition*, *25*(6), 859–866. https://doi.org/10.3758/BF03211330

Russel, L. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means.* (R package version 1.7.4-1). https://CRAN.R-project.org/package=emmeans

Russell, D., Peplau, L. A., & Ferguson, M. L. (1978). Developing a Measure of Loneliness. *Journal of Personality Assessment*, *42*(3), 290–294. https://doi.org/10.1207/s15327752jpa4203_11

Saer, D. J. (1923). THE EFFECT OF BILINGUALISM ON INTELLIGENCE. *British Journal of Psychology. General Section*, *14*(1), 25–38. https://doi.org/10.1111/j.2044-8295.1923.tb00110.x

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126. https://doi.org/10.1037/met0000275

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on*

Psychological Science, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Schmid, M. S., & Dusseldorp, E. (2010). Quantitative analyses in a multivariate study of language attrition: The impact of extralinguistic factors. *Second Language Research*, *26*(1), 125–160. https://doi.org/10.1177/0267658309337641

Schmid, M. S., & Keijzer, M. (2009). First language attrition and reversion among older migrants. *International Journal of the Sociology of Language*, *2009*(200). https://doi.org/10.1515/IJSL.2009.046

Schrauf, R. W. (2003). A protocol analysis of retrieval in bilingual autobiographical memory. *International Journal of Bilingualism*, *7*(3), 235–256. https://doi.org/10.1177/13670069030070030201

Schrauf, R. W., & Durazo-Arvizu, R. (2006). 11. Bilingual Autobiographical Memory and Emotion: Theory and Methods. In A. Pavlenko (Ed.), *Bilingual Minds* (pp. 284–311). Multilingual Matters. https://doi.org/10.21832/9781853598746-013

Schrauf, R. W., & Rubin, D. C. (1998). Bilingual Autobiographical Memory in Older Adult Immigrants: A Test of Cognitive Explanations of the Reminiscence Bump and the Linguistic Encoding of Memories. *Journal of Memory and Language*, *39*(3), 437–457. https://doi.org/10.1006/jmla.1998.2585

Schrauf, R. W., & Rubin, D. C. (2000). Internal languages of retrieval: The bilingual encoding of memories for the personal past. *Memory & Cognition*, *28*(4), 616–623. https://doi.org/10.3758/BF03201251

Schrauf, R. W., & Rubin, D. C. (2001). Effects of voluntary immigration on the distribution of autobiographical memory over the lifespan. *Applied Cognitive Psychology*, *15*(7), S75–S88. https://doi.org/10.1002/acp.835

Schrauf, R. W., & Rubin, D. C. (2004). The 'language' and 'feel' of bilingual memory: Mnemonic traces. *Estudios de Sociolinguistica*, *5*(1), 21–39. https://doi.org/10.1558/sols.v5i1.21

Schweizer, T. A., Ware, J., Fischer, C. E., Craik, F. I. M., & Bialystok, E. (2012). Bilingualism as a contributor to cognitive reserve: Evidence from brain atrophy in Alzheimer's disease. *Cortex*, *48*(8), 991–996. https://doi.org/10.1016/j.cortex.2011.04.009

Scots Language Center. (2011). *Brief Analysis of the 2011 Census Results*. Scotslanguage.Com. http://media.scotslanguage.com/library/document/SLC%20Analysis%20of%20Census%202011%20for%20Scots.pdf

Sellers, R. (2013). How Sliders Bias Survey Data. *Alert! Marketing Research Association*, *53*(3), 56–57. https://greymatterresearch.com/wp-content/uploads/2019/09/Alert-Sliders-2013.pdf

Senn, S. (2002). *Cross-over trials in clinical research* (2nd ed). J. Wiley.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New Directions for Program Evaluation*, *1993*(60), 13–57. https://doi.org/10.1002/ev.1660

Shapira, N., Barak, A., & Gal, I. (2007). Promoting older adults' well-being through Internet training and use. *Aging & Mental Health*, *11*(5), 477–484. https://doi.org/10.1080/13607860601086546

Sibbald, B., & Roberts, C. (1998). Understanding controlled trials: Crossover trials. *BMJ*, *316*(7146), 1719–1720. https://doi.org/10.1136/bmj.316.7146.1719

Siegel, J. (2010). *Second Dialect Acquisition* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511777820

Singleton, D. (2005). The Critical Period Hypothesis: A coat of many colours. *IRAL - International Review of Applied Linguistics in Language Teaching*, *43*(4), 269–285. https://doi.org/10.1515/iral.2005.43.4.269

Sjoberg, D., D., Whiting, K., Curry, M., Lavery, J., A., & Larmarange, J. (2021). Reproducible Summary Tables with the gtsummary Package. *The R Journal*, *13*(1), 570–580. https://doi.org/10.32614/RJ-2021-053

Skype. (n.d.). Microsoft. https://www.skype.com

Smith, B. (2013). Disability, sport and men's narratives of health: A qualitative study. *Health Psychology*, *32*(1), 110–119. https://doi.org/10.1037/a0029187

Smith, B. (2016). Narrative Analysis. In E. Lyons & A. Coyle (Eds.), *Analysing qualitative data in psychology* (2nd edition, pp. 202–221). SAGE.

Smith, G. C. S., & Pell, J. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ*, *327*(7429), 1459–1461. https://doi.org/10.1136/bmj.327.7429.1459

Smith, J. A., & Eatough, V. (2016). Interpretative Phenomenological Analysis. In E. Lyons & A. Coyle (Eds.), *Analysing qualitative data in psychology* (2nd edition, pp. 50–67). SAGE.

Smith, J. A., Harré, R., & van Langenhove, L. (1995). Idiography and the case study. In J. A. Smith, R. Harré, & L. van Langenhove (Eds.), *Rethinking Psychology* (pp. 59–69). Sage Publications.

Smith, J. A., & Osborn, M. (2015). Interpretative phenomenological analysis. In J. A. Smith (Ed.), *Qualitative psychology: A practical guide to research methods* (3rd edition, pp. 25–53). SAGE.

So, S. (2008). *Remembering the Kanji 1: A Complete Course on How Not to Forget the Meaning and Writing of Japanese Characters* . 5th ed. by HEISIG, JAMES W.ï¿½ *Remembering the Kana: A Guide to Reading and Writing the Japanese Syllabaries in 3 Hours Each* . 3rd ed. by HEISIG, JAMES W., HELMUT MORSBACH, & KAZUE KUREBAYASHI. *The*

Modern Language Journal, *92*(4), 663–664. https://doi.org/10.1111/j.1540-4781.2008.00793_20.x

Sorensen, A. P. Jr. (1967). Multilingualism in the Northwest Amazon. *American Anthropologist*, *69*(6), 670–684. https://doi.org/10.1525/aa.1967.69.6.02a00030

Sparkes, A. C., & Smith, B. (Eds.). (2014). *Qualitative research methods in sport, exercise and health: From process to product*. Routledge.

Spreng, R. N., & Levine, B. (2006). The temporal distribution of past and future autobiographical events across the lifespan. *Memory & Cognition*, *34*(8), 1644–1651. https://doi.org/10.3758/BF03195927

Sriwatanakul, K., Kelvie, W., Lasagna, L., Calimlim, J. F., Weis, O. F., & Mehta, G. (1983). Studies with different types of visual analog scales for measurement of pain. *Clinical Pharmacology and Therapeutics*, *34*(2), 234–239. https://doi.org/10.1038/clpt.1983.159

Steffen, E. (2016). Ethical Considerations in Qualitative Research. In E. Lyons & A. Coyle (Eds.), *Analysing qualitative data in psychology* (2nd edition, pp. 31–44). SAGE.

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *Journal of the American Statistical Association*, *54*(285), 30. https://doi.org/10.2307/2282137

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician*, *49*(1), 108–112. https://doi.org/10.1080/00031305.1995.10476125

Stern, Y. (2002). What is cognitive reserve? Theory and research application of the reserve concept. *Journal of the International Neuropsychological Society*, *8*(3), 448–460. https://doi.org/10.1017/S1355617702813248

Stern, Y., Arenaza-Urquijo, E. M., Bartrés-Faz, D., Belleville, S., Cantilon, M., Chetelat, G., Ewers, M., Franzmeier, N., Kempermann, G., Kremen, W. S., Okonkwo, O., Scarmeas, N., Soldan, A., Udeh-Momoh, C., Valenzuela, M., Vemuri, P., Vuoksimaa, E., & and the Reserve, Resilience and Protective Factors PIA Empirical Definitions and Conceptual Frameworks Workgroup. (2020). Whitepaper: Defining and investigating cognitive reserve, brain reserve, and brain maintenance. *Alzheimer's & Dementia*, *16*(9), 1305–1311. https://doi.org/10.1016/j.jalz.2018.07.219

Sterne, J. A. C. (2001). Sifting the evidence-what's wrong with significance tests? Another comment on the role of statistical methods. *BMJ*, *322*(7280), 226–231. https://doi.org/10.1136/bmj.322.7280.226

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Stickler, U., & Shi, L. (2013). Supporting Chinese speaking skills online. *System*, *41*(1), 50–69. https://doi.org/10.1016/j.system.2012.12.001

Stickler, U., & Shi, L. (2015). Eye movements of online Chinese learners. *CALICO Journal*, *32*(1), 52–81. https://doi.org/10.1558/calico.v32i1.25964

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 479–498. https://doi.org/10.1111/1467-9868.00346

Storey, L. (2016). Doing Interpretative Phenomenological Analysis. In E. Lyons & A. Coyle (Eds.), *Analysing qualitative data in psychology* (2nd edition, pp. 68–83). SAGE.

Strauss, A. L., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Sage.

Stuart-Smith, J. (2004). Scottish English: Phonology. In E. W. Schneider & B. Kortmann (Eds.), *A handbook of varieties of English: A multimedia reference tool*. Mouton de Gruyter.

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1. https://doi.org/10.2307/2331554

Sullivan, M. D., Janus, M., Moreno, S., Astheimer, L., & Bialystok, E. (2014). Early stage second-language learning improves executive control:

Evidence from ERP. *Brain and Language*, *139*, 84–98. https://doi.org/10.1016/j.bandl.2014.10.004

Sun, L., Liu, M., Hu, J., & Liang, X. (2014). A Chinese Character Teaching System Using Structure Theory and Morphing Technology. *PLoS ONE*, *9*(6), e100987. https://doi.org/10.1371/journal.pone.0100987

Svensson, E. (2000). Comparison of the Quality of Assessments Using Continuous and Discrete Ordinal Rating Scales. *Biometrical Journal*, *42*(4), 417–434. https://doi.org/10.1002/1521-4036(200008)42:4<417::AID-BIMJ417>3.0.CO;2-Z

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing Three Sources of Misresponse to Reversed Likert Items. *Journal of Marketing Research*, *45*(1), 116–131. https://doi.org/10.1509/jmkr.45.1.116

Swihart, D. W. (2004). *Success with Chinese: A communicative approach for beginners: listening & speaking*. Cheng & Tsui Co.

Tang, Y.-Y., Ma, Y., Wang, J., Fan, Y., Feng, S., Lu, Q., Yu, Q., Sui, D., Rothbart, M. K., Fan, M., & Posner, M. I. (2007). Short-term meditation training improves attention and self-regulation. *Proceedings of the National Academy of Sciences*, *104*(43), 17152–17156. https://doi.org/10.1073/pnas.0707678104

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*. https://doi.org/10.3389/fpsyg.2011.00123

Tao, L., Taft, M., & Gollan, T. H. (2015). The Bilingual Switching Advantage: Sometimes Related to Bilingual Proficiency, Sometimes Not. *Journal of the International Neuropsychological Society*, *21*(7), 531–544. https://doi.org/10.1017/S1355617715000521

Tashakkori, A., & Creswell, J. W. (2007). Editorial: Exploring the Nature of Research Questions in Mixed Methods Research. *Journal of Mixed Methods Research*, *1*(3), 207–211. https://doi.org/10.1177/1558689807302814

Tashakkori, A., Johnson, B., & Teddlie, C. (2021). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences* (Second Edition). SAGE Publications, Inc.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. SAGE.

Terry, G. (2016). Doing Thematic Analysis. In E. Lyons & A. Coyle (Eds.), *Analysing qualitative data in psychology* (2nd edition, pp. 104–118). SAGE.

The Open University. (2019). *Scots language and culture* [Educational]. Open.Edu. https://www.open.edu/openlearncreate/course/view.php?id=2705

Tinsley, T., & Board, K. (2017). *Languages for the future: The foreign languages the United Kingdom needs to become a truly global nation*. British Council.

Tipping, S. A., & Whiteside, M. (2015). Language Reversion among People with Dementia from Culturally and Linguistically Diverse Backgrounds: The Family Experience. *Australian Social Work*, *68*(2), 184–197. https://doi.org/10.1080/0312407X.2014.953187

TNS-BMRB. (2010). *Public attitudes towards the Scots language*. Scottish Government Social Research. http://www.nls.uk/scotgov/2009/298037/9780755977949.pdf

Tourtellot, A. (2007). 4th Annual Places Rated: 111 Islands. *National Geographic Traveler*, *November/December*, 109–127. https://www.nationalgeographic.com/traveler/pdf/nd07placesratedislands.pdf

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society*, *1*(2), 179–195. https://doi.org/10.1017/S0047404500000488

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving, W. Donaldson, & G. H. Bower (Eds.), *Organization of memory*. Academic Press.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford Univeristy Press.

Tulving, E. (1984). Précis of Elements of episodic memory. *Behavioral and Brain Sciences*, *7*(2), 223–238. https://doi.org/10.1017/S0140525X0004440X

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology / Psychologie Canadienne*, *26*(1), 1–12. https://doi.org/10.1037/h0080017

Tulving, E. (1989). Remembering and Knowing the Past. *The American Scientist*, *77*(4), 361–367.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373. https://doi.org/10.1037/h0020071

Tutz, G. (2022). Ordinal regression: A review and a taxonomy of models. *WIREs Computational Statistics*, *14*(2). https://doi.org/10.1002/wics.1545

U3A – our story. (n.d.). U3A. https://www.u3a.org.uk/about/history

UN DESA. (2019). *World Population Ageing 2019: Highlights* (ST/ESA/SER.A/430). United Nations, Department of Economic and

Social Affairs, Population Division. https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf

UN DESA. (2020). *World Population Ageing 2019*. United Nations Department of Economic and Social Affairs. https://doi.org/10.18356/6a8968ef-en

UNESCO. (2010). *UNESCO Atlas of the World's Languages in Danger*. United Nations, Department of Economic and Social Affairs, Population Division. https://unesdoc.unesco.org/ark:/48223/pf0000192416

Unger, J. M., Lorish, F. C., Noda, M., & Wada, Y. (1993). *A Framework for Introductory Japanese Language Curricula in American High Schools and Colleges.* The National Foreign Language Center.

Uschner, D., Schindler, D., Hilgers, R.-D., & Heussen, N. (2018). **randomizeR**: An *R* Package for the Assessment and Implementation of Randomization in Clinical Trials. *Journal of Statistical Software*, *85*(8). https://doi.org/10.18637/jss.v085.i08

Valenzuela, M. J., & Sachdev, P. (2006). Brain reserve and dementia: A systematic review. *Psychological Medicine*, *36*(4), 441–454. https://doi.org/10.1017/S0033291705006264

Valenzuela, M. J., & Sachdev, P. (2009a). Can Cognitive Exercise Prevent the Onset of Dementia? Systematic Review of Randomized Clinical Trials with Longitudinal Follow-up. *The American Journal of Geriatric*

Psychiatry, *17*(3), 179–187. https://doi.org/10.1097/JGP.0b013e3181953b57

Valenzuela, M. J., & Sachdev, P. S. (2009b). Harnessing brain and cognitive reserve for the prevention of dementia. *Indian Journal of Psychiatry*, *51 Suppl 1*, S16-21.

Valis, M., Slaninova, G., Prazak, P., Poulova, P., Kacetl, J., & Klimova, B. (2019). Impact of Learning a Foreign Language on the Enhancement of Cognitive Functions Among Healthy Older Population. *Journal of Psycholinguistic Research*, *48*(6), 1311–1318. https://doi.org/10.1007/s10936-019-09659-6

van der Leeuw, G., Leveille, S. G., Jones, R. N., Hausdorff, J. M., McLean, R., Kiely, D. K., Gagnon, M., & Milberg, W. P. (2017). Measuring attention in very old adults using the Test of Everyday Attention. *Aging, Neuropsychology, and Cognition*, *24*(5), 543–554. https://doi.org/10.1080/13825585.2016.1226747

van der Ploeg, M., Keijzer, M., & Lowie, W. (2020). Methodological concerns and their solutions in third-age language learning studies. *Dutch Journal of Applied Linguistics*, *9*(1–2), 97–108. https://doi.org/10.1075/dujal.19036.van

van der Ploeg, M., Keijzer, M., & Lowie, W. (submitted). *A practical exploration of third-age language learning needs, beliefs, and expectations.*

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(5), 615–628. https://doi.org/10.1017/S0140525X98001733

Van Herk, G. (2018). *What is sociolinguistics?* (Second edition). Wiley Blackwell.

Vega Mendoza, M. V., Blankinship, B., Norval, R., & Bak, T. H. (2021). Language learning for people living with dementia and their carers: A feasibility study. *Alzheimer's & Dementia*, *17*(S6). https://doi.org/10.1002/alz.052821

Vega-Mendoza, M., West, H., Sorace, A., & Bak, T. H. (2015). The impact of late, non-balanced bilingualism on cognitive performance. *Cognition*, *137*, 40–46. https://doi.org/10.1016/j.cognition.2014.12.008

Victor, C., Scambler, S., Bond, J., & Bowling, A. (2000). Being alone in later life: Loneliness, social isolation and living alone. *Reviews in Clinical Gerontology*, *10*(4), 407–417. https://doi.org/10.1017/S0959259800104101

Voelkl, B. (2019). Multiple testing: Correcting for alpha error inflation with false discovery rate (FDR) or family-wise error rate? *Animal Behaviour*, *155*, 173–177. https://doi.org/10.1016/j.anbehav.2019.07.001

Vorwerg, C. C., Suntharam, S., & Morand, M.-A. (2019). Language control and lexical access in diglossic speech production: Evidence from variety

switching in speakers of Swiss German. *Journal of Memory and Language*, *107*, 40–53. https://doi.org/10.1016/j.jml.2019.03.007

Voutilainen, A., Pitkäaho, T., Kvist, T., & Vehviläinen-Julkunen, K. (2016). How to ask about patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *Journal of Advanced Nursing*, *72*(4), 946–957. https://doi.org/10.1111/jan.12875

Walton, C. (2016). Doing Discourse Analysis. In E. Lyons & A. Coyle (Eds.), *Analysing qualitative data in psychology* (2nd edition, pp. 182–201). SAGE.

Wan, Y. (2018). 略论字词关系与对外汉字教学的总体设计 (On the relationship between words and characters and the overall design of teaching characters in CFL). *国际汉语教育（中英文）(International Chinese Language Education)*, *3*, 20–30.

Ware, A. T., Kirkovski, M., & Lum, J. A. G. (2020). Meta-Analysis Reveals a Bilingual Advantage That Is Dependent on Task and Age. *Frontiers in Psychology*, *11*, 1458. https://doi.org/10.3389/fpsyg.2020.01458

Ware, C., Damnee, S., Djabelkhir, L., Cristancho, V., Wu, Y.-H., Benovici, J., Pino, M., & Rigaud, A.-S. (2017). Maintaining Cognitive Functioning in Healthy Seniors with a Technology-Based Foreign Language Program:

A Pilot Feasibility Study. *Frontiers in Aging Neuroscience*, *9*. https://doi.org/10.3389/fnagi.2017.00042

Ware, C., Dautricourt, S., Gonneaud, J., & Chételat, G. (2021). Does Second Language Learning Promote Neuroplasticity in Aging? A Systematic Review of Cognitive and Neuroimaging Studies. *Frontiers in Aging Neuroscience*, *13*, 706672. https://doi.org/10.3389/fnagi.2021.706672

Watkins, C. (2020, December 15). *2020 Duolingo Language Report: United Kingdom*. Duolingo Blog. https://blog.duolingo.com/uk-language-report-2020/

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (Eds.). (1999). *Unobtrusive measures* (Rev. ed). Sage Publications.

Wei, J. (2014). Selectivity of Second Language Attrition. *Theory and Practice in Language Studies*, *4*(8), 1603–1608. https://doi.org/10.4304/tpls.4.8.1603-1608

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Wellek, S., & Blettner, M. (2012). On the Proper Use of the Crossover Design in Clinical Trials. *Deutsches Ärzteblatt International*. https://doi.org/10.3238/arztebl.2012.0276

Wewers, M. E., & Lowe, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health*, *13*(4), 227–236. https://doi.org/10.1002/nur.4770130405

Wheeler, M. E., & Ploran, E. J. (2009). Episodic Memory. In L. R. Squire (Ed.), *Encyclopedia of Neuroscience* (pp. 1167–1172). Elsevier. https://doi.org/10.1016/B978-008045046-9.00760-9

Whereby. (n.d.). Video Communication Service AS. https://whereby.com/

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wigfield, A., Turner, R., Alden, S., Green, M., & Karania, V. K. (2022). Developing a New Conceptual Framework of Meaningful Interaction for Understanding Social Isolation and Loneliness. *Social Policy and Society*, *21*(2), 172–193. https://doi.org/10.1017/S147474642000055X

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80. https://doi.org/10.2307/3001968

Wilkinson, S. (2015). Focus groups. In J. A. Smith (Ed.), *Qualitative psychology: A practical guide to research methods* (3rd edition, pp. 199–221). SAGE.

Williams, C. (2013). Emerging development of semantic and phonological routes to character decoding in Chinese as a foreign language learners. *Reading and Writing*, *26*(2), 293–315. https://doi.org/10.1007/s11145-012-9368-5

Willig, C. (1999). Beyond appearances: A critical realist appraoch to social constructionist work. In D. J. Nightingale & J. Cromby (Eds.), *Social constructionist psychology: A critical analysis of theory and practice* (pp. 37–51). Open University Press.

Willig, C. (2012). *Qualitative interpretation and analysis in psychology*. McGraw-Hill/Open University Press.

Wilson, R. S., Krueger, K. R., Arnold, S. E., Schneider, J. A., Kelly, J. F., Barnes, L. L., Tang, Y., & Bennett, D. A. (2007). Loneliness and Risk of Alzheimer Disease. *Archives of General Psychiatry*, *64*(2), 234. https://doi.org/10.1001/archpsyc.64.2.234

Windelband, W. (1998). History and Natural Science. *Theory & Psychology*, *8*(1), 5–22. https://doi.org/10.1177/0959354398081001

Wittenberg, R., Hu, B., Barraza-Araiza, L., & Rehill, A. (2019). *Projections of older people with dementia and costs of dementia care in the United Kingdom, 2019–2040* (CPEC Working Paper 5; p. 79). November 2019 Care Policy and Evaluation Centre, London School of Economics and Political Science, Alzheimer's Society.

https://www.alzheimers.org.uk/sites/default/files/2019-11/cpec_report_november_2019.pdf

Wong, P. C. M., Ou, J., Pang, C. W. Y., Zhang, L., Tse, C. S., Lam, L. C. W., & Antoniou, M. (2019). Language Training Leads to Global Cognitive Improvement in Older Adults: A Preliminary Study. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2411–2424. https://doi.org/10.1044/2019_JSLHR-L-18-0321

Wood, J. (2014, August). Away Thinking About Things—James Kelman's fighting words. *Newyorker.Com*. https://www.newyorker.com/magazine/2014/08/25/away-thinking-things

Woodrum, E. (1984). "Mainstreaming" content analysis in social science: Methodological advantages, obstacles, and solutions. *Social Science Research*, *13*(1), 1–19. https://doi.org/10.1016/0049-089X(84)90001-2

World Health Assembly, 73. (2020). *Decade of healthy ageing: The global strategy and action plan on ageing and health 2016–2020: towards a world in which everyone can live a long and healthy life: report by the Director-General*. World Health Organization. https://apps.who.int/iris/handle/10665/355618

World Health Organization. (2015). *World report on ageing and health*. World Health Organization. https://apps.who.int/iris/handle/10665/186463

World Health Organization. (2016). *Process of translation and adaptation of instruments*. World Health Organization. http://www.who.int/substance_abuse/research_tools/translation/en/

World Health Organization. (2018). *Global Health Estimates 2016: Disease burden by cause, age, sex, by country and by region, 2000– 2016.* World Health Organization. https://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html

World Health Organization. (2020a). *Decade of Healthy Ageing: Plan of Action* (p. 26). https://www.who.int/publications/m/item/decade-of-healthy-ageing-plan-of-action

World Health Organization. (2020b). *Global Health Estimates: Life expectancy and leading causes of death and disability*. World Health Organization. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates

World Health Organization. (2021, September 2). *Dementia*. Who.Int. https://www.who.int/news-room/fact-sheets/detail/dementia

Wright, D. B. (1998). Modelling clustered data in autobiographical memory research: The multilevel approach. *Applied Cognitive Psychology*, *12*(4), 339–357. https://doi.org/10.1002/(SICI)1099-0720(199808)12:4<339::AID-ACP571>3.0.CO;2-D

Wu, Y.-H., Ware, C., Damnée, S., Kerhervé, H., & Rigaud, A.-S. (2015). Bridging the digital divide in older adults: A study from an initiative to inform older adults about new technologies. *Clinical Interventions in Aging*, 193. https://doi.org/10.2147/CIA.S72399

Xiao, M., Ge, H., Khundrakpam, B. S., Xu, J., Bezgin, G., Leng, Y., Zhao, L., Tang, Y., Ge, X., Jeon, S., Xu, W., Evans, A. C., & Liu, S. (2016). Attention Performance Measured by Attention Network Test Is Correlated with Global and Regional Efficiency of Structural Brain Networks. *Frontiers in Behavioral Neuroscience*, *10*. https://doi.org/10.3389/fnbeh.2016.00194

Yakety Yak. (n.d.). *Yakety Yak Language Cafe*. https://www.yaketyyak-languagecafe.co.uk/

Ye, L. (2013). Shall we delay teaching characters in teaching Chinese as a foreign language?: Shall We Delay Teaching Characters in Teaching Chinese. *Foreign Language Annals*, *46*(4), 610–627. https://doi.org/10.1111/flan.12049

Zhan, H., & Chen, H.-J. (2014). The Role of Technology in Teaching and Learning Chinese Characters. *Inernational Journal of Technology in Teaching and Learning*, *10*(2), 147–162.

Zhao, J. (2008). 汉语作为第二语言教学: 理念与模式 (Teaching Chinese as a second language: Concepts and models). *世界汉语教学 (Chinese Teaching in the World)*, *1*, 93–107.

Zhao, L. X., Blankinship, B., Duan, Z., Huang, H., Sun, J., & Bak, T. H. (2020). Comparing Face-to-face and Online Teaching of Written and Spoken Chinese to Adult Learners: An Edinburgh-Sheffield Case Study. *International Journal of Chinese Language Teaching*, *1*(1), 83–98. https://doi.org/10.46451/ijclt.2020.06.05

Zoom. (n.d.). Zoom Video Communications, Inc. https://zoom.us/

Appendix A. Chapter 3: Cue words

Table A1. Cue words in all languages of the experiment

English	German	Polish	French	Spanish	Greek
Summer	Sommer	Lato	Été	Verano	Καλοκαίρι
Neighbours	Nachbarn	Sąsiedzi	Voisins	Vecinos	Γείτονες
Birthday	Geburtstag	Urodziny	Anniversaire	Cumpleaños	Γενέθλια
Cat	Katze	Kot	Chat	Gato	Γάτα
Mountain	Berg	Góra	Montagne	Montaña	Βουνό
Getting lost	Sich verlieren	Zgubić się	Se perdre	Perderse	Χάνομαι
Frightened	Angst haben	Bać się	Effrayé	Asustado	Φοβισμένος
Water	Wasser	Woda	Eau	Agua	Νερό
Winter	Winter	Zima	Hiver	Invierno	Χειμώνας
Friends	Freunde	Przyjaciele	Amis	Amigos	Φίλοι
Holiday	Ferien	Wakacje	Vacances	Vacaciones	Διακοπές
Dog	Hund	Pies	Chien	Perro	Σκύλος
River	Fluss	Rzeka	Rivière	Río	Ποταμός
Competition	Wettbewerb	Konkurs	Compétition	Competición	Διαγωνισμός
Laughing	Lachen	Śmiać się	Rire	Riéndose	Γελάω
Fire	Feuer	Ogień	Feu	Fuego	Φωτιά

Appendix B. Chapter 3: Memory vividity model figures

Memory vividity model prediction visualisations for significant predictors.

Figure B1. Memory vividity model: Effect of participant age

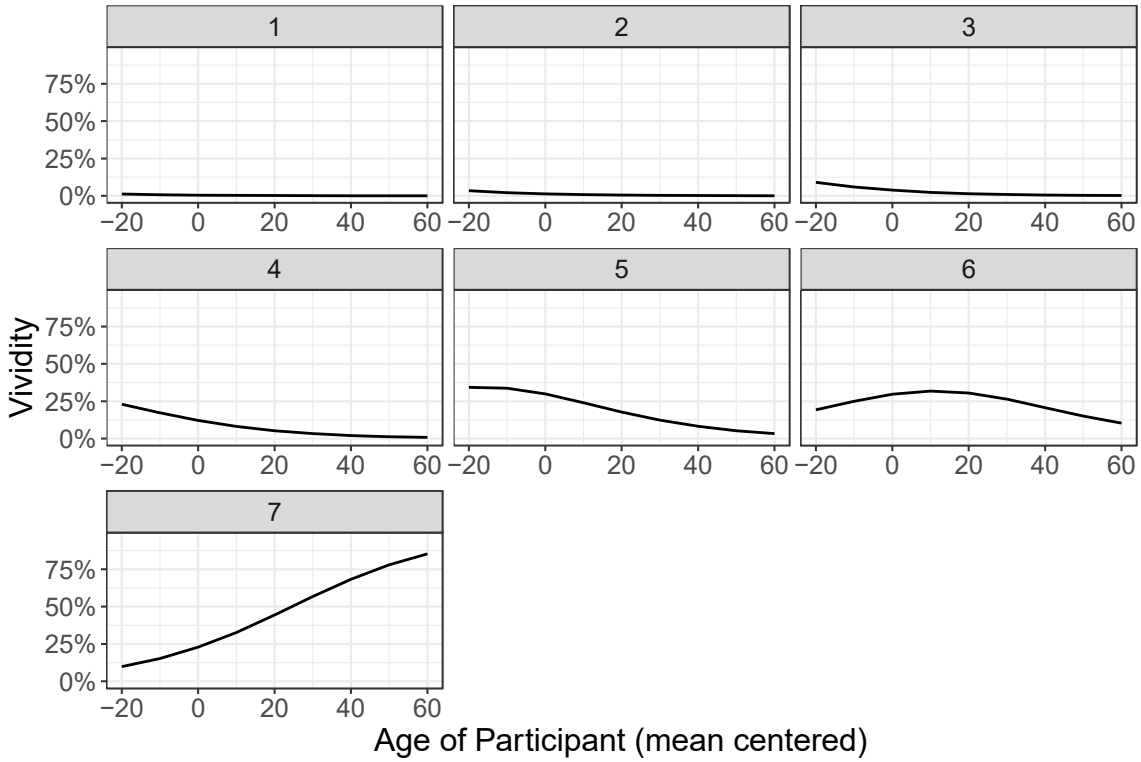


Figure B2. Memory vividity model: Effect of the age of the memory

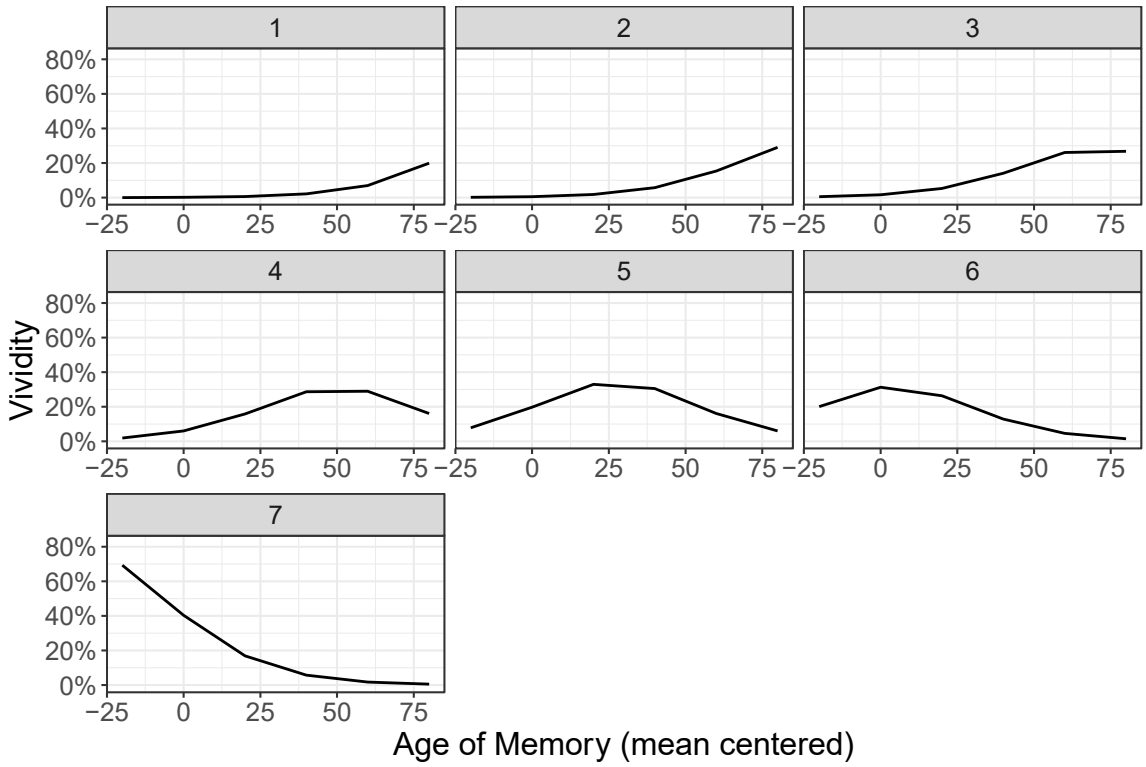
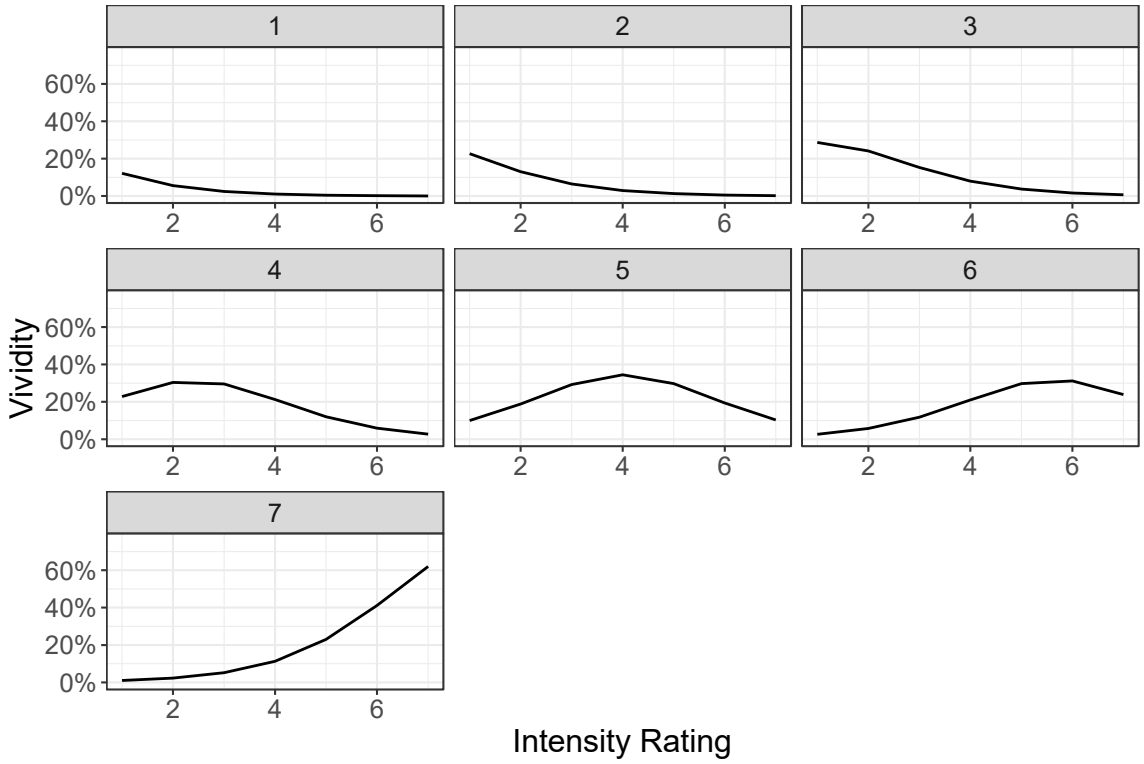


Figure B3. Memory vividity model: Effect of intensity rating



Appendix C. Chapter 3: Memory intensity model figures

Memory intensity model prediction visualisations for significant predictors.

Figure C1. Memory intensity model: Effect of age of LX acquisition

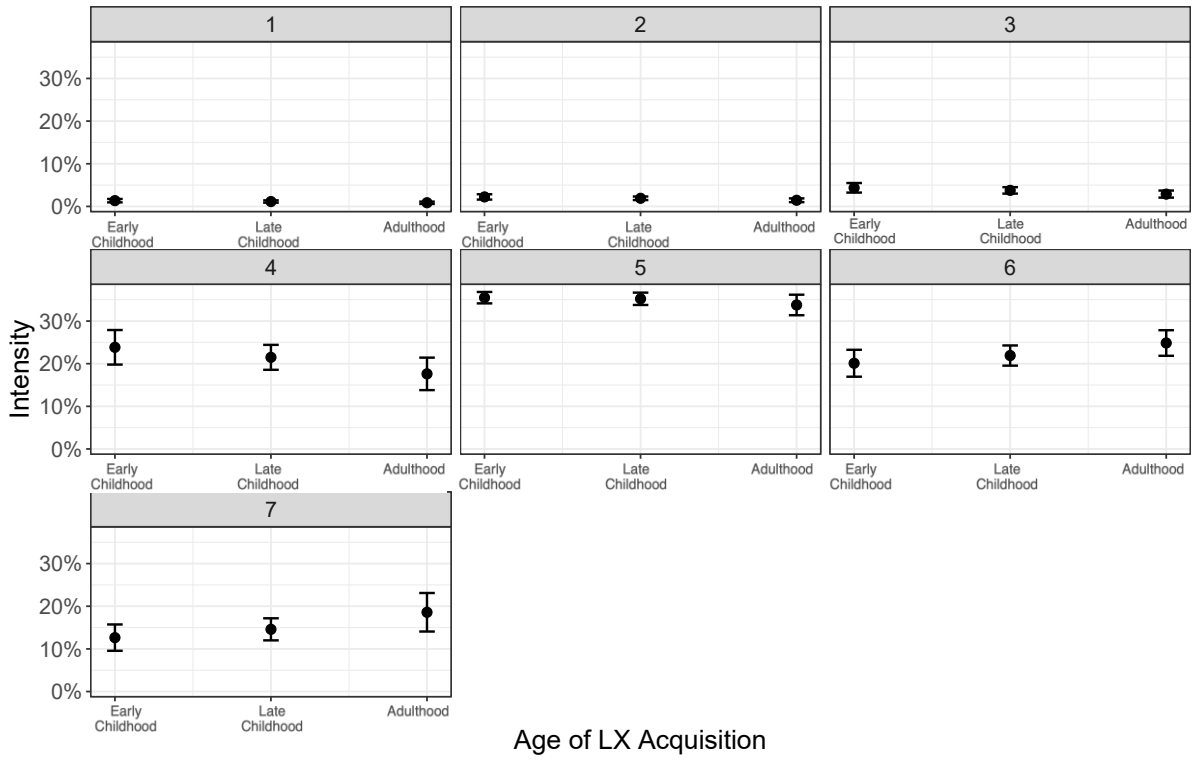


Figure C2. Memory intensity model: Effect of participant age

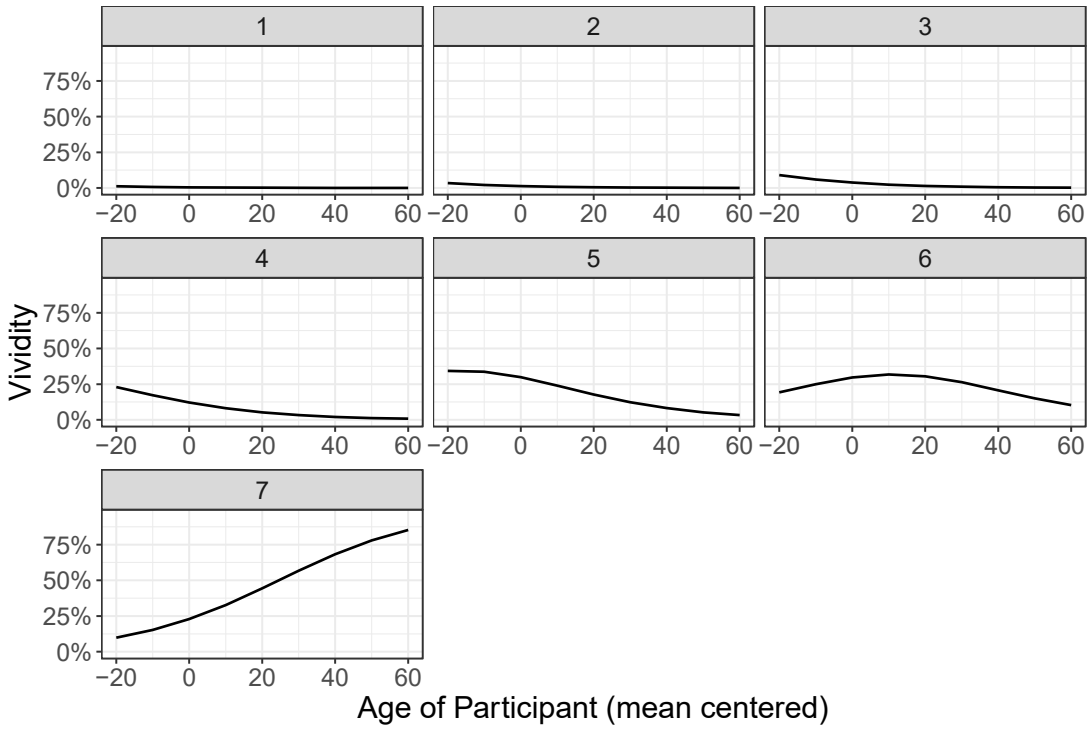
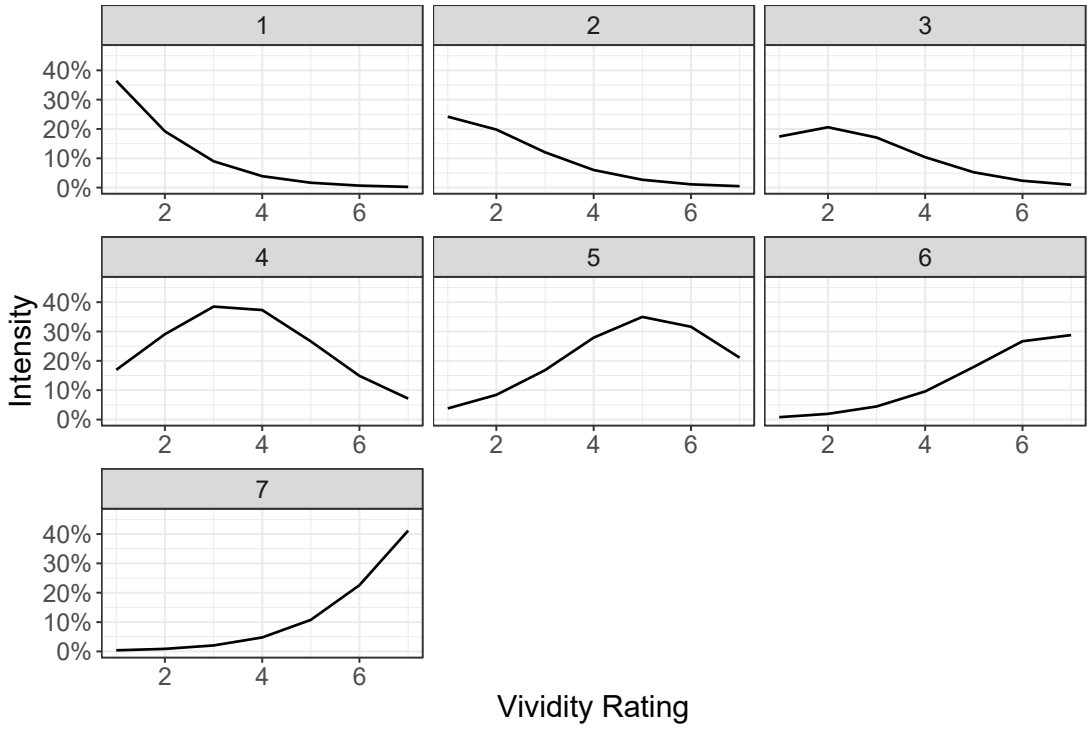


Figure C3. Memory intensity model: Effect of vividity rating



Appendix D. Chapter 3: Future events vividity model figures

Imagined future event vividity model prediction visualisations for significant predictors.

Figure D1. Future event vividity model: Effect of participant age

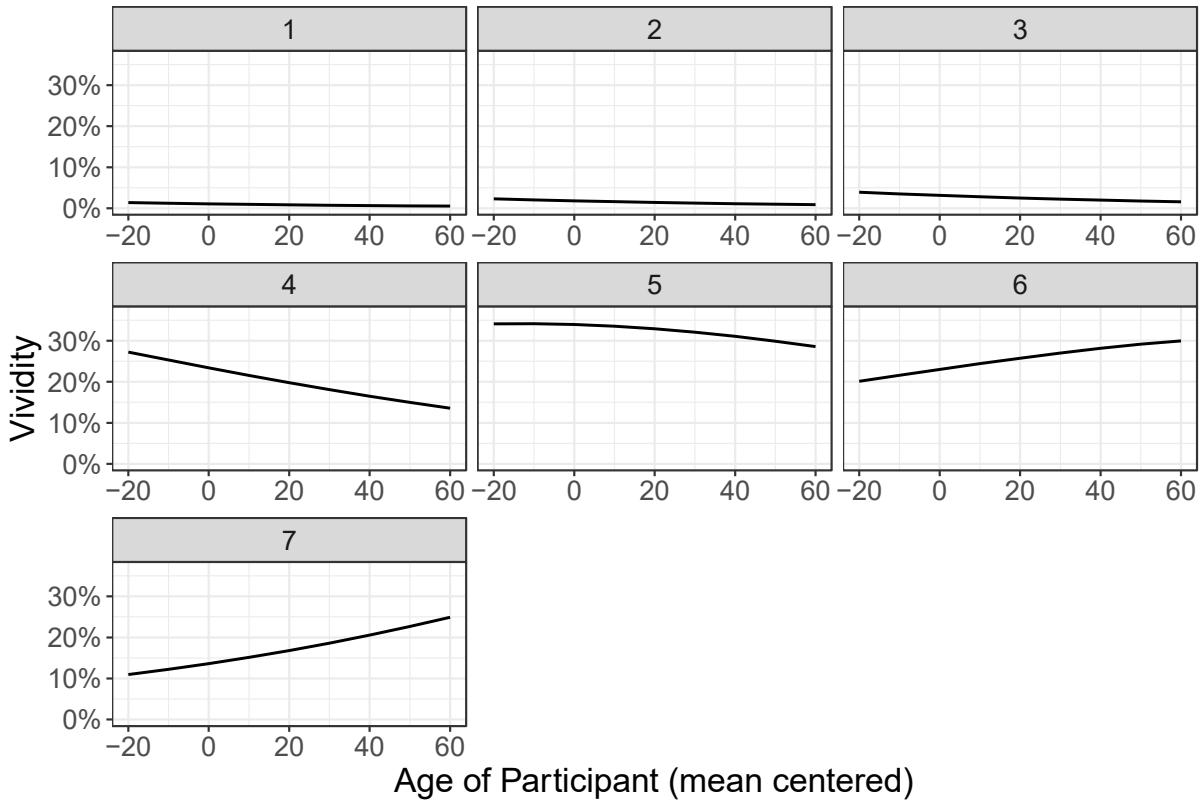
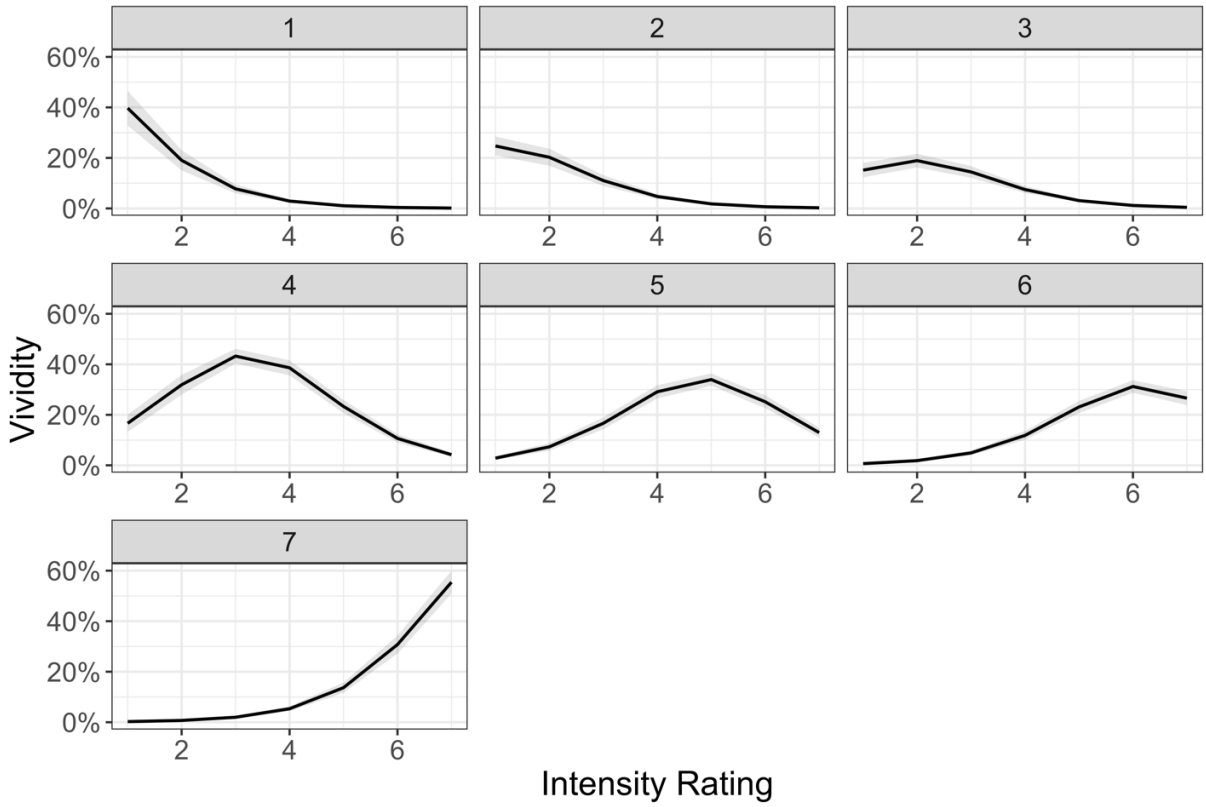


Figure D2. Future event vividity model: Effect of intensity rating



Appendix E. Chapter 3: Future events intensity model figures

Imagined future event intensity model prediction visualisations for significant predictors.

Figure E1. Future event intensity model: Effect of language of experience

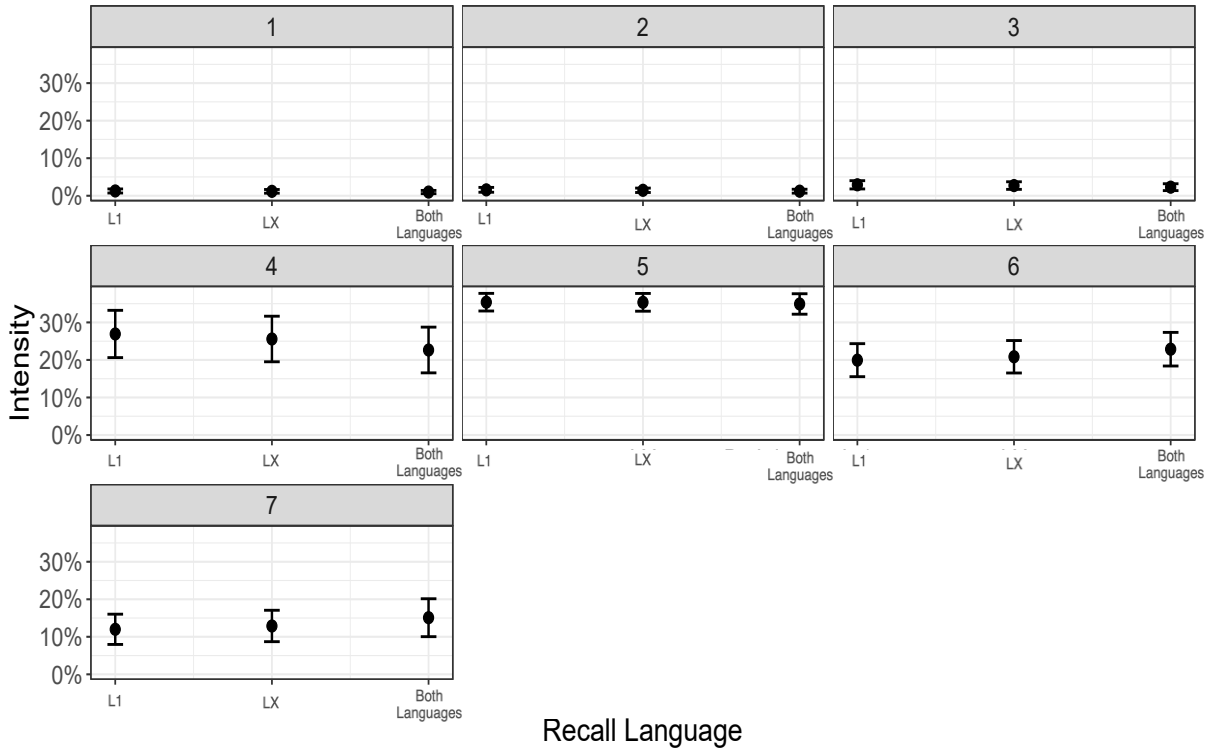
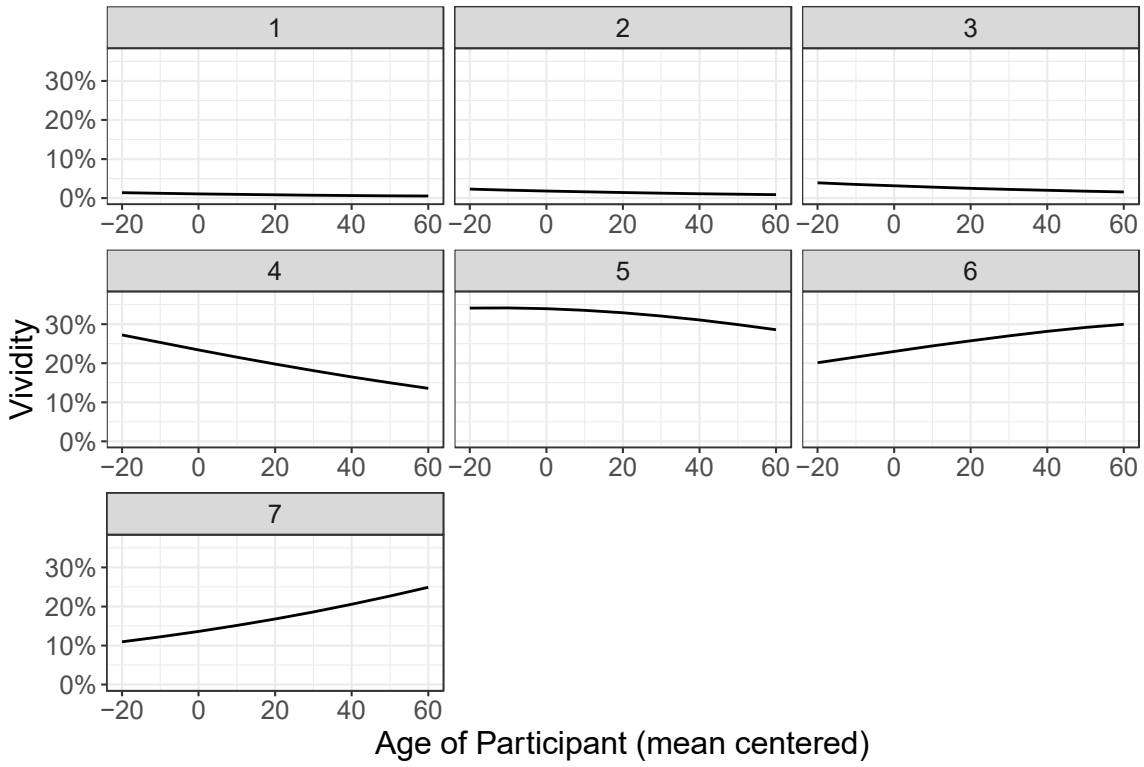


Figure E2. Future event intensity model: Effect of vividity rating



Appendix F. Chapter 4: Language History Questionnaire

In order to save space, only one language rating is included here. For each additional language, the same set of questions were asked. Additionally, the Likert scale responses were presented horizontally to participants in the online and in-person forms.

What year were you born? (enter a 4-digit birth year; for example, 1946)

_____

What is your gender?

- Male
- Female
- Prefer to self-describe _____
- Prefer not to answer

What is your dominant hand?

- Right
- Left
- Both/Ambidextrous

What is the highest level of education that you have completed?

- Primary school
- Secondary/High school
- Vocational/Professional training/degree
- University Degree
- Master's Degree
- Doctorate (e.g., PhD, MD)

At what age did you start school?

_____

What is your occupation(s)? If retired, what was your occupation(s) prior to retirement?

_____

What country were you born in?

_____

What is your current place of residence (town/city)? How long have you lived there?

_____

Have you taken any intensive residential language courses? If yes, please say which language, where it was, and for how long.

- Yes _____
- No

Please select one option below each of the following statements according to the amount of your agreement or disagreement with each item. There is no right or wrong answer.

My main motivation for learning Spanish is...

_____

It is important for members of our community to know the language that I will be learning.

- Strongly Disagree
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

Schoolchildren here should have the opportunity to study Spanish.

- Strongly Disagree
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

I would encourage others like me to learn Spanish.

- Strongly Disagree
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

Older people cannot be expected to learn a language like Spanish.

- Strongly Disagree
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

Spanish is difficult to learn.

- Strongly Disagree
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

A good language learner should have no particular trouble learning Spanish.

- Strongly Disagree
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

Learning Spanish demands less intelligence than other languages I know.

- Strongly Disagree
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

There is no good reason for anyone to need to use Spanish in the 21st century.

- Strongly Disagree
- Disagree Moderately

- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

The language I am learning has a community with its own culture.

- Strongly Disagree
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

It's really good for your brain to learn a language like Spanish.

- Strongly Disagree
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

What is/was your father's native language(s)?

_____

What is/was your mother's native language(s)?

_____

Which languages were spoken in your home and approximately what percentage of the time?

	Language name(s)	Percent of the time
Language 1		
Language 2		
Language 3		
Language 4		
Language 5		

Which languages do you know? This can be languages that you may not speak perfectly, but have learned at some point in your life, including your native language as well as English.

On the next page(s) we will ask more information about your command of each language

- Language 1 _____
- Language 2 _____
- Language 3 _____
- Language 4 _____
- Language 5 _____
- Language 6 _____
- Language 7 _____
- Language 8 _____

Language 1: _____

When was your first contact with the language?

- Since birth
- Since ____ years of age

How frequently did you use this language in:

	Never	Less than yearly	Yearly	Monthly	Weekly	Daily
Childhood						
Adulthood						
in the last 3 years						

Do you use this language in specific contexts? (e.g., family, friends, work, school, etc.)

_____

What is your current command of the language in the following categories:

	None	Beginner	Elementary	Intermediate	Advanced	Fluent
Expression						
Comprehension						
Reading						
Writing						

Select an option below as it relates to your accent in this language:

- People have great difficulty understanding me
- I often need to repeat things
- Sometimes I need to repeat things
- I don't need to repeat, but people identify me as a non-native speaker
- I am sometimes taken for a native speaker
- I am always taken for a native speaker

Any additional comments?

_____

Appendix G. Chapter 4: Feedback session interview schedule

Question: How did you find the course?

- *Thoughts?*
- *Feedback?*
- *What did you like? What would you change?*
- *Was there anything you would like or more less of?*
- *What did you think of the teacher, Max?*

Question: If you could use one word to describe the course, what would it be?

Question: How was the intensity of the course? Too little, too much, or just right?

- *In terms of meetings per week and time spent practicing the language outside of class*

Question: How did you find the length of the course? How about the length of the individual classes?

Question: What did you think about the frequency of meetings/classes?

- *Was three times a week too often, too little, or just right?*
- *Did you have homework assigned and did you do it?*

Question: How many hours of learning a week would you ideally have for a language course? What would be your ideal intensity?

- *How many hours of independent work?*
- *What kind of independent work would you be most likely to participate in outside of class?*

Question: Were the expectations that you had prior to the course start met?

- *Did you find anything surprising?*

Question: How did you find the materials/workbook?

Question: What did you think of the online implementation of the course?

- *How did it compare to other classes you have taken in person?*
- *Assuming both are equally likely, would you prefer to have classes in person or online?*

Question: What did you think about the class size?

Question: Do you think it made a difference that this was a live course rather than a series of recorded lectures?

Appendix H. Chapter 4: ETD models subset by group

The in-person data includes quiz scores, which were added to the model reported in Section 4.3.2.2 (see Table H1). The fitted model has the same random effects structure as the model with the full data (random intercept by participant), though with a different fixed effects structure.

Table H1. In-person classes ETD model with quiz scores

Predictors	**Beta**	**95% CI[1]**
Intercept	59.07	**19.0, 86.2**
Session		
Pre-course	—	—
Post-course	7.50	**0.71, 14.6**
Age of Participant	-0.08	-0.47, 0.45
Quiz Score	0.08	-0.31, 0.54
Composite Language Background Score	0.60	-0.12, 2.31
Classes Missed (out of 12)	-5.51	-16.6, 9.94
Hours Practicing Spanish	1.08	-0.10, 3.80
Gender		
Female	—	—
Male	3.47	-21.5, 24.6
Years of Education	0.320	-2.57, 3.99
Observations	56	

Predictors	Beta	95% CI[1]
ICC	.61	
Marginal R^2 / Conditional R^2	.15 / 0.67	
[1]CI = Confidence Interval, case-based bootstrap with 5,000 samples		

The online data includes Raven's APM and the Friendship scale, which were added to the model reported in Section 4.3.2.3 (see Table H2). The fitted model has the same random effects structure as the model with the full data (random intercept by participant), though was fitted with a different fixed effects structure.

Table H2. Online classes ETD model with Raven's APM and Friendship Scale scores

Predictors	Beta	95% CI[1]
Intercept	75.8	**31.8, 100**
Session		
Pre-course	—	—
Post-course	11.79	**2.14, 20.0**
Age of Participant	-0.16	-0.69, 0.53
Raven's APM Score	0.17	-0.11, 0.52
Change in Friendship Scale Score	0.87	-2.18, 2.37
Composite Language Background Score	-0.09	-0.39, 0.27
Classes Missed (out of 12)	-4.29	-10.2, 3.22

Predictors	**Beta**	**95% CI[1]**
Hours Practicing Spanish	0.27	-0.26, 0.99
Gender		
Female	—	—
Male	2.01	-9.01, 19.8
Years of Education	-1.19	-5.41, 4.68
Observations	56	
ICC	.23	
Marginal R^2 / Conditional R^2	.22 / .40	

[1]CI = Confidence Interval, case-based bootstrap with 5,000 samples

Appendix I. Chapter 4: ETR models subset by group

The in-person data includes quiz scores, which were added to the model reported in Section 4.3.2.2 (see Table I1). The fitted model has the same random effects structure as the model with the full data (random intercept by participant).

Table I1. In-person classes ETR with quiz scores

Predictors	**Beta**	**95% CI[1]**
Intercept	22.3	0.89, 41.7
Session		
Pre-course	—	—
Post-course	16.79	**6.43, 28.2**
Age of Participant	-0.66	**-0.94, -0.39**
Quiz Score	0.50	-0.19, 1.11
Composite Language Background Score	0.27	-1.13, 2.32
Classes Missed (out of 12)	-9.43	**-16.7, -2.33**
Hours Practicing Spanish	1.77	**1.02, 2.53**
Gender		
Female	—	—
Male	-3.72	-16.9, 9.19
Years of Education	0.67	-1.56, 2.93
Observations	56	
ICC	.28	

Predictors	Beta	95% CI[1]
Marginal R^2 / Conditional R^2	.50 / 0.64	

[1]CI = Confidence Interval, case-based bootstrap with 5,000 samples

The online data includes Raven's APM and the Friendship scale, which were added to the model reported in Section 4.3.2.3 (see Table I2). The fitted model has the same random effects structure as the model with the full data (random intercept by participant), though with a different fixed effects structure.

Table I2. Online classes ETR model with Raven's APM and Friendship Scale scores

Predictors	Beta	95% CI[1]
Intercept	76.2	**37.4, 102**
Session		
Pre-course	—	—
Post-course	14.6	**5.36, 23.6**
Age of Participant	-1.39	-1.88, -0.64
Raven's APM Score	0.05	-0.27, 0.43
Change in Friendship Scale Score	1.36	-1.56, 3.14
Composite Language Background Score	-0.27	-0.61, 0.13
Classes Missed (out of 12)	-7.69	-15.4, 0.46
Hours Practicing Spanish	0.94	**0.24, 2.20**

Predictors	**Beta**	**95% CI[1]**
Gender		
Female	—	—
Male	8.49	-5.92, 30.2
Years of Education	2.08	-3.71, 7.51
Observations	56	
ICC	.43	
Marginal R^2 / Conditional R^2	.49 / .71	

[1]CI = Confidence Interval, case-based bootstrap with 5,000 samples

Appendix J. Chapter 5: Demographics of all randomised participants

Across the three rounds of experimental group classes, a total of 81 participants completed the first research session and were allocated to one arm of the study. These 81 participants were randomised into either the Speaking-Writing or Writing-Speaking order of classes, using stratified randomisation based on age, gender, and education level.

Table J1. Demographics of all randomised participants

		Class Order	
Characteristic	**Overall**, N = 81[1]	**SW**, N = 40[1]	**WS**, N = 41[1]
Round			
Round 1	39 (48%)	20 (50%)	19 (46%)
Round 2	21 (26%)	10 (25%)	11 (27%)
Round 3	21 (26%)	10 (25%)	11 (27%)
Participant Age	30 (18, 60)	30 (20, 60)	31 (18, 51)
(Missing)	3	1	2
Handedness			
Left	5 (6.4%)	2 (5.1%)	3 (7.7%)
Right	73 (94%)	37 (95%)	36 (92%)
(Missing)	3	1	2
Gender			
Female	42 (52%)	21 (52%)	21 (52%)
Male	37 (46%)	19 (48%)	18 (45%)

Characteristic	Overall, N = 81[1]	Class Order	
		SW, N = 40[1]	**WS**, N = 41[1]
Prefer to self-describe	1 (1.3%)	—	1 (2.5%)
(Missing)	1	—	1
Education Level			
Primary school	1 (1.3%)	—	1 (2.6%)
Secondary/high school	4 (5.1%)	3 (7.7%)	1 (2.6%)
Vocational training/ professional degree	4 (5.1%)	1 (2.6%)	3 (7.7%)
Undergraduate Student	17 (22%)	9 (23%)	8 (21%)
University Degree	17 (22%)	12 (31%)	5 (13%)
Postgraduate Student	10 (13%)	5 (13%)	5 (13%)
Master's Degree	17 (22%)	7 (18%)	10 (26%)
PhD Student	2 (2.6%)	—	2 (5.1%)
Doctorate	6 (7.7%)	2 (5.1%)	4 (10%)
(Missing)	3	1	2
No. Languages Before Class	4.40 (1, 8)	4.15 (1, 8)	3.92 (1, 8)

[1]n (%); Mean (Range)

Appendix K. Chapter 5: ANT general analysis

Presented below is the model table and resulting interaction plots for the analysis of the ANT raw data (i.e., not the network scores).

Table K1. ANT model table

Characteristic	Beta	95% CI[1]
(Intercept)	621	**573, 669**
Cue Type		
No	—	—
Center	-13	**-26, -0.42**
Double	-21	**-34, -8.4**
Spatial	-66	**-79, -54**
Flanker Condition		
Neutral	—	—
Congruent	13	**0.73, 26**
Incongruent	98	**85, 111**
Session		
Pre-Course 1	—	—
Post-Course 1	4.7	-15, 24
Pre-Course 2	55	**19, 90**
Post-Course 2	46	**3.7, 89**
Group		
Writing-Speaking	—	—

Characteristic	Beta	95% CI[1]
Speaking-Writing	-16	-82, 51
Cue Type x Flanker Condition		
Center x Congruent	-18	**-36, -0.30**
Double x Congruent	-13	-31, 5.2
Spatial x Congruent	-3.9	-22, 14
Center x Incongruent	-4.1	-22, 14
Double x Incongruent	-4.6	-23, 14
Spatial x Incongruent	-11	-29, 7.2
Cue Type x Session		
Center x Post-Course 1	-27	**-46, -8.5**
Double x Post-Course 1	-28	**-47, -8.8**
Spatial x Post-Course 1	-23	**-42, -4.0**
Center x Pre-Course 2	-27	**-48, -6.2**
Double x Pre-Course 2	-29	**-50, -8.5**
Spatial x Pre-Course 2	-54	**-75, -34**
Center x Post-Course 2	-44	**-65, -24**
Double x Post-Course 2	-33	**-54, -12**
Spatial x Post-Course 2	-54	**-75, -33**
Flanker Condition x Session		
Congruent x Post-Course 1	-11	-30, 8.4
Incongruent x Post-Course 1	-40	**-59, -20**
Congruent x Pre-Course 2	-23	**-43, -1.9**

Characteristic	Beta	95% CI[1]
Incongruent x Pre-Course 2	-49	**-70, -28**
Congruent x Post-Course 2	-6.2	-27, 15
Incongruent x Post-Course 2	-66	**-87, -45**
Cue Type x Group		
Center x Speaking-Writing	-15	-33, 2.3
Double x Speaking-Writing	-18	-36, 0.83
Spatial x Speaking-Writing	-3.0	-20, 15
Flanker Condition x Group		
Congruent x Speaking-Writing	-2.8	-20, 15
Incongruent x Speaking-Writing	-14	-32, 3.5
Session x Group		
Post-Course 1 x Speaking-Writing	-13	-41, 15
Pre-Course 2 x Speaking-Writing	-31	-80, 19
Post-Course 2 x Speaking-Writing	-15	-75, 45
Cue Type x Flanker Condition x Session		
Center x Congruent x Post-Course 1	16	-11, 43
Double x Congruent x Post-Course 1	17	-9.6, 44

Characteristic	Beta	95% CI[1]
Spatial x Congruent x Post-Course 1	0.09	-27, 27
Center x Incongruent x Post-Course 1	20	-6.9, 47
Double x Incongruent x Post-Course 1	25	-2.4, 52
Spatial x Incongruent x Post-Course 1	26	-1.6, 53
Center x Congruent x Pre-Course 2	22	-7.4, 51
Double x Congruent x Pre-Course 2	26	-1.6, 55
Spatial x Congruent x Pre-Course 2	25	-3.8, 55
Center x Incongruent x Pre-Course 2	7.3	-22, 37
Double x Incongruent x Pre-Course 2	9.6	-20, 39
Spatial x Incongruent x Pre-Course 2	11	-18, 41
Center x Congruent x Post-Course 2	25	-4.8, 54
Double x Congruent x Post-Course 2	7.4	-22, 37
Spatial x Congruent x Post-Course 2	-14	-44, 15
Center x Incongruent x Post-Course 2	28	-0.21, 52

Characteristic	Beta	95% CI[1]
Double x Incongruent x Post-Course 2	21	-8.9, 51
Spatial x Incongruent x Post-Course 2	29	-0.76, 59
Cue Type x Flanker Condition x Group		
Center x Congruent x Speaking-Writing	1.4	-23, 26
Double x Congruent x Speaking-Writing	-5.9	-31, 19
Spatial x Congruent x Speaking-Writing	-16	-41, 8.8
Center x Incongruent x Speaking-Writing	9.7	-15, 35
Double x Incongruent x Speaking-Writing	14	-11, 40
Spatial x Incongruent x Speaking-Writing	-2.6	-28, 22
Cue Type x Session x Group		
Center x Post-Course 1 x Speaking-Writing	14	-12, 41
Double x Post-Course 1 x Speaking-Writing	17	-9.2, 44
Spatial x Post-Course 1 x Speaking-Writing	10	-16, 37
Center x Pre-Course 2 x Speaking-Writing	19	-9.6, 49

Characteristic	Beta	95% CI[1]
Double x Pre-Course 2 x Speaking-Writing	28	-2.5, 52
Spatial x Pre-Course 2 x Speaking-Writing	-2.2	-31, 27
Center x Post-Course 2 x Speaking-Writing	28	-2.2, 57
Double x Post-Course 2 x Speaking-Writing	28	-2.0, 58
Spatial x Post-Course 2 x Speaking-Writing	4.4	-25, 34
Flanker Condition x Session x Group		
Congruent x Post-Course 1 x Speaking-Writing	7.5	-19, 34
Incongruent x Post-Course 1 x Speaking-Writing	32	5.0, 59
Congruent x Pre-Course 2 x Speaking-Writing	16	-13, 45
Incongruent x Pre-Course 2 x Speaking-Writing	20	-9.4, 49
Congruent x Post-Course 2 x Speaking-Writing	-3.0	-33, 27
Incongruent x Post-Course 2 x Speaking-Writing	24	-5.9, 54
Cue Type x Flanker Condition x Session x Group		
Center x Congruent x Post-Course 1 x Speaking-Writing	3.0	-35, 41

Characteristic	Beta	95% CI[1]
Double x Congruent x Post-Course 1 x Speaking-Writing	-3.3	-41, 34
Spatial x Congruent x Post-Course 1 x Speaking-Writing	8.1	-29, 46
Center x Incongruent x Post-Course 1 x Speaking-Writing	-4.6	-43, 33
Double x Incongruent x Post-Course 1 x Speaking-Writing	-29	-67, 8.9
Spatial x Incongruent x Post-Course 1 x Speaking-Writing	-19	-57, 19
Center x Congruent x Pre-Course 2 x Speaking-Writing	2.9	-38, 44
Double x Congruent x Pre-Course 2 x Speaking-Writing	-30	-71, 11
Spatial x Congruent x Pre-Course 2 x Speaking-Writing	-3.2	-44, 38
Center x Incongruent x Pre-Course 2 x Speaking-Writing	-10	-52, 31
Double x Incongruent x Pre-Course 2 x Speaking-Writing	-24	-66, 17
Spatial x Incongruent x Pre-Course 2 x Speaking-Writing	13	-29, 54
Center x Congruent x Post-Course 2 x Speaking-Writing	7.1	-35, 49
Double x Congruent x Post-Course 2 x Speaking-Writing	8.2	-34, 50
Spatial x Congruent x Post-Course 2 x Speaking-Writing	31	-11, 73

Characteristic	**Beta**	**95% CI[1]**
Center x Incongruent x Post-Course 2 x Speaking-Writing	-23	-65, 20
Double x Incongruent x Post-Course 2 x Speaking-Writing	-28	-71, 14
Spatial x Incongruent x Post-Course 2 x Speaking-Writing	-16	-59, 26
No. Obs.	46,801	
ICC	.53	
Marginal R^2 / Conditional R^2	.08 / .57	

[1] CI = Confidence Interval, case-based bootstrap with 5,000 samples

Figure K1. ANT model cue type by flanker condition interaction

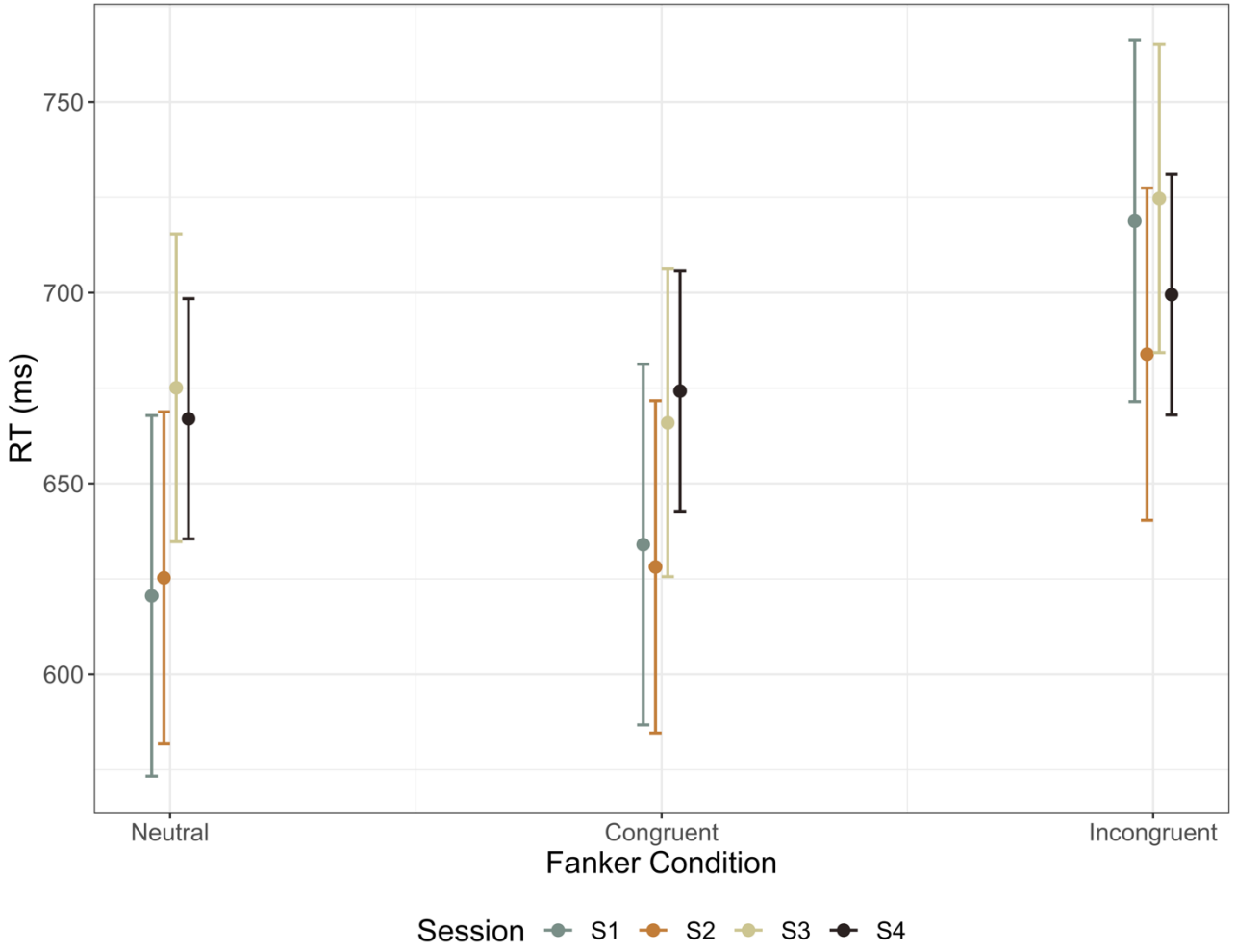


Figure K2. ANT model cue type by session interaction

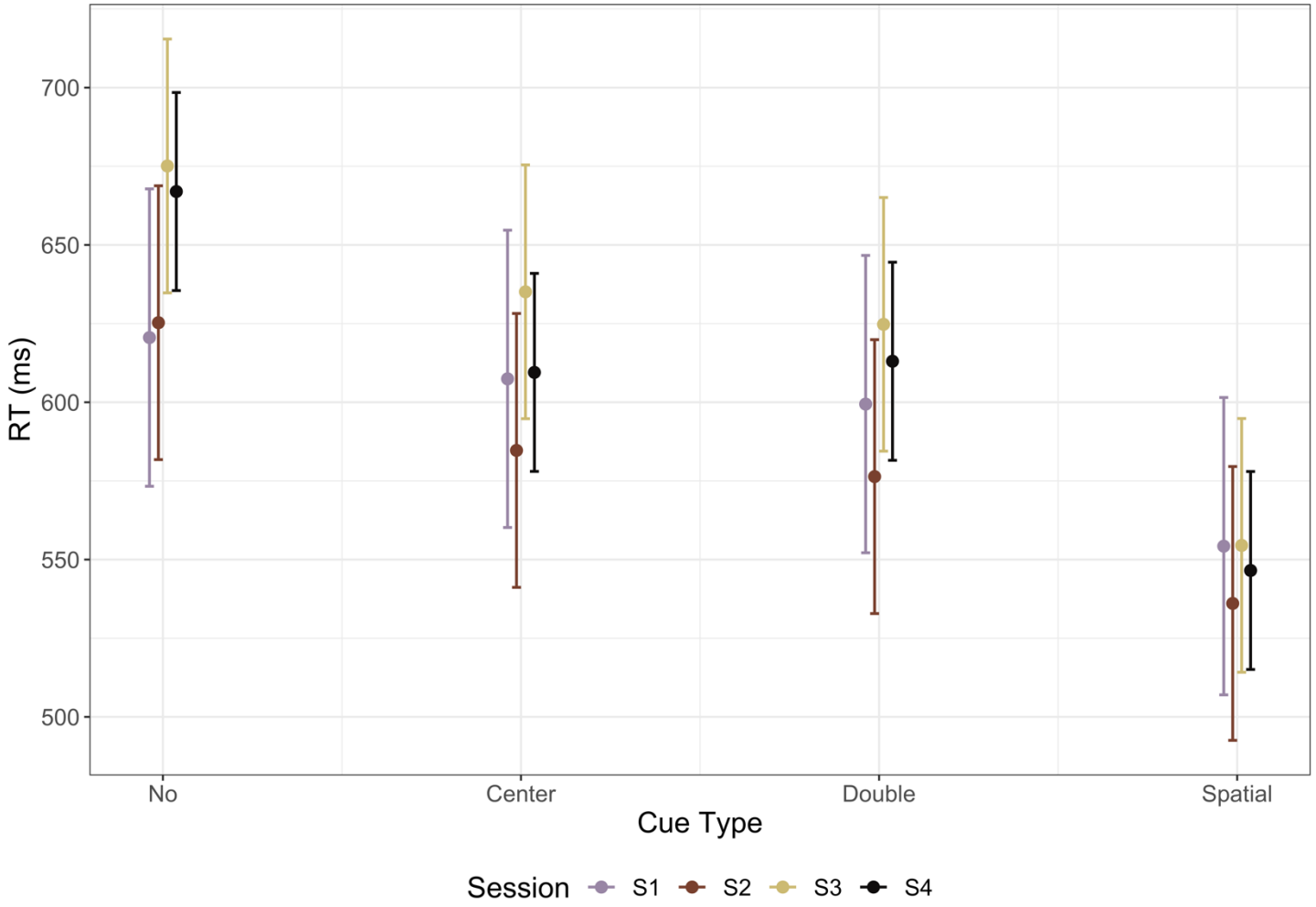


Figure K3. ANT model flanker condition by session interaction

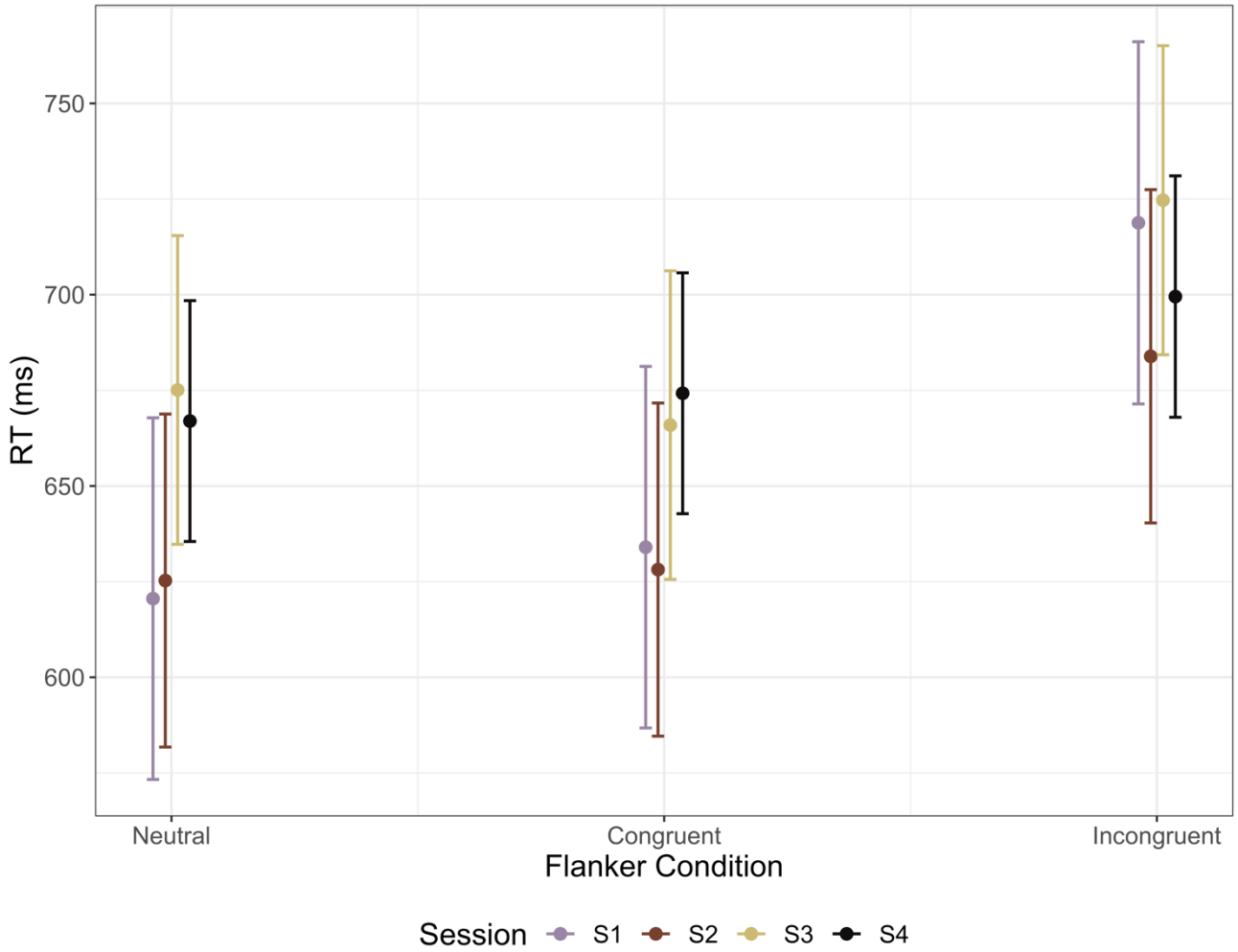
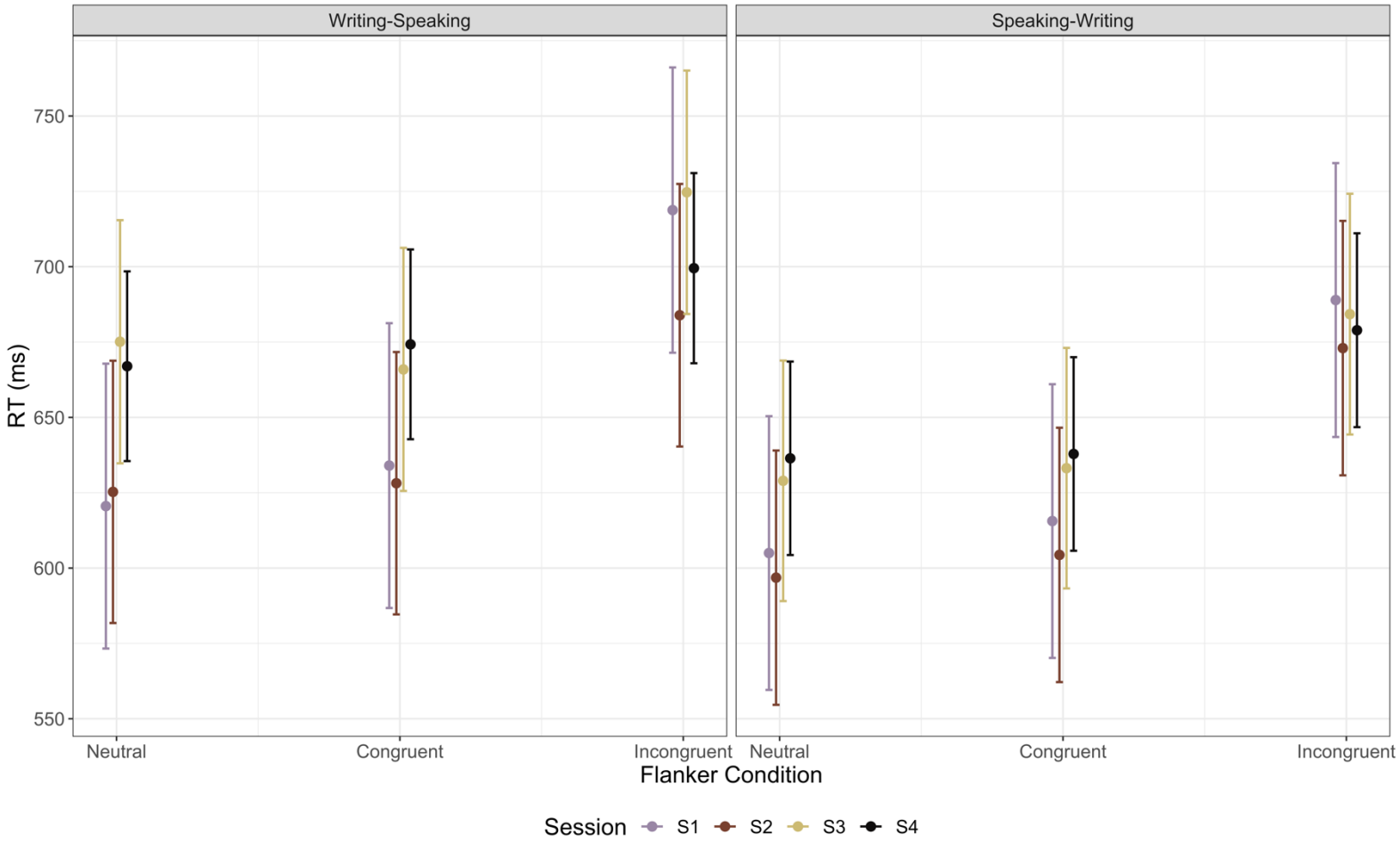
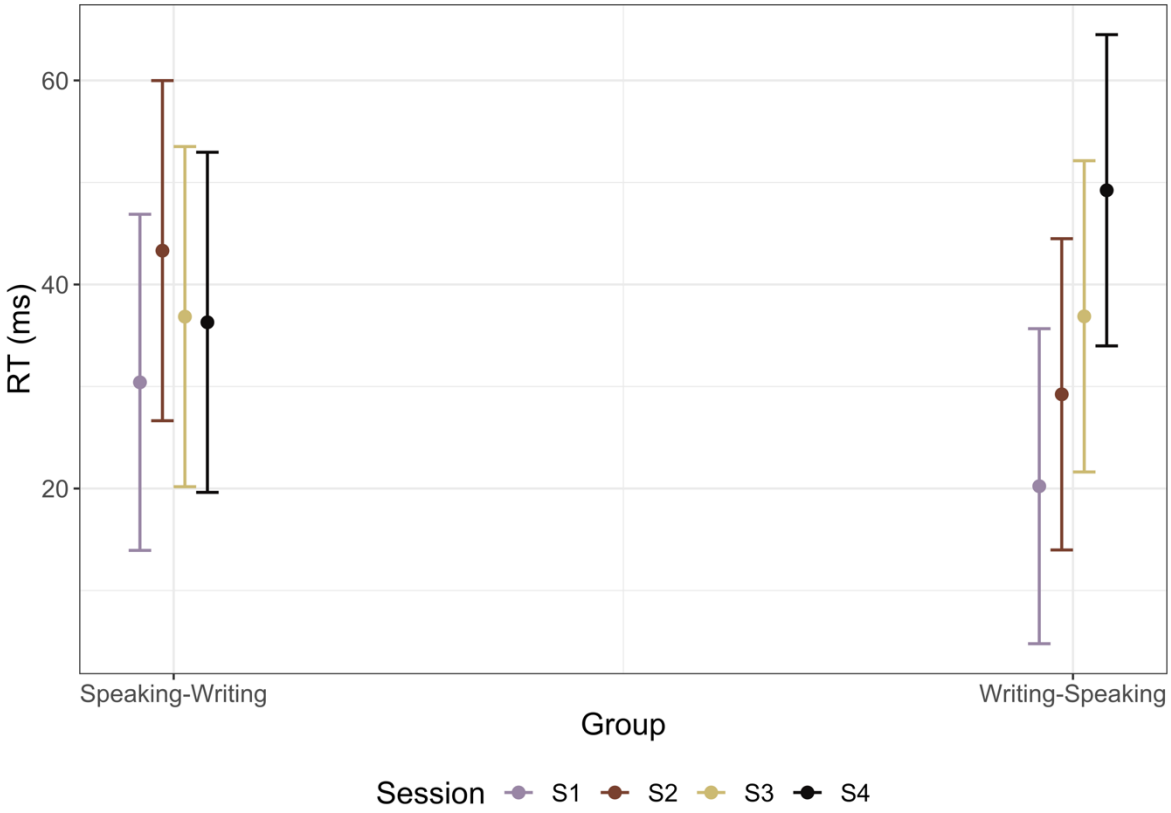


Figure K4. ANT model flanker condition by session by group interaction



Appendix L. Chapter 5: Alerting model interaction visualisation

Figure L1. Alerting network model group by session interaction



Appendix M. Chapter 6: Language attitude questions

It is important for members of my community to know Scots.

- Disagree Strongly
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

Schoolchildren here should have the opportunity to study Scots.

- Disagree Strongly
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

I would encourage others like me to speak Scots.

- Disagree Strongly
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

Learning Scots demands less intelligence than other languages I know.

- Disagree Strongly
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

There is no good reason for anyone to need to use Scots in the 21st century.

- Disagree Strongly
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

Scots has a community with its own culture.

- Disagree Strongly
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

Scots is important for the cultural heritage of Scotland.

- Disagree Strongly
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

Scots is important for my own cultural heritage.

- Disagree Strongly
- Disagree Moderately
- Disagree Slightly
- Agree Slightly
- Agree Moderately
- Agree Strongly

In 50 years time, I think Scots will be spoken by

- More people than now

- About the same number of people as now
- Fewer people than now

In 50 years time, I would like for Scots to be spoken by

- More people than now
- About the same number of people as now
- Fewer people than now

Please place a mark on the line below each of the following statements according to the amount of your agreement or disagreement with each item. There is no right or wrong answer.

	Strongly Disagree 0 10 20 30 40 50 60 70 80 90 100 Strongly Agree
Scots is a language in its own right.	
Scots is difficult to learn.	
It's really good for your brain to use/speak Scots.	
Scots should be spoken but not written.	
Scots is appropriate to use in official documents/communications.	

Note: The Likert scale responses were presented horizontally to participants in the online form.

Appendix N. Chapter 6: Feedback session interview schedule

Question: How did you find the course?

- Thoughts?
- Feedback?
- What did you like? What would you change?
- Was there anything you would like or more less of?

Question: If you could use one word to describe the course, what would it be?

Question: How was the intensity of the course? Too little, too much or, just right?

- In terms of meetings per week and time spent practicing the language outside of class

Question: How did you find the length of the course? How about the length of the individual classes?

Question: What did you think about the frequency of meetings/classes?

- Was three times a week too often, too little or, just right?
- Did you have homework assigned and did you do it?

Question: How many hours of learning a week would you ideally have for a language course? What would be your ideal intensity?

- How many hours of independent work?

- What kind of independent work would you be most likely to participate in?

Question: Were the expectations that you had prior to the course start met?

- Did you find anything surprising?

Question: Did the course change your mind or opinions in any way?

Question: How did you find the materials/workbook?

Question: Did you feel comfortable trying to pronounce things in Scots or does it feel strange? If it feels strange why?

- Is it the same when you try to pronounce things in German or Spanish or French? Or does it feel different?
- Scots group – Do you feel comfortable hearing other people trying to pronounce things in Scots or does it feel strange? If it feels strange why?
    - Is it a difference when it is English native speakers or native speakers of other languages?

Question: What did you think of the online implementation of the course?

- How did it compare to other classes you have taken in person?
- Assuming both are equally likely, would you prefer to have classes in person or online?

Question: What did you think about the class size?

Question: Do you think it made a difference that this was a live course rather than a series of recorded lectures?

Appendix O. Chapter 6: Additional visualisation of language attitude questions

Additional visualisations of language attitude questions. In Figure P1 the overall LE group attitudes towards the Likert-style questions are visualised. In Figure P2 the change in response to the VAS style attitude questions by group is visualised.

Figure O1. Overall LE group responses to the language attitude questions asked on a Likert-style scale

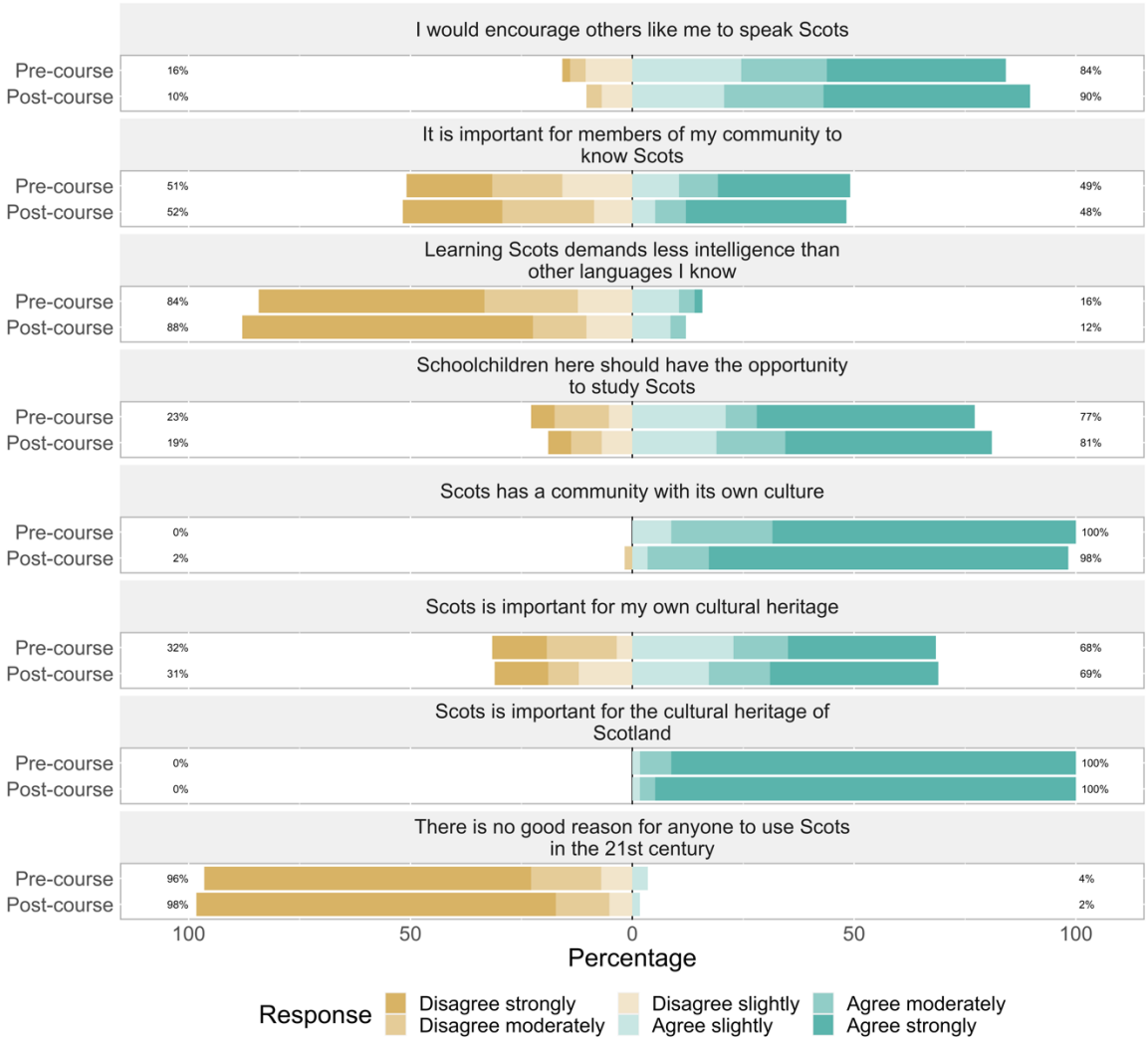
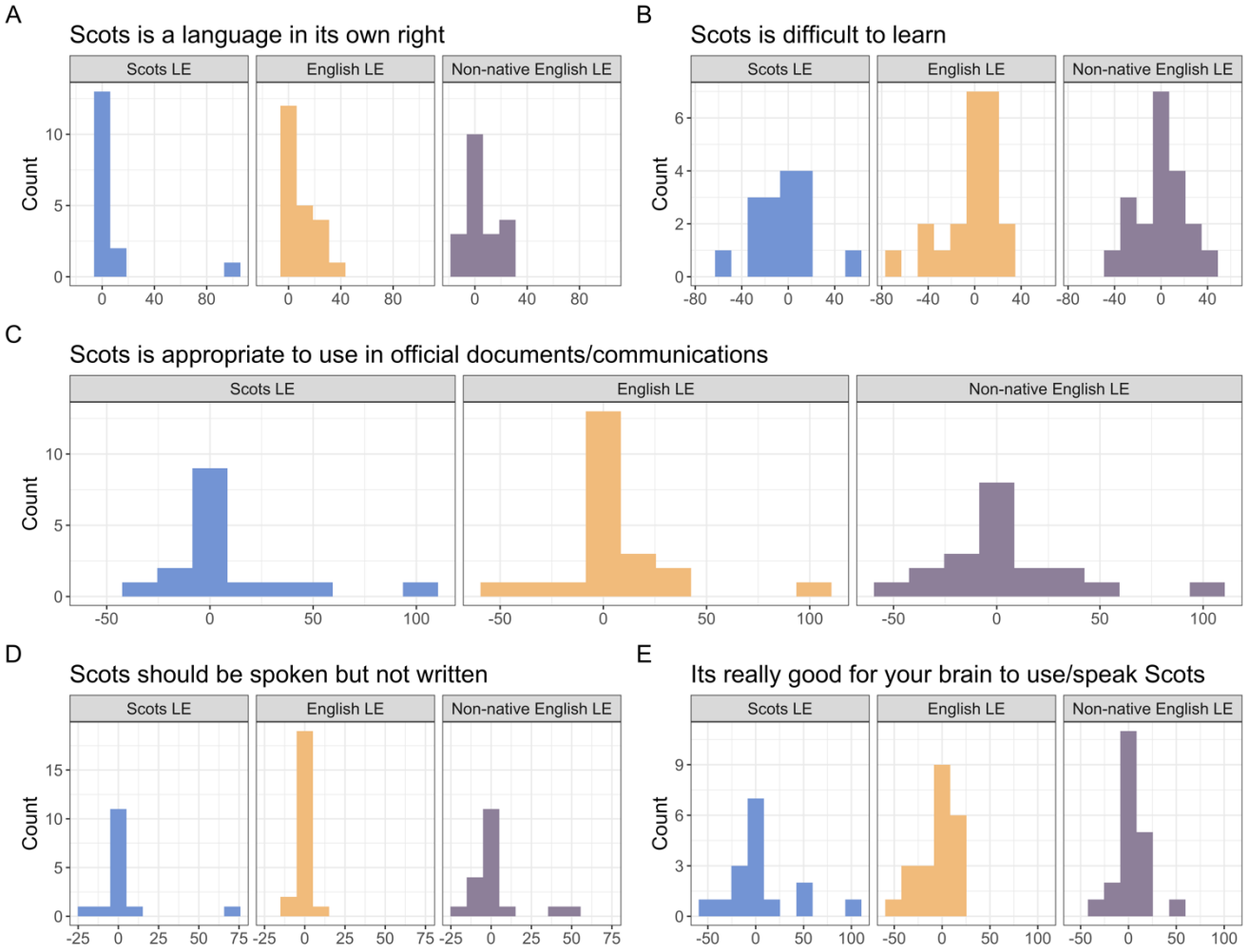


Figure O2. Difference in responses to language attitude questions as VAS



Appendix P. Chapter 7: Data descriptions

Table P1. Participants and datasets

Participant Name	Age	Dataset	Source
Aad	75	Dutch	Exit interview at post-test
Avery	65	Scottish	Spanish classes - post-course focus group
Cornelia	67	Dutch	Email to the teacher
Daisy	64	Scottish	Spanish classes - post-course focus group
Hedwig	80	Dutch	Drop-out interview
Hendrika	85	Dutch	Focus group
Hugo	55	Scottish	Scots classes - questionnaire after the course
Isaac	67	Scottish	Spanish classes - feedback after testing sessions
Jaap	62	Dutch	Exit interview at post-test
Jude	63	Scottish	LFA classes - questionnaire after the course
Matilda	67	Scottish	Spanish classes - questionnaire after the course
Penelope	63	Scottish	Spanish classes - post-course focus group

Phoebe	54	Scottish	Scots classes - feedback after testing sessions
Poppy	56	Scottish	Scots classes - feedback after testing sessions
Renske	71	Dutch	Spontaneous feedback during lessons
Scarlett	57	Scottish	Spanish classes - questionnaire after the course
Sien	86	Dutch	Daily diary
Truus	69	Dutch	Daily diary

Table P2. Sources of data

Source	Number of participants	Average amount of data	Key themes
Daily diary	Dutch: 26	Dutch: 57 dairy comments of about 2 lines (text in diary)	Dutch: Difficulties with technology in order to do the homework; course enjoyment.
Drop-out interview	Dutch: 10 Scottish: 6	Dutch: 5 minutes per person (phone) Scottish: 4 sentences (email)	Dutch: Technology; course difficulty; healthy issues. Scottish: Health issues; scheduling difficulties; technical issues.
Email to teacher	Dutch: 1	Dutch: 5 lines (email)	Dutch: Social contact.
Exit interview at post-test	Dutch: 16	Dutch: 5 minutes per person (Zoom)	Dutch: Course enjoyment; social contact; technology; daily purpose.
Feedback after testing session	Scottish: 6	Scottish: 8 minutes (Zoom)	Scottish: daily purpose, social contact; course design; online learning
Focus group	Dutch: 10 (2 groups of 5; recruited through senior networks in the Netherlands)	Dutch: 2 hours (Google Meet)	Dutch: Course design.

Post-course focus group	Scottish: 15 (9 groups total ranging from 4 to 8 people including all ages)	Scottish: 33 minutes (Zoom)	Scottish: Impact of the teacher; course enjoyment; technology; social contact; patience; daily purpose; course design; online learning
Spontaneous feedback during lessons	Dutch: 11	Dutch: 3 comments per class (33 in total)	Dutch: Technology; patience; social contact.
Questionnaire after the course	Dutch: 16 Scottish: 30	Dutch: 3 sentences (text in questionnaire) Scottish: 3 sentences (text in questionnaire)	Dutch: Technology; social contact; meaning; daily purpose. Scottish: Impact of the teacher; course enjoyment; technology; social contact; patience; daily purpose; online learning