Statistical Power for Small Effect Sizes:
An investigation of backward priming in Mandarin-English bilinguals

Xiao Xiao Li
B.A.&Sc. McGill University, 2018

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF ARTS

in the Department of Linguistics

University of Victoria

We acknowledge and respect the Lək̓ʷəŋən (Songhees and Esquimalt) Peoples on whose territory the university stands, and the Lək̓ʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Statistical Power for Small Effect Sizes:
An investigation of backward priming in Mandarin-English bilinguals

Xiao Xiao Li
B.A.&Sc. McGill University, 2018

Supervisory Committee

Dr. John Archibald, Department of Linguistics
SUPERVISOR

Dr. Sonya Bird, Department of Linguistics
DEPARTMENTAL MEMBER

Abstract

Backward priming, or L2 to L1 priming, is a small but important effect for understanding the structure of the bilingual lexicon. A meta-analysis of priming in bilingual populations has shown that while the backward priming effect is quite small, it is qualitatively but not quantitatively different from the forward (L1 to L2) priming effect (Wen & Van Heuven, 2017). The empirical evidence for this view has come from various groups of bilinguals, including Japanese-English (Nakayama et al., 2016) and Korean-English (Lee et al., 2018) bilinguals, but not yet with Mandarin-English bilinguals: In this population, the effect is inconsistently significant. In response to this, researchers have raised the question of whether the existing studies were underpowered, given the small backward priming effect. Using a simulation-based power analysis, I show that this is most likely the case, as roughly 5400 observations per condition are necessary to detect a small backward priming effect. Previous work collected an average of 453 observations per condition, making it very unlikely for their statistical tools to be able to detect the effect. Based on this, I recommend that future work in this field conduct power analyses *a priori*, using the results as a guideline rather than a strict criterion for adequate power. Adopting this practice can help make experiments more replicable and future work in this direction is crucial for developing our understanding of the structure of the mental lexicon.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. John Archibald, for all his support, wisdom, and invaluable feedback through the process of writing this thesis. This project started as a very different idea involving human participants, and I am so grateful to have had John's guidance through the process of evolving it to its present state. His constant encouragement and gentle nudges to keep this project within the scope of an MA were crucial for finishing the thesis, and I feel very privileged to have been able to learn from him. John, thank you for your endless patience and being so generous in sharing your time with me, I know that I have grown so much as a researcher through our collaborations.

I would also like to thank my committee member Dr. Sonya Bird for her insights on all the different versions of this thesis. Her contributions were critical for shaping the structure of the final product, and I am so thankful to her for sharing her time, wisdom, and experience with me. Sonya, thank you for pushing this thesis into domains that I could not have imagined; some of the parts of this thesis that I am most proud of are a direct result of your generous and precise feedback.

Additionally, I am enormously grateful to my external examiner, Dr. Christine Shea, for both her thoughtful feedback on the thesis and for her willingness to tackle challenges across time zones to make the defense happen. Thank you for your valuable insights and kindness through the process, as well as the spirited discussion.

I would also like to express my gratitude for the University of Victoria Graduate Fellowship, the Linguistics Research Fund, the Faculty of Graduate Studies Travel Grant, the CUPE 4163 Conference Award, and the University of Victoria Graduate Award for their financial support over the years. Because of these, I had the opportunity to travel to conferences and shake the hands of the researchers that previously only existed in papers that I had cited, which enormously enriched my graduate studies.

My time at UVic was also greatly shaped by the other professors in the department that I was lucky enough to learn from. Alongside Dr. Archibald and Dr. Bird, I would also like to thank (alphabetically) Dr. Li-Shih Huang, Dr. Hua Lin, and Dr. Leslie Saxon for showing me the wide diversity of what linguistics can be and contribute to the world, as well as teaching with empathy, grace, and kindness. I would also like to thank Dr. Alexandra D'Arcy for taking me on as a research assistant in the Sociolinguistics Research Lab and giving me my first taste of what goes into a large-scale research project. Thank you also to the past and present members of the Lx Phonology Lab, led by Dr. Archibald, Eloisa Cervantes Galindo, Jie Deng, Martin Desmarais,

Willem Kuun, Matthew Patience, and Junyu Wu, it has been a delight to share in each other's discoveries and celebrate each other's milestones. Additionally, I would like to thank our department secretaries, Emmanuelle Guenette, Jenny Jessa, and Maureen Kirby, for all their help over the years in making any administrative headache disappear with a kind email and a warm smile.

Thank you also to my wonderful cohort, Caroline Allen, Emmanuelle Buaillon, Rain Mao, Christiana Moser, and Reza Shirani, as well as classmates in other cohorts, including Monica Connors, Paul Emme, Kate Garrison-Riker, Amber Goss, Keun Kim, Tess Nolan, Mona Sawan, and Yi Wang, for all the laughter and commiseration shared over the years. To all of you, thank you for always being there to talk through research puzzles, lend a sympathetic ear, and gossip over spring rolls. Grad school can sometimes feel deeply isolating, and the camaraderie and care that you showed me ensured that I never felt like I was an island, despite living on an island.

Outside of linguistics, I would like to thank friends near and far for their kindness and generosity over the years, particularly in letting me talk at them about linguistics and statistics until I had talked my way out of writer's block. To Aaron Acosta, Tavish Armstrong, Camille Beaudoin, Sid Boegman, Alexe AJ Daclag, Heather McTaggart, Kaylyn Olynyk, Jake Ryan, Patrick Ryan, Koby Song-Nichols, Galen Voysey, and Trevor Walker, thank you for your endless encouragement and for always believing in me, especially when I didn't. And to Mattias Graham: Thank you for the bowls of soup during every critical moment of this process, for your optimistic reminders of the light at the end of the tunnel, and for always knowing what song to play.

Finally, I would like to thank my mother, Zhen Wang Brown, for her unconditional love and care, and for inspiring my interest in the bilingual lexicon many years ago. Thank you for passing down your determination, your wit, and your sense of humour, all of which were necessary to finish this thesis. Every opportunity I have been lucky enough to pursue is because of you.

Dedication

To Zhen Wang Brown

With love and admiration

Introduction

This project started with a straightforward question: How do multiple languages share the same space in one mind? How does a bilingual keep track of these languages? And how do we test this physiological phenomenon that we cannot physically examine? These questions led to this investigation into the bilingual lexicon, more specifically, bilingual lexical access. When more than one language is involved, how might one language influence the other? To explore these possibilities, I focused on how behavioural tasks have been used to examine online lexical processing, specifically lexical decision tasks, which are tasks where a participant is asked to determine whether a target is a real word. These tasks have been used for decades to explore this topic (Dannenbring & Briand, 1982; Wen & van Heuven, 2017), and are often combined with priming (i.e., presenting a brief stimulus to the participant before the target, with or without concealment). This version, called a priming lexical decision task, is particularly useful to study how accessing one of a bilingual's languages can be used to influence access to the other. For example, this task has been used to show how a bilingual's second language can be used to speed up access to their first language across various language pairs (Beauvillain & Grainger, 1987; Dimitropoulou et al., 2011; Gollan et al., 1997; Lee et al., 2018; Nakayama et al., 2016), however, there seems to be inconsistent evidence for this effect in Mandarin-English bilinguals.

To better understand the discordant results with Mandarin-English bilinguals, Chapter One of this thesis will review the topic of masked priming and lexical access, showing that while the forward priming effect (using a bilingual's first language to prime their second) is strong and consistent, the backward priming effect (L2 to L1 priming) is

more elusive. While it has been demonstrated in late bilinguals including Greek-English bilinguals (Dimitropoulou et al., 2011), Korean-English bilinguals (Lee et al., 2018), and Japanese-English bilinguals (Nakayama et al., 2016), there is a noticeable lack of consistent significant results reported from Mandarin-English bilinguals despite the number of experiments investigating this population (Chen et al., 2014; Jiang, 1999; Jiang & Forster, 2001; Luo et al., 2013; Wang & Forster, 2015; Witzel & Forster, 2012; Xia & Andrews, 2015). Only two of the fifteen studies included above have found a significant effect of backward lexical priming, which leads to the question of why Mandarin-English bilinguals differ from the other bilingual populations in this regard.

In Chapter Two of this thesis, various reasons for the discrepancy will be explored, ranging from issues with using adequately-proficient bilinguals (Nakayama et al., 2016) or not using an appropriate priming protocol (Lee et al., 2018) to a critique of the statistical methodology of the previous work (Lee et al., 2018; Nakayama et al., 2016; Wen & van Heuven, 2017). A review of proficiency testing in the Mandarin-English bilingual literature will show that while the participants might be a little under the proficiency criteria established by Nakayama et al., proficiency alone is not enough to explain the lack of significant results. Similarly, a review of the literature shows that the issues with the priming protocol raised by Lee et al. were not present in the previous work, leading to the question of statistical power.

Having assessed other potential reasons for the inconsistency between Mandarin-English bilinguals and other cross-script bilinguals, the possibility of statistical power being responsible will drive the present project. Defined as the ability of a test to detect a given effect under the assumption that the effect exists in the population, statistical power

is strongly influenced by the number of observations in any given study. We can imagine this as detecting a signal in the noise: if the data are very noisy, or if the signal is very weak, we need more data to be sure of what we are seeing. In practice, this often results in researchers collecting more data to increase statistical power, particularly when attempting to detect small effects (akin to a weaker signal). This is the case with the backward priming effect, as a meta-analysis of the bilingual lexical priming literature showed the average backward priming effect to be quite small at 11.38ms (Wen & van Heuven, 2017). To compare, the average forward priming effect in the meta-analysis was 44.35ms, or almost three times as large as the backward priming effect. This leads to the question of what a properly-powered experiment examining this small effect size might look like. How many observations are necessary to detect such a small effect? In the field of psycholinguistics, Brysbaert and Stevens (2018) have proposed that a minimum of 1600 observations is necessary to detect a large effect, but it is not clear how many more are needed when the effect is small. This current project aims to answer this question, and test whether the recommendation of 1600 observations is adequate.

To answer this, Chapters Three and Four will present the statistical background necessary to understand how to conduct a power analysis and discuss the methodology undertaken in this project. As the inconsistent results are based on detecting a statistically significant effect, we must first review what it means for a test to reach statistical significance. This follows a recommendation from Lindsay (2020) for researchers to review their understanding of inferential statistics and the analytical tools typically used in psychology, and so Chapter Three will discuss the basics of Null Hypothesis Significance Testing (NHST), including statistical tests, calculating *p*-values from

statistical tests, and how this affects the statistical power of a given experiment. The design of experiments typically used in backward masked priming experiments (i.e. repeated measures experiments) will also be reviewed, as this influences the way that statistical power can be calculated. Chapter Four will review how to calculate statistical power for repeated measures experiments and explain why the current project has chosen to use a simulation-based method of calculating power to address the question of how many observations is enough to detect a small effect.

The results of the power analysis will be presented in Chapter Five, showing that somewhere in the neighbourhood of 5400 observations is necessary to detect an 11.38ms priming effect. This is much higher than the 1600 recommended by Brysbaert and Stevens (2018), however, this project will echo their recommendation that researchers do not use this number as a strict cut-off. Rather, the results are intended to guide researchers' intuitions about what might or might not be enough to detect a small effect. The key takeaway for researchers should be that *a priori* power analyses are crucial to incorporate into their experiments: Each experiment is different, and there is no way of determining what will be sufficient in each case without knowing the specific details of the effects that they are trying to detect and the populations that they are sampling from.

Chapter Six will contextualize the results of the power analysis and explore how the results might fit into an open science framework. This will include a discussion of what it might mean for models of the mental lexicon if it is the case that previous work with Mandarin-English bilinguals was underpowered, as well as some limitations of the analysis. There will also be a short discussion of Bayesian statistics to explore the possibility that there really is something different about Mandarin-English bilinguals that

is causing them to behave differently from other cross-script bilinguals. The chapter will conclude with a discussion of how these results should encourage us to adopt some key practices of open science, as well as think critically about the statistical tools that we use, including in cases where it might not be possible (or feasible) to reach adequate statistical power.

Chapter One: Masked priming and lexical access

When a Mandarin-English bilingual hears English, it is easy to see why they would activate the Mandarin translation; this is the process through which adults start learning a new language. But when a Mandarin-English bilingual hears a word in Mandarin, would they activate the English translation? Studies on Greek-English (Dimitropoulou et al., 2011), Japanese-English (Nakayama et al., 2016), and Korean-English (Lee et al., 2018) bilinguals suggest that we would see such 'backwards' activation. Yet the literature on Mandarin-English bilinguals is inconsistent and inconclusive. Why does this happen? To better understand this inconsistency in the literature, this chapter aims to provide background and context about the backward priming effect and what we currently know about bilingual lexicon. First, I will discuss the types of bilinguals included in this project, and then provide a general overview of the methodology that is used to investigate the bilingual lexicon with a focus on backward masked priming tasks. I will then review our current understanding of the architecture of the bilingual lexicon, highlighting the inconsistent results with Mandarin-English bilinguals in the literature. This chapter is intended to give background context for understanding potential reasons for the discrepancy in the next chapter.

1.1 What kinds of bilinguals?

This project focuses on Mandarin–English bilinguals as a subset of cross-script bilinguals (bilinguals whose two languages use different scripts, such as Korean-English bilinguals), often broadly differentiated by the age of acquisition for the second language (AoA). Most of the bilinguals studied in the Mandarin–English literature are late

bilinguals (bilinguals with one language that they have known from birth, who have acquired their second language after childhood), and unless otherwise specified, 'bilingual' in this thesis refers to late bilinguals. Some work also refers to early bilinguals (bilinguals with one language that they have known from birth, who started acquiring their second language during childhood) and simultaneous bilinguals (bilinguals who started learning two languages from birth), but these exceptions will be noted (Li, 2000). To indicate which language was learned first for early and late bilinguals, a hyphen will separate their first language (L1) and their second language (L2). For example, a Mandarin speaker from birth who learned English later would be referred to as a Mandarin–English bilingual and not an English–Mandarin bilingual.

Defining bilingualism is an immense and complex topic that goes beyond the scope of this project, however, the working definition of bilingualism in this project will be a person who makes regular use of and has the ability to function conversationally in two languages (Li, 2000). This is a measure of fluency, rather than proficiency, which focuses on whether bilinguals readily access and produce their second language. Operationalizing fluency is typically done by recording measures like speech and articulation rate or the number of pauses that a given bilingual makes in natural speech, however, there is no assessment of declarative knowledge. This is in contrast with proficiency, which is typically measured in performance on tests assessing language production and comprehension, and emphasizes the bilingual's ability to recall specific phonological, morphological, and/or syntactic knowledge about their second language (Thomson, 2015). Additionally, defining and determining proficiency are broad and complex topics involving many factors, for example, the varying degrees of familiarity

bilinguals have with their languages in different contexts for bilinguals who use one language at work but another among family and friends. To use the construct of proficiency to define whether any given study's participants were bilingual enough to be included in this project would make applying the criterion impossible. Fluency, on the other hand, is much easier to operationalize, as it focuses on use of language and ease of production. Since this project aims to investigate the bilingual lexicon and lexical access, using fluency as a way of assessing bilingualism also helps keep the focus on how bilinguals access and use their second language, rather than assessing whether they have acquired specific language skills. Defining bilingualism this way generally excludes beginners and those who are just starting to learn their second language but encompasses highly proficient bilinguals who work and maintain social relationships in their second language and even some intermediate speakers who might not have mastery over complex grammatical structures, but nonetheless communicate in their second language readily and easily. For example, a bilingual who is fluent but not necessarily highly proficient could be an immigrant who learned their L2 as an adult to function in their new country, but who still works primarily in their first language. They would probably need support to engage with highly formal, academic text in their second language, but they could read a newspaper article with ease and are nonetheless fluent in their L2 on account of using it in their day-to-day life. This type of participant would be considered fluent enough to be included in this project, however, determining whether they would be proficient enough is out of the scope of this thesis. Additional issues with operationalizing proficiency and proficiency testing will also be discussed in later chapters.

1.2 The masked priming paradigm

Furthermore, this project focuses on Mandarin–English bilinguals' performance in masked priming lexical decision experiments. This is a type of behavioural task that is non-invasive and takes very few resources to conduct, making them a useful tool for investigating phenomena like lexical activation and access, which are important issues in the overall puzzle of determining the structure of the bilingual lexicon. Specifically, this project will focus on masked priming in lexical decision tasks, as they have been widely used to determine the relationships between words in a bilingual's first and second languages, and their ubiquity makes it relatively easier to compare results between experiments. To help do this, the next section will provide an overview of the different portions of a masked priming task, as well as directional priming.

Priming tasks are a general umbrella of behavioural tasks in which a participant is presented with a stimulus for a brief amount of time (the prime) before another stimulus (the target) that they are asked to respond to. Studies using these tasks usually use a set of primes that have a specific relationship to a set of targets (the critical primes), and a set of primes that are matched with the critical primes on form characteristics such as length and initial letter, but do not share a relationship with the targets (the control primes). In the analysis, participants' response times to the critical primes are compared to their response times to the control primes to determine if there was an effect of the prime, and this difference in response time is reported as the priming effect (i.e. how much the critical primes were able to influence the response to the targets).

Masked priming adds to this by using nonsense stimuli (i.e. masks) to conceal the existence of the prime from the participant. In this type of protocol, masks are presented

either before the prime (a forward mask), after the prime (a backward mask), or both before and after the prime with the aim of activating partial awareness of the prime while avoiding the participant's conscious recollection of it. This is done so that researchers can examine unconscious processes of lexical processing in behavioural tasks and minimize the risk of the participant altering their behaviour based on the prime[1]. The type of stimuli used as a mask depends on the critical stimuli in any given experiment, as both prime and mass should be the same type of stimulus (eg. an auditory prime would necessarily need an auditory mask). In this project, the primes are always lexical items (i.e. strings) and so the masks are nonsense strings of letters and symbols: For example, in a masked priming study, Wen and van Heuven (2018) displayed a 500 millisecond (ms) forward mask "$@#$£@£%", then the prime for 59ms, then a 24ms backward mask "%$%£@£$#". In this protocol, the use of both masks contributed to the participants reporting not being aware of the prime, despite demonstrating a 17ms difference between response times to critical and control primes (i.e. a priming effect of 17ms). A hypothetical unmasked version of this experiment would be one where only the prime is displayed for 59ms, and then the target presented immediately after.

Masked priming tasks can be described in terms of the *durations* of the prime and any masks, but this is more commonly referred to as stimulus-onset asynchrony (SOA) and inter-stimulus interval (ISI): SOA is the time between the *onset* of the prime and the onset of the target, whereas the ISI refers to the time between the *offset* of the prime and onset of the target (Lee et al., 2018; Wen & van Heuven, 2017), demonstrated in *Figure*

[1] the rationale for priming as a technique will be discussed later in this chapter

1.1. These definitions are crucial for comparing experimental methodologies and results between studies using masked priming tasks.

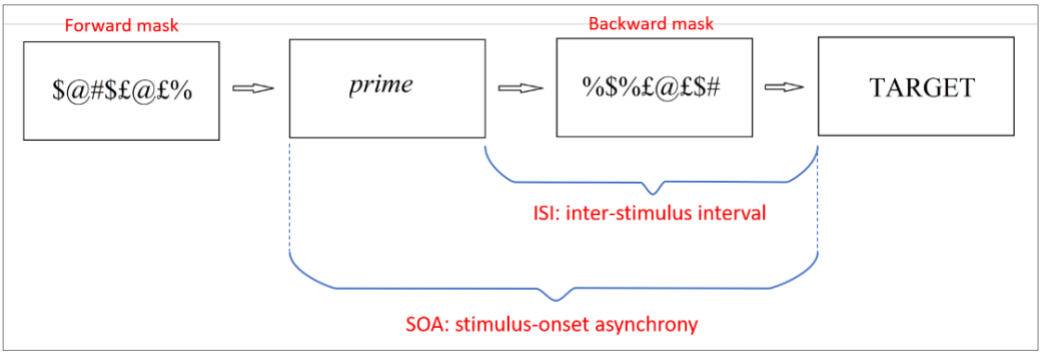

Figure 1.1: elements of a masked priming experiment

For example, in their previously referenced 2018 study, Wen and van Heuven had an SOA of 83ms (59ms prime + 24ms backward mask), but an ISI of 24ms (backward mask only). Additionally, masked priming studies often vary the length of the these different parts of a trial: In the classical three-field masking protocol used in Forster and Davis (1984), there is no backward mask, equivalent to an ISI of 0. In a meta-analysis of masked priming lexical decision experiments, Wen and van Heuven (2017) found that ISI in this field ranged from 0ms to 200ms, with some studies doing so to specifically investigate the effect of varying SOA on the priming effect (Jiang, 1999; Lee et al., 2018; Wang & Forster, 2015). Some preliminary results indicate that longer SOAs might be necessary in experiments with lower-proficiency bilinguals (Lee et al., 2018), however, there is not yet enough data to determine if there exists a single ideal ISI for all masked priming experiments, or if different ISIs may be used, depending on the phenomenon examined.

These parameters of a masked priming task could be manipulated in both studies with monolinguals and bilinguals, but with bilinguals, the languages of the prime and

target can also be varied to demonstrate cross-language priming. Priming from the L1 to the L2 is referred to as forward priming, and from the L2 to the L1 is referred to as backwards priming (Lee et al., 2018). The cross-language priming effect can be achieved with a variety of methods: Using direct translation primes in a different language from the targets is the most straightforward, and employed by many researchers (for example, Dimitropoulou et al., 2011; Jiang, 1999; Jiang & Forster, 2001; Nakayama et al., 2016). Alternatively, another method for manipulating priming direction can be to use primes and targets that have no relationship in the presented language, but when translated into another language, they share phonological or semantic properties. This was used in a study by Thierry and Wu (2007) where they used unrelated primes and targets in English, but the Mandarin translation of each prime and target pair shared a first character, creating a phonological and semantic prime. For example, the use of the prime 火腿 'hot dog' /huo͡.tue͡/ for the target 火车 'train' / huo͡.tʂʰə͡/. The character 火 'hot' adds similar semantic information in both of these compound words, referring to the flame used to cook a hot dog or the flame of a coal-powered train, but the English words 'hot dog' and 'train' have no semantic overlap, creating a hidden prime. This can also be modified to focus on phonological priming only, and not semantic priming: Zhang et al. (2011) displayed phonologically and semantically unrelated English prime-target pairs, however, the Mandarin translations of the primes and targets had overlapping phonology. For example, the prime-target pair of 'east' and 'thing' are unrelated in English, but in Mandarin, 东 'east' /doŋ/ and 东西 'thing' /doŋ.xi/ share the first character, which makes 东 'east' a phonological prime for 东西 'thing'. Notably, this is not a semantic prime, as they chose opaque compound words for targets. This means that each target was

composed of two words that were not semantically related to the meaning of the compound. These studies are presented to show the range of phenomena that priming studies can be used to examine, although this project will focus on strictly orthographic priming using translation equivalents, presented visually and not auditorily.

1.3 Non-selective lexical access

Now that we have a thorough understanding of what goes into a masked priming experiment, we can turn our attention to explaining why they work. When thinking of lexical access, or the speed and ease with which a word form is processed, the factors that might impact this are generally limited to the word form, for example, the phonology of the word or the semantic category it belongs to. Priming, then, allows us to manipulate these factors and test whether we can influence the access of a target word from by using a prime that shares some properties with the target. For example, Lupker (1979) showed that competing semantic information can influence lexical access to forms in the same category. In a picture-naming task with words superimposed on the pictures, participants were significantly slower to name the picture when a word was in the same semantic category as the picture, like when the word ‘dog’ was superimposed on an image of a mouse. This is thought to happen because the close semantic form interferes with access to the actual picture, which led to various models of lexical access addressing the details of how and why this interaction happens. When thinking about testing this effect in bilinguals, compared to monolinguals, we can then ask if this effect might be present cross-linguistically. If it is the case that each language is separate in a bilingual’s lexicon, then a bilingual doing this task in one of their languages should not be affected by a

competing picture name in one of their other languages, but this is not reflected in the literature: Rusted (1988) demonstrated the same semantic interference effect in Mandarin-English bilinguals, regardless of whether the word was in English or Mandarin. This example of cross-linguistic priming shows us that accessing lexical forms in one of a bilingual's languages can affect their access of lexical forms in their other language, demonstrating that the bilingual lexicon is not language-selective. Ergo models of the bilingual lexicon must include some interaction between L1 word forms and L2 word forms (López, 2020; Wu et al., 2013).

Other cases of non-selective lexical access have been demonstrated repeatedly in the literature, for example, in studies investigating the effect of interlingual homographs on lexical access. This term refers to words that share a form across languages, but not meanings, like the German word *Rat*. This looks the English word 'rat' but means advice or counsel, leading to questions about what meaning might be accessed when words like these are used in priming study. To examine this, Beauvillain and Grainger (1987) looked at cross-linguistic priming using interlexical homographs[2] and demonstrated that interlingual homographs, when presented in one of a bilingual's languages, can be used to facilitate access to word forms in another. In a primed lexical decision task with French English bilinguals, they showed that participants were faster to respond to the English target in trials where the French prime was a homograph for an English word that was related to the target, like in a trial where the French word *coin* 'corner' was used to prime the English target 'bank'. Other investigations of interlexical homographs have

[2] These words are called interlexical homographs in the Beauvillain and Grainger (1987) study, although they are now more commonly described as interlingual homographs. To make this study easier to understand in the contemporary context, the term interlingual homographs will be used from now on.

been conducted with variations in methodology and language pairs (de Groot et al., 2000; Elston-Güttler et al., 2005; Kerkhofs et al., 2006) and all found a significant effect of the interlingual homograph prime on the response measure, indicating that the participants were experiencing facilitation from one of their languages when completing a task in the other. The consistency of this effect in the literature reinforces the idea that the lexicons of the two languages of a bilingual are intertwined and shared with each other, i.e. words from both languages are able to influence each other, as they all found a significant effect of the interlingual homograph. These data are not compatible with a selective view of lexical access, where bilingual lexicons are like two monolingual lexicons and separate from the other (Potter et al., 1984), as selective views would predict no difference between the homograph and non-homograph primes. However, questions were raised about whether the participants were really accessing words in another language, since the forms were shared across languages. To address this, we can examine whether this effect is consistent across other types of behavioural tasks or as physiological measures.

This idea of non-selective lexical access is additionally supported by data ranging from behavioural (picture-word interference and eye-tracking) studies to neurological (ERP and fMRI) studies. For example, in a picture-naming study similar to Lupker's (1979) investigation of semantic interference, Costa and Caramazza (1999) demonstrated that Spanish-English bilinguals experience the same semantic inhibitory effect as the monolinguals in Lupker's study. The bilinguals were slower to name a picture if the superimposed word was in the same semantic category as a picture, regardless of the language of the word (for example, if a picture of a table was overlaid with the English word 'chair' or the Spanish translation *silla*). This supports work by Hermans et al.

(1998, in Wu & Thierry, 2017) and Rusted (1988) using a similar task, as well as results from Spivey and Marian's (1999) eye-tracking experiment, where they found that Russian-English bilinguals completing a task in Russian experienced interference from English. Participants sat at a table with four objects and heard instructions in Russian asking them to move one of them. The names of the other three objects on the table were not phonologically similar to the name of the target in Russian, however, one object had an English translation that shared initial phonetic features with the target. This task was then repeated with a distractor object that did not share any phonetic features with the target. For example, in a trial where a stamp (called *марку* /maɪku/ in Russian) was the target, the critical distractor was a marker (*фломастер* /flɐmasʲtʲɪr/): There is no phonetic overlap between the distractor and the target in Russian, but the English translation for this distractor starts with the same phonemes as the target. This phonetic overlap in translation is what made this a critical distractor, compared to the control distractor for this trial, ruler (*линейка* /lʲɪˈnʲejkə/), which does not sound like the target in English nor in Russian. Through comparing gaze duration for the same target in the critical and control distractor conditions, Spivey and Marian found that the bilinguals glanced at the critical distractors with a hidden phonological overlap significantly longer than the control distractors that were not phonetically similar to the target object in English nor Russian. This result is analogous to the longer reaction times in the earlier Costa and Caramazza experiment, and indicates that these distractor words slowed down processing of the target significantly.

In addition, electrophysiological studies have demonstrated that bilinguals in monolingual language production tasks have increased activation in brain regions

essential to executive function, as compared to monolinguals (Blanco-Elorrieta & Pylkkänen, 2016; Abutalebi & Green, 2008, Crinion et al., 2006, in Wu & Thierry, 2017). These regions govern tasks such as response selection and conflict monitoring, which are critical factors in limiting the language output of the bilingual Green & Abutalebi, 2013). And this effect is salient in non-production tasks: Thierry and Wu (2007) demonstrated that Mandarin Chinese–English bilinguals reading English words activated Mandarin lexical representations unconsciously in a priming experiment. They showed pairs of English words to Mandarin-English bilinguals and asked them to determine whether the second word was related to the first word, but some of the English word pairs they read contained a hidden overlap when translated into Mandarin. In English, the primes and targets had no semantic nor phonological overlap, but when translated into Mandarin, the prime and target shared the first character, creating a phonological and semantic prime. They found Mandarin-English bilinguals had a smaller N400 ERP in response to targets preceded by these hidden primes, as compared to targets paired with primes that had no hidden prime. The same effect was found in English monolinguals in this study in response to semantically-related primes (compared to semantically unrelated primes). The authors do not focus on whether a lower N400 ERP means that processing was easer or more difficult for the hidden prime trials for Mandarin-English bilinguals, rather, the key finding is that they demonstrated ERP patterns of semantic priming when they were exposed to primes that only existed in their L1, while being tested in their L2. The participants did not report being aware of the hidden translation primes when asked after the experiment, suggesting that the activation of the L1 when reading in the L2 happens automatically.

Upon reviewing the evidence for non-selective lexical access, we can now put *separate-storage* or *selective-access* models of the bilingual lexicon to rest. These models, which imagine the bilingual lexicon as simply a dictionary of entries with one section corresponding to one language (e.g. Mandarin) and another section corresponding to another language (e.g. English), do not account for the data presented that demonstrates interaction between a bilingual's two languages. For example, the Concept Mediation Model, proposed by Potter et al. (1984), understands the bilingual lexicon to consist of separate L1 and L2 lexicons connected to a shared conceptual bank but no direct linkage between the L1 and L2 lexicons. While this model was supported by data from their single-language word-naming tasks, it does not propose a way of understanding phenomena where the L1 and L2 interact, like the experiments with interlingual homographs (Beauvillain & Grainger, 1987; de Groot et al., 2000; Gollan et al., 1997), Costa and Caramazza's (1999) picture-naming studies, the eye-tracking work from Spivey and Marian (1999), nor the electrophysiological data (Thierry & Wu, 2007) demonstrating non-selective lexical activation. These models cannot make predictions about the outcomes of experiments such as masked priming experiments, where researchers use a bilingual's use of one language to influence their access of another, and they will not be further considered as they do not explain the empirical evidence at hand.

1.4 Reviewing the Mandarin-English literature

Now that we have reviewed some of the key background information in the general domain of lexical access and backward masked priming, we can turn our attention to how these have been used in the literature investigating the bilingual lexicon.

Masked priming lexical decision tasks have been used in many studies on lexical access, and this project focuses on masked priming lexical decision tasks for the ease of comparison to other experiments. For example, this task has been used with Japanese–English bilinguals in Nakayama et al. (2016), Korean–English bilinguals in Lee et al. (2018), and Greek–English bilinguals in Dimitropoulou et al. (2011) to demonstrate a significant backward priming effect. However, the results with Mandarin–English bilinguals have been mixed: One experiment of five in Jiang (1999) and one experiment of three in Luo et al., (2013) found a significant effect of an L2 prime on an L1 target, but the other experiments in those studies, as well as experiments conducted by Jiang and Forster (2001), Wang and Forster (2015), Witzel and Forster (2012), and Xia and Andrews (2015) failed to find this effect. This translates to two out of fifteen published experiments with Mandarin–English bilinguals finding a backward priming effect, leading to the question of why this effect seems to be elusive in this population: Is it that Mandarin-English bilinguals are somehow different from the other bilinguals? Or are there experimental and methodological factors that might explain this? Could it be possible that this effect exists in this population, but it has evaded detection?

To address the general question of whether the backward priming effect exists, Wen and van Heuven (2017) performed a meta-analysis that compared results of masked priming lexical decision experiments (using translation primes) with late bilinguals of many language combinations. They evaluated 31 experiments (across 20 studies) examining forward priming and 33 experiments (across 18 studies) examining backward priming, and calculated a standardized priming effect size, Cohen's *d* (Cohen, 1962) for all experiments. This value is calculated based on the reported test statistic (either *t* or *F*)

and the number of participants in each experiment (*n*), which allowed Wen and van Heuven to compare the magnitude and the direction of the priming effect in each experiment, even though the studies had different numbers of participants and different ways of calculating an effect size. The results indicate that the backward priming effect is likely to exist, and it is quantitatively smaller than the forward priming effect. They found an effect size of 0.86 for the forward priming effect, which can be considered a large effect, but found an effect size of 0.31 for the backward priming effect, which is often considered in the small range. They conclude that there is a quantitative, but not qualitative difference separating the forward priming effect and the backward priming effect. However, the data from Mandarin-English bilinguals counter this overall finding, as only two out of fifteen studies with this population report a significant backward priming effect, despite being the most well-studied group: A total of 450 Mandarin-English bilinguals participated in the studies included in the meta-analysis, significantly higher than the second and third most studied groups with 122 Japanese-English bilinguals and 108 Greek-English bilinguals included in the analysis.

These results then lead to the question of why studies of Mandarin-English bilinguals so rarely report a significant backward priming effect, compared to other cross-script bilinguals. In their meta-analysis, Wen & van Heuven (2017) suggest that the lack of a consistently significant backward priming effect with Mandarin-English bilinguals could be due to the low number of items per condition (i.e. stimuli per condition): The average number of items per condition for experiments with other cross-script bilinguals is 38.3, but for studies with late Mandarin-English bilinguals[3], the average number of

[3] This excludes one study (Wang, 2013) that was included in the meta-analysis but examined simultaneous and early Mandarin-English bilinguals, who are not part of the population of interest for the present project

items is only 16.6 (see *Table 1.1*). Based on this, we might conclude that future studies looking for a backward masked priming effect use at least 39 items per condition, given that there have been significant effects found in those experiments.

	Number of Participants	Items per Condition	Observations per Condition	Priming Effect (ms)
All backward priming experiments	28.94	24.89	720.62	11.38
Mandarin-English bilinguals	28.13	16.56	453.25	9.88
Other cross-script bilinguals	33.75	38.25	1311.50	9.30

Table 1.1: Average values for number of participants, items, observations per condition, and priming effect across all backward masked priming experiments, experiments with Mandarin-English bilinguals, and experiments with cross-script bilinguals in Wen & van Heuven (2017)

However, it is important to consider other explanations. Both Nakayama et al. (2016) and Wen & van Heuven (2017) propose that the statistical practices of previous work with Mandarin-English bilinguals could be responsible for this, drawing attention to the low numbers of stimuli used in experiments. And both sets of authors, as well as others, note that work with bilinguals often focuses on proficiency but fails to assess their participants' proficiency in their second language, making it difficult to determine whether participants in previous experiments with Mandarin-English bilinguals were at a high enough level of proficiency in English to expect a backward priming effect (Nakayama et al., 2013). Furthermore, Lee et al. (2018) raise the possibility of previous work with Mandarin-English bilinguals not using a long enough SOA in their priming protocol to detect an effect. The next chapter will examine these proposed causes and explain the rationale for focusing on power, in this project.

Chapter Two: Backward masked priming and Mandarin-English bilinguals

In the previous chapter, I reviewed lexical priming and discussed non-selective lexical access to explain why priming experiments demonstrate the effects that they do, as well as touched on previous work examining the backward priming effect. In this chapter, I will build on this and discuss why it is important for our understanding of the bilingual lexicon to investigate the discrepancy and explore the reasons proposed in the literature for the inconsistent behaviour of Mandarin-English bilinguals, compared to other cross-script bilinguals. To do this, I will discuss proficiency testing and issues with the masked priming protocol itself to explain why the current project has chosen to focus on statistical power as the explanation to investigate empirically.

2.1 Models of the bilingual lexicon

Now that we have discussed how previous results with Mandarin-English bilinguals stands out in the literature on backward masked priming, we can start thinking about what this means for our theories of the mental lexicon. First, as established, non-selective lexical access is assumed. This means that the relevant models are aimed at accounting for cross-linguistic priming, however, there is still a debate on whether there is a quantitative or qualitative difference between the forward and backward priming effect. In terms of predicting a qualitative difference, there are several models that are based on lexical activation: For example, in the Revised Hierarchical Model (Kroll, Judith, & Stewart, 1994), the authors hypothesize that the L1 and L2 lexicons are separate but both connected to the conceptual store, albeit weakly for the L2. Additionally, they propose that there is a strong link from the L2 to the L1 lexicon and a

weak link between the L1 lexicon and the L2 lexicon, given that the process of learning a language starts with translation equivalents. The weak links are thought to grow stronger with use and proficiency in the L2, but the process of lexical access remains the same for both languages. The BIA+ model (Dijikstra & van Heuven, 2002) offers a slightly different way of understanding the process of lexical access, but again, maintains that this is the same for all words, regardless of language. They explain that the observed asymmetry in the magnitude of the priming effect is due to differences in baseline activation between the L1 and the L2: L1 lexical forms are more activated by default, but the baseline activation of L2 words can be raised with proficiency and frequency of use. This is quite similar to the Distributed Representational Model from Schoonbaert et al., (2009), which is also based on activation, but of semantic nodes that are connected to word forms rather than lexical items themselves, with weaker linkages between L2 word forms and semantic nodes than between L1 word forms and semantic nodes. All of these explanations are in-line with the results of Wen and van Heuven's (2017) meta-analysis, which concluded that the backward priming effect is quantitatively smaller than the forward priming effect, but not qualitatively different.

However, there are other models of the bilingual lexicon that suggest that the backward masked priming effect would only exist under specific conditions, predicting a qualitative and not quantitative difference between forward and backward priming effects. These have been proposed to try to explain the lack of a consistent backward masked priming effect with lexical decision task in the literature: For example, the SENSE model from Finkbeiner et al. (2004) suggests that lexical access is based on the semantic density of a given lexical entry. Under this model, an L2 prime won’t activate

enough semantic information to activate the L1 representation in a lexical decision task, but it might activate enough information in a semantic task, since the task is based on the actual driver of word form recognition (for example, in a semantic categorization task). Alternatively, Witzel and Forster (2012) propose the Episodic L2 model, which posits that there is a qualitative difference in how L1 words and L2 words are accessed. This model states that L2 words are stored in episodic, rather than long-term memory, which impacts their ability to activate (and therefore prime) L1 word forms. Following from this, they suggest that a backward priming effect can only be found in an episodic memory task (for example, recalling words from a list), rather than a lexical decision task. Since both the SENSE model and the L2 Episodic Memory model predict forward priming but not backward priming in lexical decision tasks, they propose a qualitative, rather than quantitative, difference between L1 lexical access and L2 lexical access. This makes them incongruous with the conclusion from the meta-analysis that there is a quantitative, not qualitative difference between the forward and backward priming effect, and this also puts these models at odds with some of the literature that does show a backward priming effect (Chen et al., 2014; Dimitropoulou et al., 2011; Lee et al., 2018; Nakayama et al., 2016; Wen & van Heuven, 2018; Xia & Andrews, 2015). However, their predictions have been supported by other empirical evidence, mostly from Mandarin-English bilinguals demonstrating no or limited backward masked priming effect in lexical decision tasks (Jiang, 1999; Jiang & Forster, 2001; Luo et al., 2013). Therefore, it is important to investigate this exceptional behaviour to assess whether there exists empirical support for these theories.

2.2 Potential causes for the incongruent results

Now that we have established the importance of continuing to investigate this effect, we can consider proposals for why this effect has not been consistent in the literature with Mandarin-English bilinguals. Low number of stimuli used in previous studies (Nakayama et al., 2016; Wen & van Heuven, 2017), a lack of stringent proficiency assessment (Dimitropoulou et al., 2011; Lee et al., 2018; Nakayama et al., 2016), and low SOA (Lee et al., 2018) have all been proposed, and to help separate them, I will categorize these proposals into methodological factors (proficiency and SOA) and statistical factors (low number of stimuli). This is done since the statistical factors are a part of every experiment included in this project, including the experiments examining the methodological factors. I will also operationalize low number of stimuli as experimental power, in this section. As a brief recap, experimental power refers to the likelihood of an experiment detecting an effect that exists in the population that it samples from, assuming that the effect exists. This is influenced by factors such as the size of the effect, the number of observations included in the analysis, and the variance in the data. My reasoning for using low number of stimuli as a proxy for statistical power will be developed in the next chapter, as I will show that changing the number of stimuli and participants are the best options for researchers in this field for increasing power.

2.2.1 Proficiency and the priming effect

First, I will consider a lack of strict proficiency testing and reporting as a potential contributor to the lack of a consistent backward masked priming effect in the Mandarin-English literature. As noted by Dimitropoulou et al. (2011), Lee et al. (2018), and

Nakayama et al. (2016), it's crucial to report and assess participant proficiency when conducting studies with bilinguals, as proficiency can affect the way that participants behave in their L2. For example, Dimitropoulou et al. found that while L2 proficiency did not affect the priming effect size in their study of forward and backward priming in Greek-English bilinguals, it did have an effect on their response latency and the error rates. This, however, runs counter to Nakayama et al., who found that proficiency has a strong effect on the priming effect: In their experiments, a significant backward priming effect was found only in their higher-proficiency participants and not their lower-proficiency group. This is further supported by results from their previous study (Nakayama et al., 2013), which also suggests that there is a strong effect between proficiency and the size of the backward priming effect. On the surface, this seems like a contradiction, although digging deeper into the details of the way that proficiency was assessed in each study reveals that the participant groups in Nakayama et al. were fairly widely spread out in terms of proficiency, with their higher-proficiency groups scoring in the 96th and 98th percentile of the TOEIC (Test of English Communication) and their lower-proficiency group scoring in the 75th percentile of the same test. Comparatively, the bilinguals in Dimitropoulou et al. might be considered more like different bands of intermediate bilinguals; they used three groups of bilinguals (lower-proficiency, medium-proficiency, and higher-proficiency) who gave self-rated proficiency ratings of 5.8, 6.8, and 7.6 out of 10, respectively. This is despite Dimitropoulou et al. controlling for proficiency by separating these groups based on their score on the British Council Examination, out of a possible 9 points, the lower-proficiency group scored between 5 and 6, the medium-proficiency group scored between 7 and 8, and the higher-proficiency

group scored 9 points. The close spread of the self-rated proficiencies raises questions about whether their participants were truly at different proficiency levels, compared to the participants in Nakayama et al., which might have led them to conclude that there was no effect of proficiency on the priming effect. Their results on the forward masked priming task also support this argument, as all three groups produced very similar forward priming effects (28-31ms), which would not be predicted if they were at different levels of proficiency (Nakayama et al., 2013). Additionally, it could also be the case that these two studies are a poor comparison, as the average age of acquisition for the participants in Dimitropoulou et al. was 7 for all participant groups, whereas the participants in Nakayama et al. had an average age of acquisition of between 9.2 and 11.3, depending on the group, and this may have affected their respective results. This highlights how it is crucial to look closely at how studies operationalize the construct of "proficiency" when comparing studies investigating this phenomenon.

What implications does all this have for understanding the Mandarin-English literature? It looks like there are two options: If we go by the results from Dimitropoulou et al. (2011) and assume that proficiency does *not* influence the priming effect, then the proficiency of the Mandarin-English bilinguals examined in the literature should not matter. Regardless of proficiency, they should have shown the effect, provided that other experimental criteria were satisfied. However, if we assume that proficiency *does* influence the backward priming effect, following Nakayama et al. (2013, 2016), then it behooves us to examine the proficiency measures used in previous experiments with Mandarin-English bilinguals to see whether this factor might explain the Mandarin exceptionality in the literature. In reviewing the previous studies, it seems like most

experiments were conducted with international students from China in English-speaking countries, so the proficiency measure was a minimum TOEFL score of 550 out of 677 (Jiang, 1999; Jiang & Forster, 2001; Wang & Forster, 2015; Witzel & Forster, 2012). Other experiments used IELTS scores as proficiency measures (Xia & Andrews, 2015), requiring a minimum of 6 out of a possible 9, and others relied on self-reported proficiency measures (Chen et al., 2014; Luo et al., 2013; Zhang et al., 2011). The use of self-rated proficiency is slightly concerning, as it is known to be a highly variable indicator of proficiency, but the proficiency assessments using tests make it much easier to compare the participants.

According to Educational Testing Service, which produces and administers the TOEFL, the IELTS, and the TOEIC, a score of 550 on the TOEFL is approximately equal to an IELTS score of 6.5 (Educational Testing Service, 2023a). This indicates that Mandarin-English bilinguals across the experiments were probably similar in terms of proficiency. However, to contextualize these results with the studies on proficiency becomes more difficult, as Nakayama et al. (2013, 2016) operationalized proficiency using the TOEIC. This is a two-part test, and scores are reported in terms of a listening and reading portion, with a range of 5–495 possible points, and a speaking and writing portion, scored from 0–200 points (Educational Testing Service, 2023b). Adding a test taker's score in both portions of the test gives their overall score, which is reported by both Nakayama et al. (2013) and Nakayama et al. (2016). To directly compare them, we would have to know their participants' average scores on each portion of the test, as it would be possible for a test taker to score very highly on the listening and reading portion but much lower on the speaking and writing portion, or vice versa, and arrive at the same

score. However, we can approximate based on the performance descriptors given for the TOEFL and the TOEIC. A TOEFL score of 550 indicates intermediate to high-intermediate proficiency in the B1–B2 range in the Common European Framework of Reference (CEFR) (Educational Testing Service, 2021), but a TOEIC score of 800 or 850, like the advanced participants in Nakayama et al. (2016), indicates advanced proficiency (Educational Testing Service, 2018a, 2018b) in the C1 and up range of the CEFR (Educational Testing Service, 2019). This corresponds with the percentiles given in Nakayama et al. (2016), who reported that their participants were in the 96th, 98th, and 75th percentile for their respective experiments. Conversely, a score of 550 on the TOEFL was in the 42nd percentile of all test takers in the 2015 version of the test, which was the last year that Mandarin-English bilinguals were tested for a backward priming effect (Educational Testing Service, 2016). Taken together, these suggest that the Mandarin-English participants were less proficient than the highly-proficient Japanese-English participants in Nakayama et al. (2016), but it is impossible to say how much less proficient they were, as they did not use the same proficiency assessments. Thus, it is possible that variation in proficiency of Mandarin-English participants plays a part in the lack of a consistent backward priming effect, but the degree of this influence is unknown.

2.2.2 Stimulus-Onset-Asynchrony and the backward priming effect

In response to these slightly murky results, Lee et al. (2018) propose an explanation based on the methodology of a backward masked priming experiment itself, as they were able to detect a statistically significant backward masked priming experiment in low proficiency Korean-English bilinguals with a 150ms SOA, but not

with a 60ms SOA. They achieved this by using a 50ms prime in both conditions but used a 100ms blank screen before target presentation in the 150ms SOA condition and a 10ms screen before the target in the 60ms SOA condition. This is the first study to report cross-script backward masked priming with low-intermediate bilinguals, as their participants had an average score of 655 out of a possible 999 on the TOEIC, the same test that was used to differentiate the bilingual groups in Nakayama et al. (2016). Note that this is even *lower* than the low-proficiency group in Nakayama et al., who had an average TOEIC score of 710, putting them in the 75th percentile. Crucially, the backward priming effect was only significant in the version of the experiment with an SOA of 150ms, leading Lee et al. to conclude that the longer SOA is necessary for less-proficient participants to activate the L2 word forms. This is corroborated by Nakayama et al., as they also used an SOA of 60ms and did not find a significant backward priming effect in their low-proficiency group. Thus, we can examine the previous Mandarin-English literature through the lens of SOA: If it is the case that previous experiments with Mandarin-English bilinguals did not control for proficiency thoroughly enough, they might have also not been using an adequate SOA to detect a backward priming effect in their participants. However, the average SOA of the previous experiments was 161ms, according to Wen and van Heuven's (2017) meta-analysis, with most experiments using an SOA of 150ms or above (Chen et al., 2014; experiment 4 in Jiang, 1999; experiments 1 and 2 in Jiang & Forster, 2001; Luo et al., 2013; Witzel & Forster, 2012; Xia & Andrews, 2015). Only four experiments out of the 16 included in the meta-analysis used an SOA of shorter than 150ms, one of which found a significant backward priming effect with a 50ms SOA (experiment 1 in Jiang, 1999) and the others producing null effects

(experiments 2 and 3 in Jiang, 1999; experiment 3 in Jiang & Forster, 2001; Wang, 2013; Wang & Forster, 2015). This suggests that SOAs were generally long enough in previous work, even if the participants were not as proficient as the participants in Nakayama et al..

While the high SOAs used in previous experiments with Mandarin-English bilinguals might seem incongruous with Lee et al. (2018)'s findings that a higher SOA is necessary for activating the backward priming effect in low-proficient bilinguals, I propose an explanation based on experimental power. By experimental power, I mean the likelihood of detecting a significant effect in an experiment on a sample, assuming that the effect exists in the underlying population. Experimental factors such as number of observations per condition, as well as effect size, are strong influences on this, and I will explore why this is the case in later chapters. The effect of including more items, however, is reflected in Lee et al.'s experiments testing the effect of a 150ms SOA versus a 60ms SOA, they used 30 items and either 30 or 43 participants: This is significantly higher than the average numbers of participants and items for experiments with Mandarin-English bilinguals, which use an average of 16.6 items and 28 participants. As Wen and van Heuven's (2017) meta-analysis establishes, the low number of items used in these experiments is a significant predictor of effect size, which is a strong influence on the power of a given experiment. Given that experimental power being a foundational requirement for detecting an effect that exists, these results demonstrate that even if experiments are using bilinguals that are adequately proficient and using an SOA that gives participants adequate time to mentally process the primes, they must still have a sufficiently high level of power to be able to reliably detect an effect.

As a demonstration of this, we can turn to Xia and Andrews' (2015) experiment with moderately-proficient Mandarin-English bilinguals. Their participants were undergraduate students at the University of Sydney who reported a minimum of 6 out of 9 on the IELTS and had an average self-rated proficiency of 4.9/7 on their English skills, and they had been living in Australia for a minimum of one year, with an average of three years. They rated their Mandarin skills as significantly higher than their English and reported using both languages roughly equally, despite living in an L2 environment. In their lexical decision task protocol, they used a 50ms prime and an 150ms backward mask ("&&&&&&&&") for an SOA of 200ms. They used 34 participants in experiment 1B and 30 participants in experiment 2B, but only 16 items per condition in each experiment. The participants demonstrated a 12ms and 14ms backward priming effect in the respective experiments, which did not reach significance. Priming effects of this range were significant in experiments 1 and 2 in Dimitropoulou et al., (2011), who reported a 14ms priming effect in both experiments, and smaller priming effects were significant in experiment 1 of Nakayama et al. (2016) and experiment 1 of Lee et al. (2018), both of which reported a 10ms priming effect. The crucial difference between these experiments and Xia & Andrews' (2015) experiment is that these experiments used significantly higher numbers of items per condition, namely, 58 for Dimitropoulou et al. and 30 for both Nakayama et al. and Lee et al.. While avoiding post-hoc power speculation, this finding, combined with Wen and van Heuven's meta-analysis, suggest that stringent proficiency testing and an appropriate SOA are not enough, we must also take care to conduct properly powered experiments to detect a backward masked priming effect.

2.3 Research Question

With all of this in mind, we can now consider the concept of statistical power and how it may have affected the previous work on Mandarin-English bilinguals. Recall that Wen and van Heuven's 2017 meta-analysis of backward masked priming experiments included 33 experiments with both cross-script and same-script bilinguals, and studies with Mandarin-English bilinguals comprised 15 of those experiments. This resulted in a total of 450 Mandarin-English participants across all studies, making this population the most well-studied among all the language pairs included in the meta-analysis and far ahead of the second and third most studied populations of 122 Japanese-English bilinguals and 108 Greek-English bilinguals in the meta-analysis. However, only two of these fifteen studies reported a significant backward masked priming effect (one experiment of the five conducted by Jiang (1999) and one experiment of the three in Luo et al. (2013)). This runs counter to the results of the meta-analysis, which concludes that the backward masked priming effect is significant, and also counter to the significant results from other cross-script bilinguals (for example, Japanese–English bilinguals in Nakayama et al. (2016), Korean–English bilinguals in Lee et al. (2018), and Greek–English bilinguals in Dimitropoulou et al. (2011)). Such anomalies raise the question of why there have not been more conclusive results with this particular population. One possible interpretation is that there is something inherently different about Mandarin-English bilinguals, compared to these other bilinguals, in that for some reason they are less sensitive to backward masked priming. While there might be some merit to this, as there could be an effect of the other languages using (at least partially) an alphabet-based writing system compared to the Mandarin logographic system, I will argue that the

differences in statistical methods between the studies drives this discrepancy (after all, most Mandarin speakers are taught the alphabetic pinyin writing system in primary school and it must be used when typing in Mandarin on an alphanumeric keyboard). Namely, the other cross-script studies that found significant effects used 30-60 stimuli items per condition whereas the studies with Mandarin-English bilinguals used only 12-16 items per condition. Knowing that the statistical power of an experiment is the likelihood of finding a significant effect, and knowing that this is influenced by the number of stimuli items per condition, we can then ask whether these studies were appropriately powered to be able to find an effect. This will not be a post-hoc power analysis of each individual experiment, which is a tautological exercise—if an experiment is underpowered, then it is most likely not going to reach significance due to the lack of power (Hoenig & Heisey, 2001). Rather, this project will ask what the requirements are for an adequately-powered experiment that can investigate the small effect sizes typically found in backward masked priming. This can help us determine whether the lack of consistency in finding a significant backward priming effect in Mandarin-English bilinguals is a result of statistical issues in previous work, or due to an actual lack of a behavioural effect in this population. To probe these questions, simulations based on large datasets from masked priming lexical decision task experiments will be used to determine the approximate ranges of stimuli and participants necessary in the design of an adequately powered experiment. This project will aim to demonstrate how *a priori* power analyses can help improve statistical reliability and how statistical planning should be incorporated into future studies interested in examining this effect.

Chapter Three: Statistical Background

In the previous chapters, the literature on backward masked priming and the bilingual lexicon was reviewed to highlight how results from studies with Mandarin-English bilinguals are anomalous when compared to other types of cross-script bilinguals. Similarly, this chapter will review the statistical background necessary to understand why future studies with Mandarin-English bilinguals need to consider statistical power if they wish to address this inconsistency. In this chapter, I discuss the Null Hypothesis Significance Testing (NHST) framework, giving a brief overview of inferential statistics and frequentism before discussing statistical testing and *p*-values, focusing on what is meant by a statistically significant result. This is followed by a discussion of how sample size impacts statistical power to explain why this project focuses on sample size to increase statistical power, particularly for small effects.

Parts of this section may seem too rudimentary, as most of this should be common knowledge for those working in this field. However, in a review of steps that can be taken to increase transparency in psychology, Lindsay (2020) presents evidence suggesting that fundamental ideas in NHST such as *p*-values are often misunderstood by psychologists and encourages researchers to review the basic tenets of inferential statistics to ensure that we understand the tools that we use to produce knowledge. With this in mind, these core statistical concepts are covered so that the logical progression from setting up an experiment to obtaining a result from a statistical test is made clear. This chapter is intended to serve as a reference for any researchers looking to design an experiment looking at the backwards masked priming effect in Mandarin-English bilinguals, including student researchers who may have less familiarity with these tools.

3.1 What is statistical significance?

First, to understand why experiments with Mandarin-English bilinguals have been generally unsuccessful in demonstrating a significant backward priming effect, we must understand what it means to achieve statistical significance. This concept comes from the way that statistics has traditionally been done in psychology and psycholinguistics, which uses an approach called Null Hypothesis Significance Testing (NHST). The goal of NHST, as with all types of inferential statistics, is to infer meaningful conclusions about a population based on a representative sample, but NHST is a specific way of doing inferential statistics that relies on specific assumptions about probability (Perezgonzalez, 2015). Given that almost all studies in Wen & van Heuven (2017) used statistical tests in the NHST framework to determine the presence or absence of a backward priming effect, this section aims to provide the statistical background necessary to investigate potential statistical explanations for the lack of significant backward priming results with Mandarin-English bilinguals.

3.1.1 Outline of an NHST experiment

For a typical experiment in the NHST framework, there is a straightforward series of events: Researchers collect data, divide it into groups, and compare them to determine if there is a meaningful difference between the groups[4]. To do this, they develop two hypotheses about their data: The null hypothesis (H_0), which states that there is no

[4] It is possible to compare data gathered to a theoretical distribution, however, it is much more common in the literature to compare groups (for example, comparing bilinguals' and monolingual controls' performances on a language task). This section will use examples and language that reflect this more common usage to help explain the logic behind hypothesis testing, which is generally applicable to both testing between groups and comparing to a distribution.

difference between the groups, and the alternate hypothesis (H_1), which states that there exists a difference between the groups. In NHST, only the null hypothesis is considered, and evidence is assessed to determine if the null hypothesis is *accepted* (no evidence for a difference between the groups) or *rejected* (there exists evidence for a difference between the groups). To test whether there exists enough evidence to reject the null, researchers must choose a statistical measure to represent each group, such as mean ($\bar{x}$) or variance (s^2), and compare them based on this measure. To do this, a test statistic based on the statistical measure chosen is computed for each group and the difference between them is interpreted as a quantification of the difference between the groups. If the difference between the test statistic values for two groups is large enough, according to a pre-determined criterion, the difference is declared significant, and the null hypothesis is rejected. For example, in a *t*-test, the test statistic is the mean, so the null hypothesis in an experiment using a *t*-test is that the difference between the means of the groups is zero. This may seem straightforward; however, many statistical assumptions and decisions have been made throughout this process, and the next sections will aim to explain those assumptions.

3.1.2 Inferential statistics

First, I will go over the topic of inferential statistics, which is the foundation of all statistical analyses performed in experimental psycholinguistics. This is the process of drawing conclusions about a population by taking representative samples from that population and measuring characteristics from the sample like the mean ($\bar{x}$) and variance (s^2) to infer those characteristics at the population level. There is always a degree of

uncertainty in approximating the population statistic the sample, as each sample of a population is randomly selected and will yield slightly different estimates. For example, if the average rainfall in a given area for the month of June is 30mm over 50 years, we might expect that the average rainfall for any given year to fall in the range of 20-40mm. However, if we only measure a few years of average rainfall that are closer to 20mm, we might assume that the average rainfall overall is near 20mm rather than the actual value of 30mm. As a result of sampling, inferential statistics is also concerned with quantifying the *confidence* of a given claim about a population (Gelman & Hill, 2006).

There are two main approaches to quantifying the certainty of a claim and they differ in their approaches to probability. This section will discuss *frequentism*, which is the view of probability behind NHST. An alternative approach that is becoming more widely-used (Bayesian statistics) will be discussed briefly in a later chapter. Frequentism is concerned with the frequency of sampling from a given population to estimate the probability of any given event. It relies on an assumption that as more samples are measured from a given population, the more the measurements will converge on the underlying probability in the population: For example, when flipping a coin, the underlying probability of landing on 'heads' is 0.5, or 50%. A frequentist view states that as the number of coins flipped approaches infinity, half of the coins will land on 'heads', but there is no meaningful probability of any given coin toss landing on 'heads' (Vasishth & Nicenboim, 2016). This philosophy is demonstrated in NHST in the criterion for acceptability of evidence: When maintaining the same criterion for significance, the same t-value is more likely to indicate a significant result if it comes from a group with thirty

participants rather than three participants. This is a way of modelling results from finite data points, since no experiment can collect an infinite number of data points.

3.1.3 Statistical tests and error

For an experiment using the NHST framework, statistical testing is critical to interpreting the results of the experiment: The test statistics by themselves are not useful for answering the questions that the researcher is interested in, they only become useful when interpreted in relation to a test statistic from another group. For example, if a researcher is interested in comparing arithmetic test scores from students who ate breakfast compared to students who did not eat breakfast, they would need to conduct a statistical test to determine if the difference is due to their experimental manipulation or if the students in one group just randomly did better on the test that day. In addition, the nature of sampling is always subject to error, so the researcher must decide before doing the test what margin of error they are willing to accept. Furthermore, there are different *sorts* of errors which must be taken into account. There are two main types of statistical error commonly discussed: Type I and Type II errors.

Type I error is also known as a false positive, which is when the null hypothesis is rejected despite there being no actual difference in the samples that are being compared. For example, if a test comparing brain activity between two dead salmon rejects the null hypothesis (i.e., declared that there was a difference between the two dead fish brains), this would be a Type I error. When conducting statistical tests, this is reflected in the alpha (α) level of a test, which is used to determine the critical test statistic value for declaring that there is a significant difference between the groups. By convention,

researchers often set α at 0.05, which reflects a 5% chance that the test has detected a difference between the samples that does not, in fact, exist between the populations. This is a way to combat the potential for error discussed earlier, where measuring characteristics from non-representative samples might mean that a difference is detected between the two samples when, in fact, they come from the same underlying population.

Conversely, a Type II error is a false negative, where the test fails to reject the null hypothesis but there actually does exist a difference between the populations. This would be like if a test comparing the brain activity of dead salmon to a live salmon *failed* to reject the null hypothesis (i.e., declared that there is no difference between the live and dead fish brains). This is often referred to as the beta (β) for a given test, which reflects a researcher's tolerance for a false negative. A commonly accepted value for β is 0.2, which refers to a 20% chance that the researcher falsely fails to reject the null hypothesis. This is more commonly referred to in terms of statistical power, which is equal to 1 – β and can be thought of as the long-term likelihood of finding a significant effect that exists in the population through the process of testing samples from that population.

It is also important to note that α and β are *likelihoods* of these types of errors appearing, whereas the occurrence of an error (whether Type I or Type II) is a discrete event. Parameters such as α and β are chosen for each test to combat the likelihood of a test returning a Type I or Type II error, but determining whether the error was, in fact, an error can only be understood when we know the state of the population, often after more experimentation or measurement. For example, if we imagine that a new species of bacterium has been discovered, the researchers reporting it as a new species have measured characteristics of this sample and conducted tests that indicate that it is

different from its closest relatives. Since it is a previously-unknown species, it is impossible to know if the tests are resulting in any errors. However, if we conduct more experiments measuring the same characteristics of this bacterium and gather more evidence indicating that is not different from its closest relatives, we can look back on the first results as a Type I error. This is another reflection of frequentism in NHST, which is more concerned with long-run probabilities rather than discrete events.

3.1.4 Statistical testing and p-values

Now that the parameters of the test are set, we can begin to discuss the concept of statistical significance. Every statistical test in NHST is based on a test statistic that can be visualized as a probability distribution, where values on the y-axis reflect the likelihood of the corresponding x-axis test statistic value, assuming the null hypothesis. This can be imagined as a sort of histogram: Imagine graphing the height of all the people in a full lecture hall, with height on the x-axis and number of people on the y-axis, and trying to predict the height of someone outside the door to the room. The most probable guess would be at the area of highest density in the histogram, which is what a probability distribution demonstrates. An example of a t-distribution is shown in *Figure 3.1* with probabilistic density reflecting likelihood on the y-axis.

The curves are specified by different formulae for each statistic but in the context of statistical testing, they also take a measure of an experiment's sample size into account. This is called the degrees of freedom, or *df*, and this refers to the number of independent data points that can go into any given dataset. To understand this in a more intuitive way, I will run through a quick example: Consider a set of five numbers that

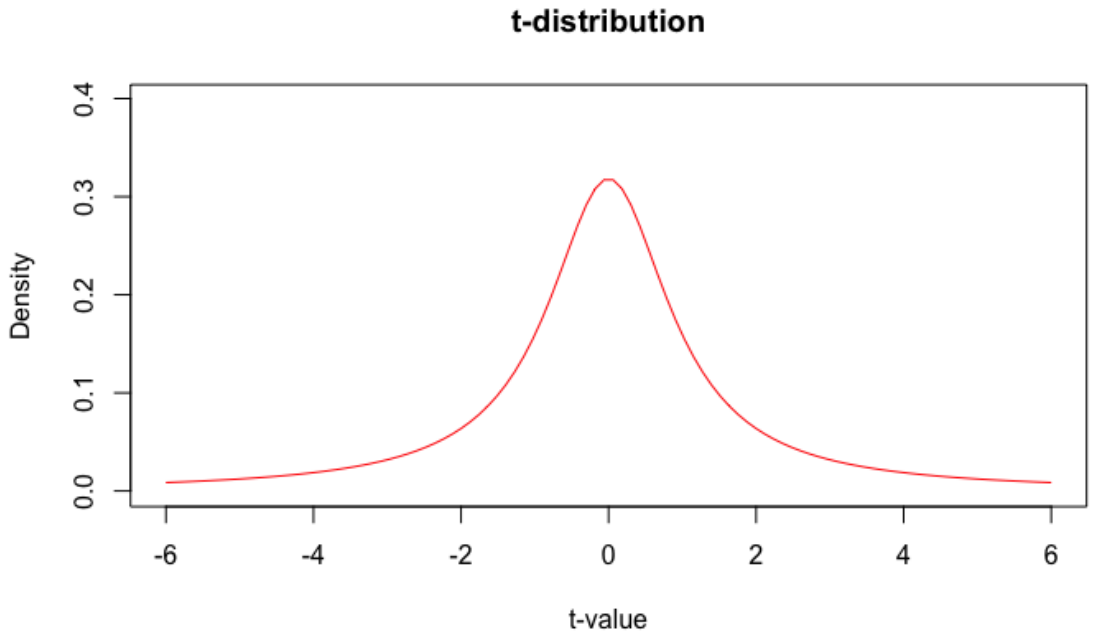

Figure 3.1: An example of a distribution for the t statistic with 1 degree of freedom

sum to 10, which means that they have a mean of 2. This could be a set of [2, 2, 2, 2, 2] but this could also be a set like [4, 8, -4, -8, 10]. If one number in any of these sets was unknown, we could calculate it by knowing that the set must sum to 10; however, this means that the unknown value must be the only value that makes the set have a mean of 2. This is what is meant by independent data points: This last data point is not independent since it must be the one that completes the parameter, whereas the others could be any value. In this example, if we were to perform a statistical test to determine whether there is a significant difference between the mean that we measured from this set and another value, we would declare four degrees of freedom: All but one of our values can change while maintaining the number of values and the sum constant. In experimental contexts when performing statistical tests, *df* is often thought of as a reflection of the number of data points that go into the analysis. This is not entirely

accurate, as *df* also depends on the number of parameters estimated, but experimental factors like number of items or participants that influence the number of data points are the largest contributor in most experiments in psycholinguistics, as the number of parameters estimated is typically quite small compared to these experimental factors (it's quite rare to estimate more than 10 parameters, but an experiment with fewer than 10 participants or items would be exceedingly rare). Thus, we can generally understand the changing of the degrees of freedom as a result of changing the sample size . This change in the shape of the distribution as degrees of freedom change clearly illustrated in the *t*-distribution, the basis of the *t*-test, which grows narrower at the ends of the curve as *df* increases, as illustrated in *Figure 3.2*.

Now, with a *df* and alpha level for a specific test, we can visualize the range of critical values of the test statistic for that test as the range of x-values where the area

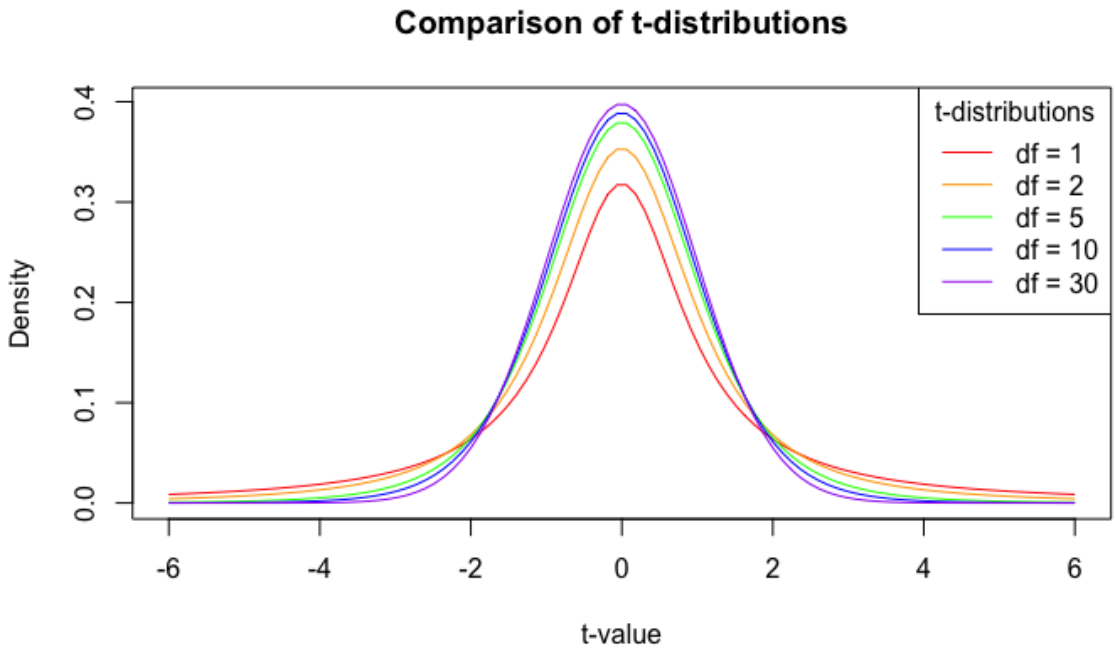

Figure 3.2: t-distributions with different degrees of freedom overlaid on each other, showing how the tails of the curve grow narrower as df increases

under the curve sums to alpha. In fact, there are two ranges here, the positive and negative ranges, for the two tails of the *t*-distribution. The sign of the *t*-statistic describes the direction of the effect, which in experimental contexts describes whether the experimental manipulation being tested increases or decreases the outcome values. Using one or two tails is a choice that the researcher can make, depending on their hypotheses. If we consider the entire *t*-distribution (i.e. positive and negative ranges), we are doing a *two-tailed* test which does not bias the test results into one causal direction. For example, if we test the effects of eating breakfast on test scores, a two-tailed test does not assume whether scores will be higher or lower after eating breakfast. Conversely, a one-tailed test that only considers one half of the *t*-distribution (i.e. only the positive or only the negative range) assumes that scores will be either higher or lower in the group that ate breakfast, depending on the tail chosen. This is less conservative than the two-tailed test, as a one-tailed test allows a larger range of critical test statistic values. Since the critical values are *t*-values where the area under the curve sums to alpha, a two-tailed test includes a range for each tail where the area under the curve sums to half of alpha. For a test where only one tail is being considered, the range is larger as the area under the curve is allowed to sum to alpha (Winter, 2019).

From here we can also see how keeping alpha the same while changing the *df* results in a larger range of values accepted as significant: A distribution with narrower tails means that the y-values are decreasing, resulting in a larger range of x-values where the area under the curve sums to alpha. This is a reflection of the frequentist approach to statistics in NHST, where increasing sample size (the largest influencer of *df*) allows us

to be more confident in our results and accept a value as significant from a sample with more participants that we may not have accepted from a sample with fewer participants.

With this visualization we can also see how researchers compute *p*-values from the results of statistical tests. Since a *p*-value refers to the likelihood of obtaining that result under the null hypothesis, we can find the test result on the distribution for the test statistic and report the area under the curve up to that value from the closest tail. If that sum is lower than alpha, we declare that the results are significant at the level of the sum, i.e., a result being significant at the $p < 0.05$ level means that the sum of the area under the curve is less than 0.05.

3.1.5 Sample size and power

Previously, the effect of sample size on the critical test statistic values was discussed in the context of degrees of freedom for a statistical test. Here, we will explore the effects of sample size on power, or the likelihood of detecting an effect in the sample that exists in the population. As previously defined, the power of a given analysis refers to 1 - β, where β is the probability of a test falsely rejecting the null hypothesis. Experiments usually aim for a power of 80% so that analyses will not be oversensitive and detect an effect where none exists (i.e. a β of 20%). It is important to run properly-powered experiments as under-powered tests run higher risks of rejecting an effect that actually exists, and non-significant results from under-powered experiments cannot be interpreted as evidence for the absence of an effect. Rather, they can only demonstrate that there were not enough data in a given experiment to determine whether an effect

exists, which does not generalize outside of that experiment. (Kirby & Sonderegger, 2018a).

Under-powered experiments also run the risk of what are known as Type M and Type S errors (Gelman & Carlin, 2014). These errors are similar to Type I and Type II errors, but they differ in the sense that they are not risks of incorrectly determining significance, rather, they are risks that the effect sizes calculated from experiments deviate from the real effect size. Type M errors refer to measured effect sizes that greatly differ in *magnitude* from the real effect size, for example, estimating an effect size of 3 when the real effect size is 0.3. Type S errors are errors in sign or *polarity* of the measured effect size, which could be like measuring a *positive* correlation between an independent and dependent variable when in reality (i.e. when compared an effect size measured from the full population), the correlation is *negative*. Like Type I and Type II errors, these are inherent risks of inferential statistics, as they arise due to non-representative sampling. However, the chances of making these errors increases dramatically as power decreases, with Type M errors increasing faster than Type S errors. The risks of these errors also decrease as power increases, underscoring again the importance of both running properly-powered experiments and critically evaluating the statistical methodology of a given study, regardless of the significance of the results (Gelman & Carlin); further emphasizing how focusing on statistical significance does not always paint the full picture of a given experiment.

With this in mind, it is alarming that statistical power in psychological experiments often hovers around 30–40% (Smaldino & McElreath, 2016), which corresponds to a 60–70% long run probability of missing an effect that exists in the

population. This is far from the generally understood best practice of aiming for a β of 20%, since power of 30–40% means a β of 60–70%, making Type II errors much more likely in the sampled studies. In conjunction with a publication bias towards significant results at the commonly accepted $\alpha = 0.05$ level (Fanelli, 2012; Franco et al., 2014; Lindsay, 2020), the proliferation of underpowered studies (which are, by definition, highly unlikely to achieve significance) can lead to researchers adopting questionable statistical practices such HARKing, or Hypothesizing After Results are Known (Kerr, 1998) or running multiple analyses on the same data until a significant test result is achieved (Simmons et al., 2011). Both of these practices increase the risk of a Type I error and undermine the strength and accuracy of the conclusions that are published (Lindsay, 2020).

The way that power is calculated for an analysis depends on the design of the study and while each calculation is different, they share the same conceptual foundations: The power of a given study is based on the difference between the populations, the variation within each dataset being compared, and the sample sizes in each population. Therefore, to increase the power of a study, one could increase the difference between the populations, decrease the variation within the datasets, or increase the sample size (Vasishth & Nicenboim, 2016). Often, researchers aim to increase the sample size to increase power, as changing the other parameters would change the phenomenon that they are interested in. For example, consider an experiment aimed at examining the differences between French-English bilinguals and English monolinguals' performance on a vocabulary task. To increase the difference between the populations, a researcher could choose less-fluent bilinguals or screen out bilinguals with large vocabularies, but

this would mean that their results can only be interpreted to hold for the population that was examined, making the results less generalizable. Considering that populations of interest are highly specific in psycholinguistic research, this is not an ideal way to increase power. Additionally, a quantifiable measure of the difference between the groups measured is often one of the aims of a given study, which means that trying to manipulate it in hopes of increasing power would change a fundamental element of the study. The second option would be to decrease variability in the datasets. To decrease variability in this sense often means attempting to elicit the same type of response from as many participants as possible through making the stimuli or participants more homogenous, however, this has a natural limit. For example, there is variation in the ways that participants respond to words in terms of their frequency or length, as well as other factors. These limitations are often incorporated into the design of experiments in terms of restrictions on words based on these form-based factors and restrictions on participants in terms of L2 experience, type of education, proficiency, and so on. But even with strict limitations on participant background, there are still many things that are out of the researchers' control—not every type of classroom education is the same, years of L2 experience can produce different results based on environment and how active participants were in their L2, and quantifying these differences is often outside of the scope of an experiment as there are near infinite ways in which participants can vary on these factors. Restrictions like this must also be balanced with the need to draw from a representative sample to be able to produce meaningful conclusions from a study. This leaves increasing sample size as the only option for increasing power in a way that does not alter the population or task in a meaningful way.

3.2 Experimental design

Now that the effect of sample size on power has been established, this section will discuss some common ways that backward masked priming experiments are designed to give context for the following chapter on calculating power for these types of experiments.

3.2.1 Assumption of independence

First, let us address the topic of independent samples versus repeated measures designs. Experimental designs can vary with respect to how participants are divided between experimental manipulation conditions, where approaches using independent sampling (i.e. between-participants) do not repeat participants between conditions, while repeated measures (i.e. within-participants) experiments expose participants to all levels of the experimental factor (Sonderegger, 2022). For example, consider an experiment where a researcher is investigating the effect of eating breakfast on test scores and has recruited ten participants. An independent samples design would split those participants into five students who ate breakfast and five who had not and compare their performance on the test. Comparatively, a repeated measures design would have all ten participants eat breakfast and take the test on a given day, and then have them repeat the test at a later date when they had not eaten breakfast. Here, eating breakfast was the experimental manipulation (also called a factor), with two levels: eaten or not eaten. In the independent measures design, each participant experienced only one level of this factor, as they either ate or did not eat breakfast. Conversely, the participants in the repeated measures design experienced all levels of this factor.

Repeated measures experiments typically have higher power than independent measures studies as they increase the sample sizes of the groups being compared and decrease the variance between each group. To appreciate the effect of sample size, consider an experiment where a fixed number of participants is recruited: If the researcher intends to compare the different levels of a given factor, then an experiment where *all* participants are exposed to *all* levels of a factor results in a larger sample size for each level, compared to splitting the participants between the levels. Moreover, repeated measures experiments also increase power due to decreased variance between the groups as each participant acts as their own control. The analyses used with these experiments can separate the variance due to participant differences from the total variance, analogous to removing some of the noise in the dataset, allowing a weaker signal to be seen more clearly. Therefore, repeated measures experiments are the most common types of experiments in studies of bilingual masked priming as this allows researchers to maximize experimental resources (i.e. recruit as few participants as possible) as well as reduce variance in the data (Sonderegger, 2022). And despite this repetition of participants in both groups, none of the examples given so far have violated the assumption of independence. This states that each datapoint in a set has no relation to any other point in the set, like if each datapoint came from a coin toss (Sonderegger, 2022). Even in the repeated measures experiment example, the assumption of independence is satisfied, as each datapoint in the sets being compared has come from a different participant and therefore has no relation to the other datapoints within the set.

Conversely, the experiments typically done to investigate backward masked priming are repeated measures that use multiple stimulus items per experimental

condition, which violate the assumption of independence. For example, consider a masked priming experiment with 20 English-French bilingual participants which examines the difference between participants' response to a single target after control versus translation primes. If all the participants complete this experiment, the researchers would have 20 responses in each condition, corresponding to one response each from 20 different people in each dataset (control vs. translation prime). This does not yet violate the assumption of independence—This is simply a repeated measures experiment, and analyses have been developed to account for the decreased variance in taking samples from the same source (for example, a simple *t*-test can be used with the data from this experiment, if the equation is set up properly). However, if 10 more targets are added to this repeated measures experiment, then the experiment violates the assumption of independence in two ways. Each participant would contribute 10 responses to each condition and each target would produce 20 observations for each condition, adding patterns into the data from the tendency for responses from the same participants or to the same words to cluster together: Adding more stimuli has greatly increased the number of observations per condition, but not by an independent process. To imagine this, think back to the coin toss example, where each datapoint does not have any likelihood of being more similar to another datapoint. In this proposed bilingual experiment, some datapoints are more likely to be similar to other datapoints, if, for example, one participant was considerably faster than the others, or if one target word was more familiar to the participants. The variance in this dataset will reflect these patterns, unlike data gathered through an independent process. This practice of *pseudoreplication*, or using statistical tests which assume independence with data not generated through

independent processes, such as *t*-tests and *F*-tests, can result in an elevated risk of Type I and Type II errors (Winter, 2011), and so this is important to for researchers to take into account when designing studies.

3.2.2 Statistical analysis for repeated measures experiments

To analyze data that do *not* conform to the assumption of independence, many researchers have now incorporated linear mixed-effects modelling into their statistical toolboxes (Baayen et al., 2008). These techniques can account for what are called *random* effects that introduce patterns into the variation of the data, such as participant and item, since responses from each participant (or item) tend to cluster together, compared to the entire dataset. Conversely, experimental manipulations are called *fixed* effects, as they introduce known sources of variation. However, before this was more common, researchers often used averaging to adapt the data to the analyses that they were doing, based on *F* ratios (Winter, 2019). Given that many of the studies in the literature incorporate these types of analyses, I will give an overview of these before discussing the linear mixed-effects modelling more commonplace now.

Analyses of Variance, or ANOVAs, use the *F* ratio as a base statistic and are common throughout the social sciences as they can compare more than two datasets and, when used properly, determine if those datasets are statistically different from each other. Such designs maintain a low risk of Type I error even when doing multiple comparisons. However, they are limited in that they can only easily incorporate one random effect, and this does not work in experiments where both participants and items are random effects. While there exists a way to incorporate two random factors using quasi-*F* ratios (Clark,

1973), this was not a common approach as it was not readily available in statistical software programs used by social sciences researchers, such as SPSS (*IBM SPSS Statistics for Windows*, 2020). For example, none of the studies that were examined in this project used this technique. To work around the difficulty of the statistical analysis without treating either participants or items as a fixed effect, researchers often performed two separate ANOVAs, one partitioned by participants (sometimes called an F1 analysis) and another partitioned by items (sometimes called an F2 analysis). An F1 analysis (by participants) averages all of a given participant's responses in a given condition into one data point for the condition, while an F2 analysis (by items) averages responses to each item into one data point for the condition, averaging across the trials. This approach transforms the data into an appropriate format but removes some of the variation in the random effects data from the final analysis, leading to underestimation of the overall variance. This can increase the risk of Type I error as the reduction in variance can decrease noise artificially, highlighting a signal that might not exist if the data had been analyzed in using a more appropriate method (Barr et al., 2013; Vasishth & Nicenboim, 2016).

This problematic result was demonstrated by Brysbaert and Stevens (2018) when they compared the results from an F1/F2 ANOVA and a linear mixed effects model on a masked priming dataset. For the same 16ms behavioural difference between prime and control conditions in this dataset, the effect sizes were $d = 0.87$ from the F1 analysis (by participants) and $d = 1.24$ from the F2 analysis (by items), but the effect size calculated from the mixed effects model was $d = 0.868$. In this experiment, there were 1020 participants and 210 items per condition, so it is not surprising that the effect size is

larger in the F2 analysis: If we think of effect size as a signal to noise ratio, we can increase the effect size by either increasing the signal or decreasing the noise. The signal was the same in each of these analyses, which means that the larger effect size in the analysis by items was due to a decrease in noise, which statistically corresponds to a decrease in variation. Since there were fewer items than participants in their dataset, the F2 analysis compressed the data into far fewer discrete values for the analysis (compared to the analysis by participants) reducing the variance and increasing the effect size. Crucially, this inflation is entirely due to the transformation of the data for the ANOVA analysis. The effect size from the mixed effects model (which did not transform the data in any way) was $d = 0.868$. Thus, the effect size from the F1 analysis of $d = 0.87$ was much closer to the mixed effects model effect size, due to the smaller degree of compression.

In response to these issues with ANOVAs, mixed effects regression models are now commonly used to analyze data with multiple dependencies, generated through processes that violate the assumption of independence (for example, when they see multiple control or multiple critical targets). This technique is based on linear regression, which is mathematically very similar to an ANOVA—each relies on partitioning the variance due to specific factors and assigning a "weight" to that factor, based on their contributions to the total variance. However, the output of a linear regression can describe the specific contribution of each level of a factor, which ANOVAs cannot. For example, in an experiment comparing participants from multiple language backgrounds, the factor "L1" would have a level for each language spoken by the participants. Since ANOVAs assess the data against the null hypothesis of "all datasets come from the same

population" and produce a binary *yes-or-no* answer to that statement, further testing is required to determine the contributions of each level. Conversely, in a linear regression model, the output is a regression coefficient for each level of each factor that, when taken as a set and compared to each other, can tell a researcher whether one level is having more of an influence on the outcome variable than another. Take an experiment comparing the test scores of English-French students, English-Mandarin students, and monolingual English students as an example. The students' language background would be the *factor*, and English-French, English-Mandarin, and English would each be the *levels*. Assuming that the English-French bilinguals are significantly different from both other groups on this test, but the English-Mandarin bilinguals and English monolinguals perform similarly, an ANOVA would just tell us that there are significant differences between the groups, and post-hoc testing would be necessary to determine between *which* groups there is a significant difference. With the right setup, a linear regression model would be able to compare all the groups against a chosen reference group (say, the monolingual English speakers) and the estimated coefficient for each of the bilingual groups could be interpreted to determine both whether they are significantly different from the monolingual group and how much they each differ. If the English-French bilinguals are significantly different from the monolingual English speakers, the coefficients would reflect that with a larger estimate for that group, compared to the English-Mandarin bilinguals. This makes the results a bit more intuitive to interpret, compared to ANOVAs. Additionally, regression modelling techniques can also be adapted to a wide variety of response data, from continuous response time data to

categorical forced choice data (unlike ANOVAs, which cannot be used to analyze categorical data), which make them popular choices for researchers.

Linear regression modelling can also be extended to account for violations to the assumption of independence without pooling data into F1 and F2 analyses, as an ANOVA would have to do. Mixed effects regression models (as previously mentioned) allow researchers to incorporate multiple random effects, such as participant and item, into regression analyses without reducing or transforming the data (Baayen et al., 2008). This is most often done with packages like lme4 (Bates et al., 2015) in the free statistical programming language R (R Core Team, 2020), in a similar way to linear regression models. However, the interpretation of model outputs (i.e., estimated coefficients) for random effects, such as participant and item, is different from fixed effects, such as control or critical target. The coefficients are still estimates of variance attributed to the effect, but since these effects are (by definition) limited to the specific participants or items that are chosen, the impact of random effects cannot be extended to say anything about its influence on the outcome variable more generally, like a fixed effect (to do so would be to draw conclusions from a specific item or specific participant, which is not generally what researchers are interested in).

Having reviewed the background for why statistical testing is performed, what is meant by statistical significance, and some common methods for analyzing data from a typical masked priming lexical decision task study, we can see how factors such as sample size and type of analysis influence power and power calculations for a given experiment. With this in mind, we can now explore the various methods available to researchers for calculating power in the next chapter.

Chapter Four: Methodology

To explore the potential effect of low statistical power in previous work with Mandarin-English bilinguals, it must first be determined what an appropriately powered experiment would look like. From *Table 1.1*, we see that the average priming effect is 11.38ms for all backward masked priming experiments, but for experiments with cross-script bilinguals (Hebrew-English, Mandarin-English, Japanese-English, Greek-English, and Korean-English), the average priming effect is lower at 9.30ms (Wen & van Heuven, 2017). This is slightly lower than the average priming effect in experiments with Mandarin-English bilinguals, which is 9.88ms, but this is not a significant difference ($t(24) = 0.90$, $p > 0.05$). In this section, I will show how simulation modelling can be used to appropriately determine experimental factors (such as number of items and number of participants) in order to reliably detect a priming effect of this size. To do this, I first discuss different methods of calculating statistical power before explaining the procedure for this project.

4.1 Calculating power

To determine the experimental conditions necessary to detect such a small effect, this project will calculate statistical power for different combinations of numbers of items and participants to see which combination can provide 80% power. This is the commonly-accepted threshold that balances the risk of Type II error against the risk of Type I error, as an overpowered test might detect an effect that does not exist in the underlying population. There are different ways to calculate power based on experimental design and type of statistical analysis, but since they're based on the same conceptual

ideas about the elements that influence power, they generally operate the same way. That is, they take estimated effect size and experimental factors[5] as input and produce the power of a given regimen as output. Essentially, this allows researchers to solve for a missing parameter if only two are known. For example, if they have a target power level in mind and an estimated effect size from a previous study, then they could use a power calculation to find the missing parameter (the experimental factors) that would let them reach that desired power level. Ideally, a researcher would use these calculations to determine power *a priori* (before the experiment) to estimate how many participants and/or items are necessary to reach a target power level, to avoid conducting an experiment that is not likely to detect an effect that exists (underpowered) or wasting resources by recruiting too many participants (overpowered).

To do this type of analysis, researchers need to first determine what sort of statistical analysis they will use for their experiment. As previously mentioned, masked priming lexical decision task experiments tend to be repeated measures experiments that violate the assumption of independence, which necessitates specific types of analyses that can properly account for the variation in the data as a result of these violations. These are often mixed effects regression models that partition out the patterns in the variation, however, choosing this type of analysis over more traditional analyses (for example, analyses based on *t*-tests, *F*-tests, or *z*-tests) makes calculating the necessary power more complicated.

[5] The use of the phrase "experimental factors" is a bit vague here intentionally, as the particular factors depend on the different ways to calculate power for different types of experiments: For example, some power calculations for repeated measures experiments specify the number of participants and the number of items, while power calculations for independent samples experiments only specify the number of participants

4.1.1 Power for ANOVA-based analyses

Formulae to calculate power for experiments based on traditional statistical tests have been available for decades, for example, Brown and Benedetti's 1977 formula to determine power in comparisons of two groups using a *z*-test, and some specialized programs such as G*Power (Faul et al., 2009) have been developed to make these calculations easier for less-technically oriented researchers: Using this program, researchers can specify two of three of their expected power, effect size, and number of participants, and the program will plug these into the formula to determine the missing variable, as shown in *Figure 4.1*[6]

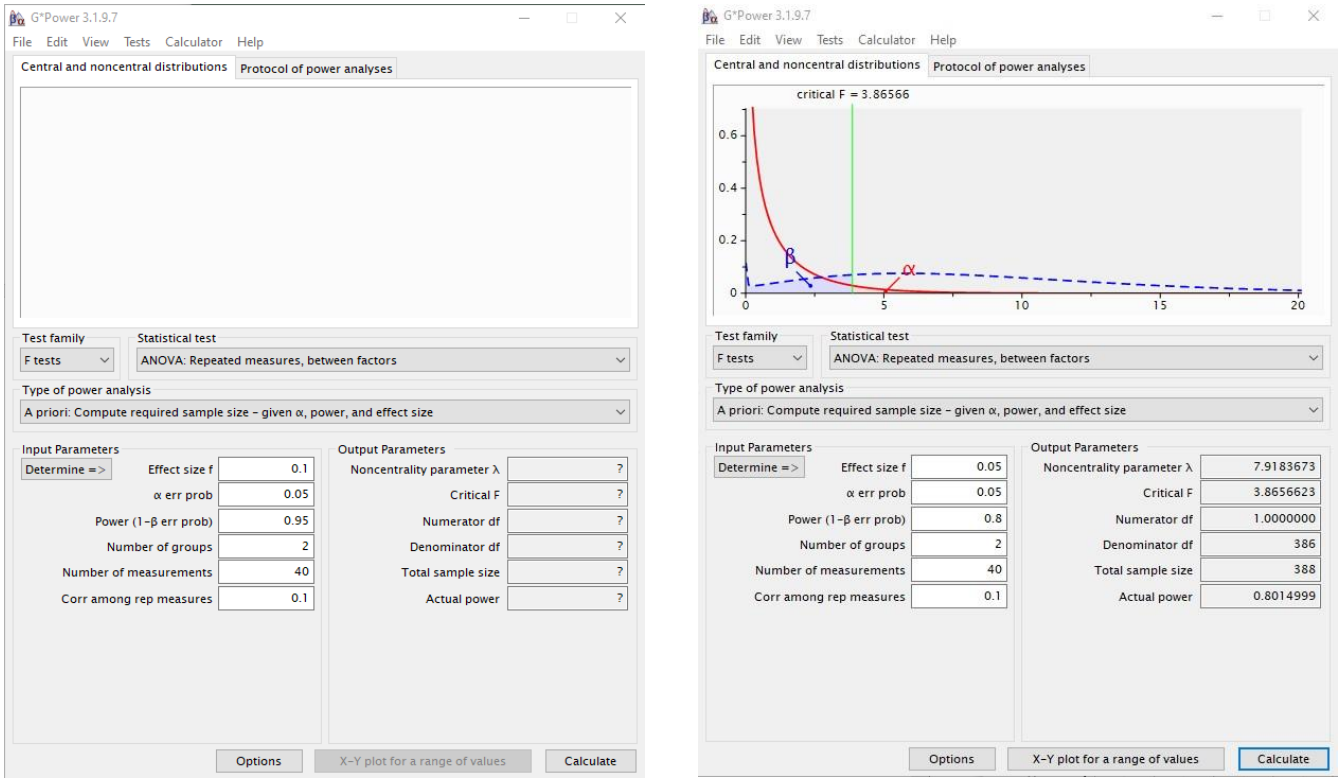

*Figure 4.1: Screenshots of the G*Power interface showing a power calculation for a one-way ANOVA*

[6] This figure is included to illustrate the usability of the graphical interface, and the values inputted are not relevant for this project.

For example, in G*Power, a researcher can just specify their estimated effect size, alpha and beta, as well as their experimental factors (i.e. number of measurements) and then G*Power estimates a sample size using the appropriate formula for the selected test. This makes it easy for researchers to calculate power *a priori* for a variety of analyses, however, G*Power (and power-calculating applications like it) do not have functionality for mixed regression models. Researchers using ANOVA-based analyses may find such apps useful, but as previously established, ANOVAs do not allow for all of the variance in the data to be captured by the model. This can sometimes lead to less accurate conclusions drawn about the effect size compared to modelling techniques that *do* incorporate all the variation, like the mixed effects regression models that have become commonplace in the field. To calculate power for these analyses, researchers must use other tools.

4.1.2 Power for mixed effects regression models

For mixed effects regression models, power calculations can be done in a variety of ways, either by using formulae developed for mixed effects models or by using methods involving simulation. In terms of formulae, Westfall et al. (2014) have developed a method that incorporates variances from random effects into the variance portion of a power formula and developed an online calculator applet based on this formula. The calculator includes the usual parameters of effect size and experimental conditions as parameters, but also includes Variance Partitioning Coefficients (VPCs), which are unique to this method of power calculation. Recall from the earlier discussion of regression models that the outputs are coefficients for each fixed and random factor.

These coefficients correspond to how much variation is attributed to each effect, and VPCs function similarly, as they estimate how much variance each random effect contributes to the model. However, this means that the output of a regression model is necessary to calculate VPCs. To use the formula and calculator from Westfall et al., the estimated coefficient for each random factor from the model is turned into a VPC by dividing the estimated coefficient for a given factor by the sum of all the estimated coefficients for random factors in the model. This technique allows researchers to incorporate variance due to random effects into a power calculation for a more precise estimate, compared to using methods developed for studies where the assumption of independence is not violated.

Alternatively, power can be estimated for these types of experiments by using a simulation. These techniques generally work by extracting (or simulating) a random sample of data from a dataset, applying a statistical test to the sample, and then repeating this process with a different random sample from the same original dataset. In each run through this process, the statistical test is the same and is applied to determine whether there is a statistically significant difference between the groups. Assuming there is a difference between the groups in the original dataset, then the number of runs where the test indicated a statistically significant result divided by the total number of runs gives the likelihood of the test detecting a difference that exists—i.e., statistical power. This method is flexible and allows researchers to perform *a priori* power calculations for any type of analysis based on significance testing, including analyses on the output from a linear mixed effects model.

To use this to calculate power *a priori* for a specific experiment investigating a specific phenomenon, researchers can adjust the parameters of the dataset to mimic the effect that they're studying. For example, the effect size in the dataset can be manipulated by adding a constant value to the data from a given group to mimic a smaller or larger difference, depending on the effect size that the researcher predicts based on previous studies of the same phenomenon. The number of participants or items to include in a study can also be tested *a priori* by adjusting the samples to include different numbers of participants or items (for example, removing participants or items from the dataset or simulating new participants or items from the existing data). This can help researchers determine the minimum requirements for an adequately powered experiment without being overpowered. And of course, choosing the correct dataset for sampling is of vital importance when using simulation techniques. For example, reaction time data from a visual perception task would not be a good basis for a researcher interested in performing an experiment with a lexical decision task, despite both being datasets of reaction times. The distributions of data could be different due to the task differences, or it could be the case that one task is more difficult, or a number of potential differences could impact the data in various ways. Additionally, reaction time data in particular is always skewed due to perceptual limitations (Lo & Andrews, 2015), so it is essential to choose a sampling dataset that is closest to the type generated by the experiment for this type of data. For example, a dataset from lexical decision task where the stimuli are single words would have a different distribution than a dataset from a grammaticality judgement where the stimuli are whole sentences, due to the differences in the length of time that is necessary for participants to process the stimuli. Differences in variation between the sampling

dataset and the experiment dataset could affect the power analysis if the researcher was applying a statistical test based on analysis of variance (for example, an ANOVA) as part of their simulations.

4.2 The current experiment

Having reviewed the options for statistical testing and calculating power, this project will use a linear mixed effects regression model to assess statistical significance and conduct an analysis of power using simulation methods, following Brysbaert & Stevens (2018). This method allows for a more generalizable understanding of the effect of sample size on power in detecting small effects, compared to the Westfall et al. (2014) calculator that would require computing post-hoc power for each study with Mandarin-English bilinguals, as it relies on repeated sampling. Brysbaert and Stevens' methodology of using the R package *simr* (P. Green & MacLeod, 2016) to perform the simulation and the dataset of reaction time data from a primed lexical decision task (Perea et al., 2015) will also be adapted for this study. This study was chosen because it was very overpowered for the size of the effect that it was trying to examine, making it a good candidate for the techniques in this study that artificially decrease the power to examine what happens to the results.

The Perea et al. (2015) dataset is from an experiment investigating the effect of orthographic case alternation (eg. cAsE aLtErNaTiOn) in priming. In this experiment, 40 monolingual participants performed a lexical decision task where there were four conditions for primes based on two factors: (1) whether they were related to the target, and (2) whether the prime was written in a mix of uppercase and lowercase letters. There

were 120 real word targets in total, repeated across the prime conditions. The authors found a significant main effect of repetition (i.e. whether the prime was the same word as the target), which is to be expected, but no effect of case alternation, and no interaction. When the non-significant effect of case alternation is removed, this results in a dataset with 40 participants, 120 items, and a 37ms difference between the two priming conditions (unrelated and repeated). After trimming for incorrect responses and outliers (trials with a reaction time lower than 250ms and above 1500ms), this leaves a dataset of 4512 trials. To adapt this dataset for re-analysis, a constant value will be added to the reaction times for trials with a repeated prime to mimic a priming effect of 11.38ms, which was determined to be the average backward masked priming effect size from the recent Wen and Van Heuven (2017) meta-analysis. This technique follows from Brysbaert and Stevens' (2018) and Kumle et al.'s (2021) recommendation as adding a constant is a linear transformation (i.e. it does not affect the relationships between datapoints in the dataset) and therefore does not distort the variation in the dataset, which is crucial for the modelling analysis (Winter, 2019).

The power simulation will be done using *simr* (P. Green & MacLeod, 2016)[7]. Specifically, this tool will be used to estimate power at different levels of a given random effect, such as number of items or number of participants. To do this, one random effect is held constant while power is calculated at different levels of the other: In this experiment, power will be estimated for an experiment with 40 participants and from 100

[7] Although other packages for calculating statistical power using simulation exist, like *mixedpower* (Kumle et al., 2021), *simr* was chosen for its flexibility and ability to estimate confidence intervals. In their paper introducing *mixedpower* and comparing its performance against other packages for simulation-based power analyses, Kumle et al. demonstrate that the power estimations in *mixedpower* and *simr* generate comparable results and advocate for the complementary use of these two similar tools.

to 200 items in intervals of 20 (i.e. at 100, 120, 140, 160, 180, and 200 items). Power will also be estimated for an experiment with 120 items and between 30 and 80 participants in intervals of 5 (i.e. at 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, and 80 participants[8]. Since the original dataset does not include this many items, the *extend* feature will be used to simulate more items and participants, based on the distributions of the existing items and participants. In this simulation, there will be 1000 runs at each level (i.e. running analyses on 1000 samples of 40 participants and 100 items, then 1000 samples of 40 participants and 120 items, and so on), based on recommendations from Kumle et al. (2021). In each run, the data will be modelled by a linear mixed effects regression model using R package *lme4*, with reaction time as the outcome and prime type as the only main effect, with random intercepts by participant and item. This model formula was selected by building a maximal model with random slopes and intercepts by participant and item (Barr et al., 2013) and comparing it to models built with random slopes by participant only and by item only, as well as the base model with only intercepts (Bates et al., 2018). The base model was chosen as it was the best fit for the distribution (Akaike, 1994). This model will also use an optimizer from the R package *optimx* for the mixed effects model to facilitate model convergence. In *lme4* syntax, the basic model looks like:

$$RT \sim prime\_type + (1 \mid participant) + (1 \mid item)$$

A *t*-test on the estimated coefficient for the main effect of prime type will be used to assess significance in each sample, with an alpha set to 0.05 for significance.

To recap, the novel contribution from this experiment will be estimated power levels from simulated experiments examining an effect size of 11.38ms with different

[8] These ranges were decided from preliminary power analyses demonstrating that fewer than 100 items or 30 participants resulted in power below 50%.

numbers of items and participants to see how power is affected by these experimental factors, specifically for masked priming lexical decision tasks. The next sections will present the results from these simulations, make recommendations for researchers looking to conduct experiments of this type, and contextualize these findings in the literature.

Chapter Five: Results

The results of the power analysis demonstrate that the minimum 1600 observations per condition for experiments using mixed effects models, as recommended by Brysbaert and Stevens (2018), is not sufficient to achieve statistical power of 80% when assessing a priming effect of 11.38ms. Power simulations of an experiment with 40 participants and between 100 and 200 items demonstrated that power reached 80% at around 145 items, or a theoretical 5800 observations per condition, as shown in Figure 5.1. To get a more precise estimate, a power calculation was performed for samples of 40 participants and 145 items in the same way as they were calculated in the power curve (counting the percentage of correctly-identified significant differences within 1000 simulations of the experiment with 40 participants and 145 items). In these analyses, the

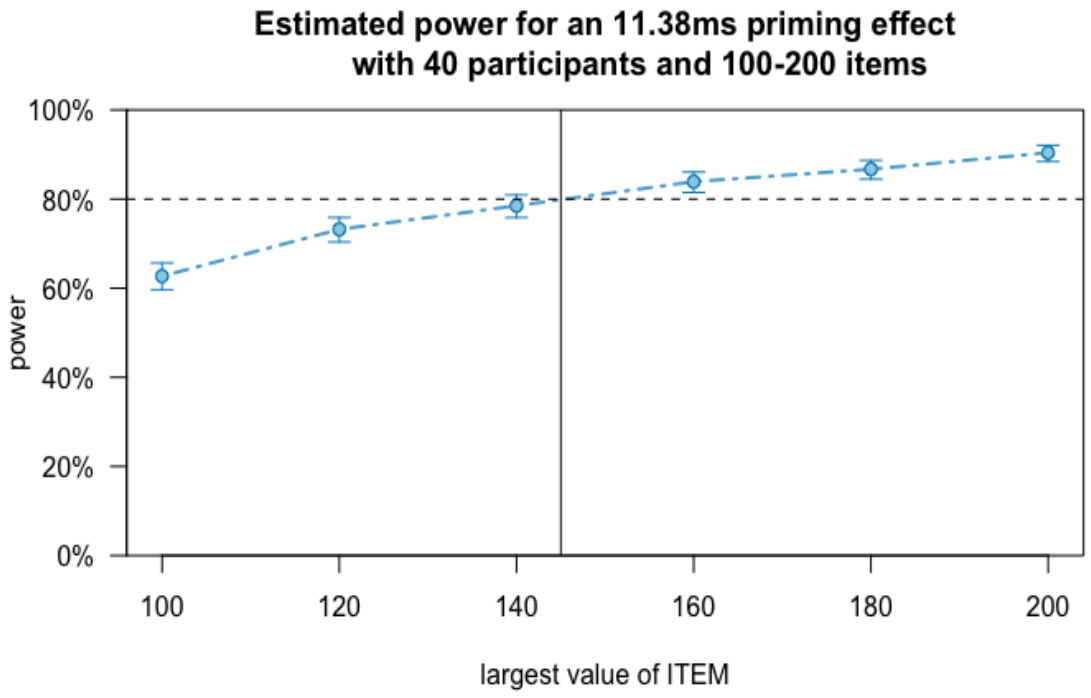

Figure 5.1: Estimated power for experiments with 40 participants and 100 to 200 items. For an 11.38ms difference between groups, about 145 items were necessary to reach about 80% power.

estimated power was 80.0% (95% CI[9] 77.4, 82.4) but this was achieved with 5444 observations. This value is smaller than the 5800 observations one would observe by multiplying the number of participants and number of items (145 items * 40 participants = 5800 observations) due to trimming of data for outliers and trials where the participant responded incorrectly. This accounts for a 6.1% data loss, which is within the range when following this procedure for trimming data; for example, Lee et al. (2018) trimmed 5.4% and 4.4% of their data in their experiments after removing trials where participants responded incorrectly (1.1% and 0.7% respectively) and outliers beyond two standard deviations (4.3% and 3.7% respectively).

Similarly, simulations of an experiment with 120 items and between 30 and 80 participants showed that 45 participants were necessary to reach a power level of about

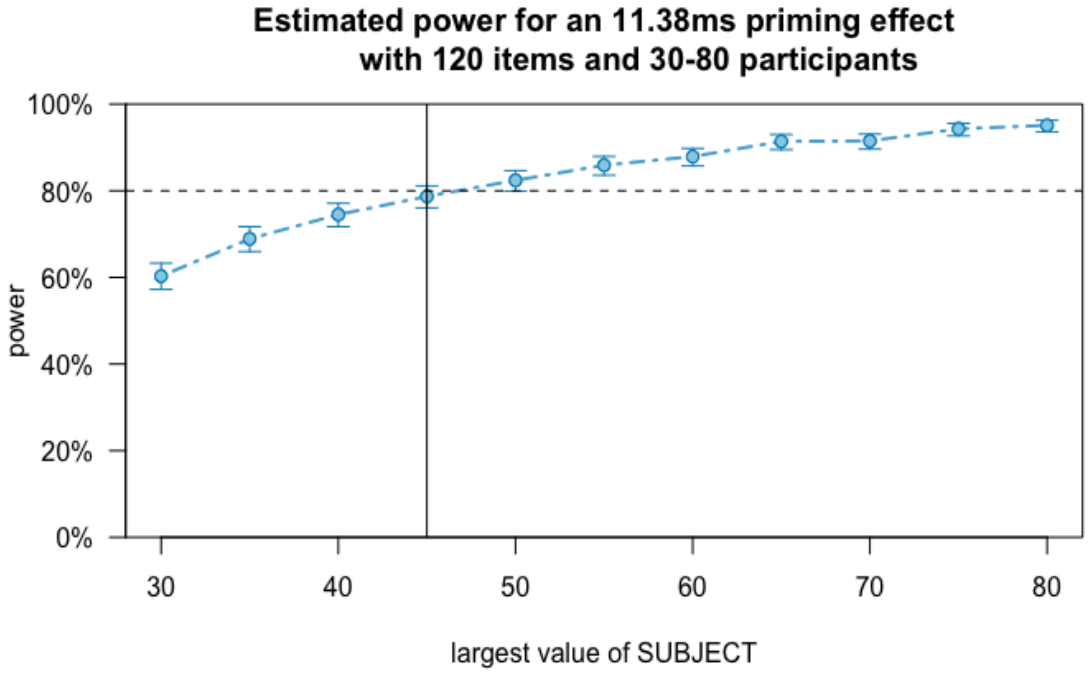

Figure 5.2: Estimated power for experiments with 120 items and up to 80 participants. For an 11.38ms difference between groups, about 45 participants were necessary to reach about 80% power.

[9] 95% confidence interval, i.e. if this experiment were repeated infinite times, 95% of the calculated power values would fall within this range

80% (*Figure 5.2*). And again, to get a more precise estimate, a power simulation was performed on a subset of the data with 120 items and 47 participants, which reached 80.4% power (95% CI 77.8, 82.8) with 5343 observations per condition. Again, this does not line up perfectly with the expected value of 5640 (120 items * 47 participants = 5640 observations) due to data trimming of 5.3%.

Taken together, the results indicate that while there is some variability within the precise number of observations necessary to reach adequate experimental power, somewhere in the neighbourhood of 5400 observations is required to reach 80% power in experiments investigating an 11.38ms priming effect. However, it is important to note that this criterion should be taken as a guideline and not as a hard rule, as the results also show that variation within the random factors of item and participant can affect the power of an experimental design (demonstrated by the different numbers of observations necessary when holding item constant vs. holding participant constant). Rather, these results should help shape researchers' intuitions around statistical power and experimental design, guiding them to perform their own power analyses *a priori* (ideally from pilot data generated using their experimental items) to most accurately model what their specific experiment will need to reach adequate power. If this is not available, publicly available open datasets (such as Perea et al. (2015)) are also a good option for researchers to approximate power *a priori*, although researchers should endeavour to find a dataset that closely mirrors their experiment.

The implications for studies with Mandarin-English bilinguals is starker. As per *Table 1.1* reviewing experimental factors in Wen and van Heuven's 2017 meta-analysis of backward lexical priming experiments, the average number of participants in studies

on Mandarin-English bilinguals was 28 and the average number of items used was 17, giving an average of roughly 453 observations per condition. This is smaller than the average for all backwards masked priming experiments (720 observations from 29 participants and 25 items, on average), and stands out even more when compared to other backwards masked priming experiments studying cross-script bilinguals, which used an average of 34 participants and 38 items, giving an average of 1312 observations per condition. And while all of these fall short of both the recommendation from Brysbaert and Stevens (2018) to aim for 1600 observations per condition and the suggestion from this experiment to use 5400 observations per condition, experiments with Mandarin-English bilinguals are especially remarkable. Given this, it is perhaps not surprising that studies with other cross-script bilinguals have reported significant backward masked priming when this effect has not been found consistently in experiments with Mandarin-English bilinguals, as this study illustrates how experimental power increases with the number of observations. This suggests that previous experiments with Mandarin-English bilinguals may have been underpowered, leading to the lack of significant results, and that studies investigating backward masked priming in other cross-script bilinguals did not suffer from this problem as much. It is then especially crucial for future studies with this population to conduct power analyses *a priori* to ensure that the results are reliable.

Chapter Six: Discussion

After conducting a simulation-based power analysis to estimate the number of observations necessary to observe a 11.38ms priming effect, we can return to the research question to determine if the lack of a consistent and significant backward masked priming effect in experiments with Mandarin-English bilinguals was caused by issues with statistical analysis, or if it is due to an actual lack of this effect in this population. The results suggest that a deficit in properly-powered experiments in the literature on this population is a more likely cause, rather than this effect not existing in the population. In the simulations, roughly 5400 observations per condition were necessary to achieve 80% power to detect an effect this small, but the average number of observations from previous experiments with Mandarin-English bilinguals was around 453, according to a meta-analysis of backward masked priming experiments (Wen & van Heuven, 2017). This is supported by literature with other cross-script bilinguals, where this effect has been consistently found significant in experiments using an average of 1312 observations per condition, and there is no basis within the currently-understood models of the bilingual lexicon to explain why a backward priming effect would exist for speakers of some languages, but not others. Given this, the current chapter will discuss the narrow implications of what this means for researchers interested in conducting these types of experiments with Mandarin-English bilinguals, as well as the broader implications of how researchers can incorporate these results into their approach to statistical analysis, including in cases where it might not be possible to conduct adequately-powered experiments.

6.1 Implications for masked priming experiments

First, I will review how the results from this experiment fit into the current literature around backward masked priming. The results from this experiment are in line with the general calls from researchers to conduct more high-powered studies in the field whenever possible (Brysbaert, 2020; Brysbaert & Stevens, 2018; Nakayama et al., 2016; Wen & van Heuven, 2017), although this experiment builds on these calls to offer specific recommendations for researchers studying small effect sizes. For example, Wen and van Heuven's meta-analysis demonstrated that items per cell (i.e. number of experimental items) significantly affected the size of the priming effect in backward masked priming studies, which led them to recommend that researchers use a "sufficient number of items per cell" to investigate this effect (p. 883). Similarly, Nakayama et al. point out that previous work in this field may have been underpowered, but do not offer recommendations for how researchers might address this. Brysbaert and Stevens build on this and give a number, suggesting that 1600 observations per condition would be sufficient, although they base this recommendation on analysis of a dataset with a 37ms priming effect, which is much larger than the average for backward masked priming experiments in Wen and van Heuven's meta-analysis (11.38ms). By using the same methodology recommended by Brysbaert and Stevens, this experiment suggests that 1600 observations is not enough when examining small effect sizes, and that somewhere around 5400 is necessary, although this depends on factors affecting variation in the data. This is not surprising, given that the number of observations necessary for reaching adequate power increases as effect size decreases. Yet given how the number of observations required to reach 80% power differed depending on whether number of

participants or number of items was held constant, the results from this experiment concur with Brysbaert and Stevens in urging researchers to conduct power analyses *a priori* with these tools, ideally using data from a pilot experiment, to determine what will be right for their experiments. Ultimately, these results are meant as a suggestion to help researchers see how adding items or adding participants might affect the statistical reliability of their results (and error rates), and to give a rough estimate of what to aim for, although each experiment will be different.

One area where these results differ from the literature, however, is how well the results reflect the experimental practices of studies that *have* found a significant backward masked priming effect in cross-script bilinguals. For example, Nakayama et al. (2016) detected a significant backward priming effect with 30 items and as few as 34 participants in their experiments, while Lee et al. (2018) found a significant backward priming effect with 30 participants and 30 items. This comes to roughly 1020 observations for Nakayama et al. (2016) study and only 900 observations for the Lee et al. (2018) study, not accounting for data trimming due to incorrect trials and outliers. This may give the impression that the results from simulations are not reliable models for how experiments would function in real life, but expecting the simulation data to match the literature exactly is an overextension of the results. The results are intended to guide researchers with a ballpark idea of what sort of data collection might be necessary for adequate power, but other experimental factors impacting the variation in the data will also heavily impact this, as each sample population will have their own characteristics. For example, both of the other experiments restricted their participants based on their scores on a test of second language proficiency in English (the TOEIC), since they were

investigating the effect of L2 proficiency on the backward priming effect. Additionally, both studies used participants with highly homogenous experiences in their first and second languages (factoring in aspects such as age of acquisition, duration of study, and type of study) which could have potentially reduced variation in their data, affecting the power of their experiments. Since neither reported performing an *a priori* power analysis, it is impossible to know whether these experiments were adequately powered, given their particular experimental conditions. For example, one experiment in Nakayama et al. reported a 22ms priming effect, which suggests that a power level of 80% could have been reached with fewer than 5400 observations, but again, whether this was reached is impossible to determine without an *a priori* power analysis.

6.1.1 Limitations of the current project

In terms of limitations of this experiment, it's important to note that the simulations in Chapter Five were conducted on data from participants in a masked priming lexical decision experiment, examining the effect of case alternation in primes on target reaction time (Perea et al., 2015). There is very little known about the participants in this experiment, apart from the fact that they were monolingual undergraduate students who were native speakers of Spanish. This highlights a key limitation of using simulation modelling for power analysis, which is that the results are dependent on the data used for the simulations, so the most accurate results will be obtained by using data from an experiment that most closely resembles a given study of interest. Given that data from bilinguals often have more variability than data from monolinguals (Brysbaert, 2020) and that power decreases with increased variability, it is possible that the recommendations

drawn from this project are underestimating the number of observations required to reach 80% power, when looking at effects this small. However, this just underscores the key takeaway from this project, which is that researchers should endeavour to perform *a priori* power analyses to ensure that their methodology is appropriate for their research questions, ideally based on data from pilot experiments using their experimental items.

6.1.2 Implications for models of the bilingual lexicon

The possibility of a lack of statistical power being responsible for previous studies with Mandarin-English bilinguals failing to detect a backward masked priming effect also helps us interpret the data with respect to empirical support for models of the bilingual lexicon. Recall that Mandarin-English bilinguals are the population in which the backward masked priming effect has been most elusive, driving the development of models like the SENSE model (Finkbeiner et al., 2004) and the Episodic L2 model (Witzel & Forster, 2012) that propose a qualitative, rather than qualitative difference between forward and backward priming. Therefore, determining whether this is due to methodology or due to a difference in the way that Mandarin-English bilinguals process language (compared to other bilinguals) can help us re-assess the amount of empirical support for them. If it is the case that issues with power and general methodology are behind this specific discrepancy, as this project and other researchers suggest (Lee et al., 2018; Nakayama et al., 2016), then it is most likely that these models, while still being useful for explaining other phenomena related to bilingual lexical access, might not account for the data at hand. For example, in the three masked priming studies with Mandarin-English bilinguals that support either the SENSE model or the Episodic L2

model (Jiang, 1999; Jiang & Forster, 2001; Luo et al., 2013), there were 10 experiments conducted to test for a backward masked priming effect. These had an average priming effect of 12.2ms, with an average of 27 participants and 14.8 items, and only two of them found a statistically significant backward priming effect. This comes to an average of 400 observations per condition to detect a priming effect of 12.2ms (on average), which shows that those studies were likely underpowered, given the results of the simulations. Additionally, there are some anomalous results within those studies, showing potential Type M and Type S errors (detecting an effect in a sample that is much larger, smaller, or in a different direction than it is in a population). For example, in Luo et al. (2013), one of their experiments reported a 48ms priming effect and another reported a -4ms priming effect. Given that the average backward masked priming effect is 11.38m, the largest backward priming effect that had been reported before this study was 28ms (Schoonbaert et al., 2009), and only one out of 19 experiments previous to this had reported a negative backward priming effect (Finkbeiner et al., 2004), both the magnitude and direction of those results raise some concerns about the statistical validity of these analyses. This is supported by evidence showing that Type M and Type S errors are likelier in studies with low power (Gelman & Carlin, 2014). These discrepancies in the empirical support for the SENSE model and the L2 Episodic Memory model raise some questions about whether these models should continue to be investigated in future studies, or if the field should focus on discerning between models that predict a quantitative and not qualitative difference.

6.1.3 A Bayesian perspective

So far, we have reviewed the previous work into Mandarin-English bilinguals and examined the analyses through the lens of statistical power, but we have not yet broached the question of whether Mandarin-English bilinguals could be, in fact, different from other types of bilinguals. This is a hypothesis that would counter every empirically-supported theory of the bilingual lexicon, as this effect is predicted to exist by those theories and has been shown to exist in other populations of bilinguals: In fact, incorporating it into current theories would require us to reject the strong empirical evidence in support of the effect based on the results of a few studies with Mandarin-English bilinguals. However, if we wanted to seriously engage with this question, we would have to step outside of the perspective of NHST and into the world of Bayesian statistics. In a Bayesian analysis, the focus is on probability distributions, not point estimates, like a *t*-value or a *p*-value (Norouzian et al., 2018; Vasishth et al., 2018; Winter, 2019). This is an understanding of statistics that understands the data as fixed, but the hypotheses as flexible, which allows us to ask questions about the probability of the hypothesis, given the data. This runs counter to frequentism-based NHST, which treats the hypotheses as fixed, and asks whether the data are probable, given the hypothesis.

To understand the steps involved in a Bayesian analysis, it is important to focus on probability distributions. Differing from an NHST analysis where researchers would decide on a null hypothesis, in Bayesian analysis, researchers must specify a prior distribution that captures information about the likelihood of the effect. The shape of this distribution is generally decided based on previous knowledge of the effect that the

researcher is testing (for example, a prior distribution for the backward masked priming effect would probably try to rule out the likelihood of an effect size larger than say, 40ms, based on the literature). In a Bayesian analysis, this prior distribution is combined with the distribution of the data gathered to generate a posterior probability distribution to determine the likelihood of that effect, illustrated in *Figure 6.1*[10].

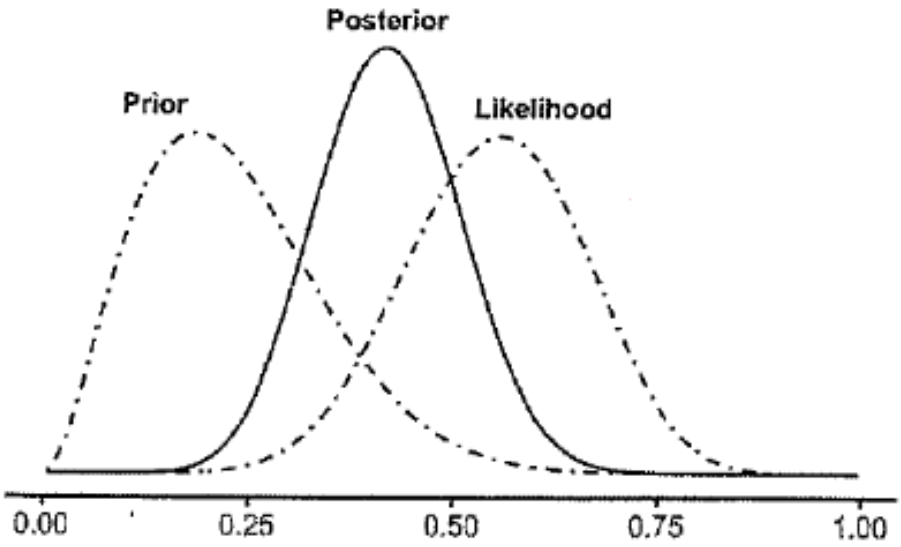

Figure 6.1: The prior, likelihood, and posterior in Bayesian analysis (Garcia, 2021)

Effects are reported in terms of areas of high probability density in the posterior, and the fact that the posterior is a probability distribution allows for researchers to reason about the confidence of their results. This reflects the one of the philosophical tenets Bayesian statistics, which states that we already have information about the world (i.e. a prior) and that we can update that information with new data. Strongly held information can be reflected in a tightly-specified prior, which requires stronger evidence to shift the

[10] From *Data Visualization and Analysis in Second Language Research*, by Guilherme D. Garcia, 2021, p.213. Copyright 2021 by Routledge

posterior, but effects that we are not so sure about can be specified as weakly informative priors, which can be updated more easily. For example, if a researcher wanted to test whether the backward priming effect exists in Mandarin-English bilinguals, they would conduct a lexical decision task as usual, but they would have the opportunity to specify some prior knowledge about this effect. For example, setting the prior distribution to a normal distribution with a mean of 0 and a standard deviation of 20ms would tell the model that values between -20ms and +20ms are likely, and that values larger in magnitude than 40ms are highly unlikely, which is reflected in the literature on backward masked priming in lexical decision tasks. Then they would gather data and fit a Bayesian model to their data, which combines their prior with the data that they gathered to generate the posterior distribution. This allows them to assess the likelihood of a given value of the effect via point hypothesis testing in the posterior.

This shift in statistical perspective from NHST to Bayesian modelling allows researchers to conduct studies testing the null hypothesis directly and ask whether the priming effect would equal zero, which is impossible in NHST: The null hypothesis is treated as a hypothesis that needs data to disprove, and so a null result can only tell us that the data fail to reject the null hypothesis, but it cannot determine whether the null hypothesis is likely. Other strengths of Bayesian inference include allowing researchers to incorporate what is already known about the effect via prior distributions and testing for effects that are predicted theoretically, as well as Bayesian models being able to better handle the complexity of data with multiple dependencies. For example, repeated measures experiments can generate data with highly complex hierarchies, which can cause convergence issues for generalized regression models, particularly when the data

are sparse. This often results in poorly-fitted models or models that fail to converge (or both), which cast doubts on the statistical validity of those analyses. Bayesian models, however, handle this complexity with ease and reflect the uncertainty due to the sparseness of the data in wide posterior distributions with large ranges of estimates for areas of high posterior density, rather than convergence errors. This allows researchers to have a quantitative understanding of the uncertainty of their results and to be able to reason about that uncertainty, which is not available information in NHST-based analyses. However, this comes at the cost of higher demands on computational resources and potentially a higher risk of user error, as these types of models are less established within the field, but accessible guides to Bayesian analysis are available (Garcia, 2021; Nicenboim & Vasishth, 2016; Norouzian et al., 2018; Vasishth et al., 2018), as well as R packages that make the process of fitting a Bayesian model very similar to fitting a generalized regression model (Bürkner, 2018). We must also keep in mind that a Bayesian model is still a model, and so other methodological considerations, such as controlling for proficiency, using appropriate masked priming protocols, and ensuring that experiments are properly-powered are also respected.

6.2 Moving towards open science

In terms of implications of these results on statistical practices in the field, there are some key takeaways for researchers examining small effect sizes, in general. When thinking about appropriate sample sizes for detecting these effects, it is important to keep in mind that low-powered experiments examining a small effect size, paired with a tendency to publish only significant results (Button et al., 2013; Fanelli, 2012; Franco et

al., 2014), can lead to the impression that these results can be replicated with the same numbers of items and participants. As Wen and van Heuven (2017) have established, the backward priming effect is small but present, even though studies have had varying degrees of success in capturing this effect. However, due to the lack of reporting around power and publication bias (among other factors), it is difficult to tell whether their results could be considered typical or atypical, as we do not know how many studies that did not produce statistically significant effects have not been published. Additionally, as Lindsay (2020) notes, replications that match the sample sizes used in the original study often fail to find statistically significant effects. As such, there have been proposals in the literature to establish more transparent methodological practices, such as publishing the proposed methodology and analysis method for a study before collecting data (i.e. preregistration) and sharing anonymized data collected from a given study, to address the concerns about questionable methodological practices in previous work. The following sections will review the role of power in replication and the movement towards open science, as well as alternative approaches to statistical inference and some broader views on statistical practice overall.

6.2.1 Replication, replicability, and power

First, we will briefly review the literature on replication and replicability to help us better understand why it's important for researchers examining the backward masked priming effect to consider statistical power. Recalling Chapter Three, *frequentism* is the statistical approach underlying NHST, which relies on repeated sampling from a population and testing those samples to determine whether an effect exists in the

underlying population. This assumes that as more data are collected, the estimated effect gets more and more accurate. If an effect is consistent across multiple experiments, then each repeated experiment strengthens our confidence in those effects being present in the underlying population. Power, then, is an important part of replicability: if an experiment is not well-powered, then it is more likely to fail to detect an effect that exists, and replications that fail to find an effect start to beg the question of whether the original effect existed in the first place. This is theoretically controlled by α and β in experimental contexts, which control for Type I and Type II errors respectively, but in practice, a 5% α level and a 20% β level does not tend to be respected. According to a review of research practices in second language acquisition by Plonsky (2013), studies tend to run multiple comparisons without adjusting α to compensate for the increased risk of Type I error that this causes, which can lead to α being higher than the commonly-accepted 5% in practice. Additionally, Plonsky warns of a "power problem" (p. 678) in the field of second language acquisition, based on the small sample sizes typically used. He reports on an empirical review of 174 studies on L2 interaction from Plonsky and Gass (2011), who estimated a post hoc power of 56% for the studies involved, based on the median number of participants (19) and average effect size ($d = 0.74$), which results in a β of 44%, much higher than the typical 20% accepted Type II error rate. These findings support other calls within the field to increase statistical power (Brysbaert, 2020; Button et al., 2013; Greenland et al., 2016; Nicenboim et al., 2018; Plonsky, 2013), as well as the main conclusion from the present project, which emphasizes how experimental power has historically been lacking. Efforts to address this are crucial for increasing the replicability of our experiments and strengthening our confidence in our results.

Additionally, the attention paid to methodology in second language acquisition over the past decade echoes the way that psychology, as a discipline, has grappled with similar concerns after some stark publications demonstrated how researchers could manipulate statistical analyses to find a significant effect where none exists. For example, in 2011, Bennett et al. drew attention to this issue when they released their fMRI study on the ability for post-mortem salmon to detect a difference between prosocial and antisocial photos. They placed a dead fish into an MRI machine and displayed images in the different conditions while measuring their brain activity, which is a protocol that they had used with human participants. From this, they concluded that there was a statistically significant difference in the brain activity of the fish when viewing images in the different conditions. This physically impossible but statistically significant effect was achieved by engaging in *p*-hacking, or performing different analyses on their data until a significant effect was found. When they corrected α to compensate for performing multiple comparisons, as is statistical best practice, the statistically significant effect disappeared. Bennett et al. independently published their work to highlight the elevated rate of Type I error when performing multiple comparisons without correction (i.e. increasing the family wise error rate) and to warn other researchers in the field of the consequences of this practice. Similarly, Bem (2011) conducted an investigation into precognition (i.e. being able to forecast the future), and found a statistically significant result in eight out of nine studies, concluding that personality traits linked to extroversion were related to participants' ability to predict the future. Both studies were conducted to exploit specific methodological weaknesses in statistical practice in psychology at the time, particularly Bennett et al.'s work demonstrating that multiple comparisons can

drastically inflate the risk of Type I error. These studies which sparked vigorous debate about methodology and replicability and growing concern about the so-called “replication crisis” (Pashler & Harris, 2012).

As a response to this, the Open Science Collaboration (2015) coordinated a large-scale series of replications of 100 studies from the 2008 issues of three major psychology journals, where teams of researchers aimed to replicate the results from the original study using a high-powered experimental protocol and the original materials where available. While 97% of the original studies had significant effects, only 36% of the replications had significant results, and only 39% of the replications were judged to have replicated the results of the original. This was the first large-scale empirical investigation into the so-called “replication crisis,” which sparked spirited discussion on the validity of older results and research into experimental design, as well as another large-scale replication attempt from Klein et al. (2018) that aimed to investigate the effects of experimental setting and cultural context on study replication. This team selected 28 studies to replicate with high-powered protocols and found that approximately 50% of them had replicated the statistically significant effect of the original, with the median effect size shrinking from $d = 0.60$ across the original studies to $d = 0.15$ in the replications. Ultimately, this variation in the effect size was attributed to the properties of the effect being studied, rather than the methodology of data collection (for example, online experiments versus experiments performed in a lab) or the cultural context in which the original data were replicated. These replication rates are much lower than expected, if we assume that the original experiments were designed to have 80% power. This harkens back to Plonsky and Gass' (2011) analysis demonstrating a rough average of 54% power in the studies

that they included in their review of methodology in second language acquisition and Smaldino and McElreath's (2016), estimation of average power in psychology hovering around 30–40%, showing that issues with underpowered experiments might be present across the board in the social sciences.

6.2.2 Open science practices

The so-called "replication crisis" and lackluster results from large-scale studies aimed at replicating effects (Klein et al., 2018; Open Science Collaboration, 2015) highlight patterns of methodological and statistical issues in psychology, which has spurred inquiry into methodological practices that enhance the replicability of results. For example, practices such as preregistering studies so that experiments with strong protocols will be published, regardless of the results, as well as sharing anonymized data openly have been proposed as ways to address these issues. These are some of the practices of the broader movement known as Open Science, which aims to make scientific knowledge "openly accessible, transparent, rigorous, reproducible, replicable, accumulative and inclusive" (Parsons et al., 2022, p. 314). While there are a number of practices within Open Science that are outside of the scope of this project, I propose that preregistering and sharing data are particularly crucial to address the lack of statistical power in backward masked priming experiments with Mandarin-English bilinguals. Given that the previous experiments in this field have generated mixed results, it is especially crucial to make sure that any further study follows the methodological recommendations of the literature in terms of having high experimental power, using an appropriate SOA, controlling and testing for proficiency, ensuring that experimental

stimuli are well-controlled, and other factors that might impact the variability of the data. For example, if a researcher wanted to preregister a backward masked priming study with Mandarin-English bilinguals, they would specify things like their research questions, experimental protocol (including details like how many participants they aim to recruit, what tools they'll use to assess linguistic background, the details of their backward masked priming task, their stimuli, and other procedural aspects), and their plan for analyzing the data, even before data have been collected. This allows for the possibility of their methodology being reviewed before data are even collected, which can help ensure that they are adhering to best practices. And while there is still a debate in the field about whether this is a worthwhile endeavour (Nosek et al., 2019; Szollosi et al., 2020), the process of pre-registering at minimum encourages researchers to think critically about their statistical practice and gives other researchers the opportunity to weigh in on any questionable methodology, which are steps in the right direction. This way, we give ourselves a better chance of generating results that will help us understand whether Mandarin-English bilinguals really are different from other types of bilinguals or whether it was methodological issues in previous work that prevented this effect from being detected, which will allow us to update our models of the mental lexicon.

Additionally, the call to share anonymized data from experiments is particularly crucial to the recommendation from this project, which is to conduct *a priori* power analyses based on pilot data. This experiment used a simulation-based power analysis, which means that the results are heavily based on the data used to generate the simulations. Thus, having data from the sampling population undergoing the same task is ideal to make sure that the results are as accurate as possible, which is why the emphasis

is placed on conducting pilot studies to assess power *a priori* as a key takeaway. In fact, one of the limitations of this experiment was not having access to data from bilinguals undergoing a lexical decision task to use for the simulations, which limits the strength of the recommendations about sample size and power: On IRIS, a repository for sharing experimental protocols and data in language sciences (Marsden et al., 2017), the only data from bilinguals in a lexical decision task is from an experiment examining word association, using semantic primes and not lexical primes (Eguchi et al., 2022), and it is not clear how the difference in prime type might affect the distribution of the data. If any of the previous studies examining the backward priming effect in bilinguals had made their data available for use on IRIS, this project could have made more accurate recommendations for reaching adequate power by using data that came from a much more similar population. Sharing data that specifically focuses on bilinguals in lexical decision tasks, using lexical primes, would give researchers who have not done a pilot study the opportunity to use simulation-based power analysis to make sure that they are conducting high-powered studies and not performing another underpowered replication of the same type of study that may have contributed to this inconsistency in the literature. After all, an underpowered replication of an underpowered study only tells us what we already know, which is that more rigorous study needs to be conducted to assess whether this effect exists.

6.3 What about when adequate power is not possible?

While this project focuses on the importance of statistical power, it is important to also consider areas of study where appropriate statistical power might never be attained.

Science is not limited to examining questions that can only be answered with hundreds of participants in a laboratory setting, and some important knowledge about the world can only be found in other ways, using techniques that are "un-ideal" from a statistical point of view. The aim of this project is not to prescribe a given method of analysis on all work done in this field, rather, it aims to help researchers reason about what might be an appropriate type of analysis for their given research question. There are some questions that inferential statistics, in general, might not be fit to answer: For example, in the study of endangered or minority languages, it might be the case that there do not exist enough speakers to reach statistical power for a given research experiment. If there is a specific effect that a researcher might be theoretically motivated to test for in this sort of experimental framework, a Bayesian analysis might help them capture the best estimate possible and reason about the uncertainty due to low sample size (Nicenboim et al., 2018), but this should be up to the researcher's judgement. Alternatively, researchers might consider adapting the tools used in this experiment to conduct *a priori* power analyses to determine if it is feasible to include more stimuli items to reach appropriate statistical power, however, this might result in a case where participants might be subjected to more stimuli than appropriate. This might prompt a researcher to conclude that statistical modelling is not the right tool for answering their research question, in which case, they could use the methodology from this project to argue for a non-parametric (or even non-inferential) approach to their analysis that breaks from the general trends towards parametric tests and regression modelling. For example, they might report descriptive statistics from a study with few participants and treat it as a case study, rather than trying to mold their project to fit the expectations of a specific type of

quantitative analysis even if they are reporting on an effect that is typically analyzed using statistical modelling. This would be a more statistically rigorous approach than attempting to perform statistical tests on a dataset that is incompatible with the framework and assumptions of NHST. After all, there is no one "right" way to analyze a given dataset, and this sort of plurality in approaches to data analysis and careful thought about whether a given analysis is appropriate for the data is encouraged by advocates for increased methodological rigor in linguistics (Roettger et al., 2019), which ultimately contributes to the reproducibility of results. The more researchers that critically question their statistical practices and engage in the practices of open science, the more information we have to work with to assess evidence for theories as a whole, outside of each individual experiment.

Conclusion

This project used a simulation-based power analysis to assess a potential statistical cause for the lack of a consistent and significant backward priming effect in masked priming lexical decision tasks with Mandarin-English bilinguals. The simulations reveal that the previous literature most likely suffered from issues with experimental power, much like many older studies in the field (Brysbaert, 2020; Nicenboim et al., 2018; Plonsky, 2013; Plonsky & Gass, 2011), and more statistically rigorous studies are necessary to determine whether there is empirical support for this effect. Finding this effect would help provide empirical evidence for models of the mental lexicon that predict this effect (Dijikstra & van Heuven, 2002; Kroll & Stewart, 1994; Schoonbaert et al., 2009) and confirm the findings from other cross-script bilinguals showing this effect (Dimitropoulou et al., 2011; Lee et al., 2018; Nakayama et al., 2016). The discrepancy in the literature is notable and worthy of further empirical study to confirm the theoretical predictions, or to give us information to update our models of the bilingual lexicon to account for the lack of empirical support in the face of rigorous testing. And of course, future studies of this effect must also properly control for confounds, such as proficiency and age of acquisition, as well as use proper lexical decision task protocols, as previous work has shown that these are also crucial to conducting rigorous experiments. The key takeaways from this project are for researchers to conduct power analyses *a priori* to ensure that they're conducting statistically-sound experiments, but also for researchers to think deeply about their experimental and statistical protocols to improve replicability in the field overall. Over the past decades of study, we have moved from understanding bilingualism as two monolinguals in one mind to understanding each bilingual as having

many complex factors that shape their experience, and our understanding of the statistical practices that we use to investigate these phenomena similarly needs to evolve. It is crucial for researchers to reason about their results and move from understanding their statistical tests as a binary tool to assess whether or not an effect exists to understanding each *p*-value as informed by all the decisions we make leading up to it so that we may continually move towards a clearer understanding of the bilingual lexicon.

Works Cited

Abutalebi, J., & Green, D. W. (2008). Control mechanisms in bilingual language production: Neural evidence from language switching studies. *Language and Cognitive Processes*, *23*(4), 557–582. https://doi.org/10.1080/01690960801920602

Akaike, H. (1994). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

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

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parsimonious Mixed Models* (arXiv:1506.04967). arXiv. http://arxiv.org/abs/1506.04967

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

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

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2011). Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. *Journal of Serendipitous and Unexpected Results*, *1*(1), 1–5.

Blanco-Elorrieta, E., & Pylkkänen, L. (2016). Bilingual Language Control in Perception versus Action: MEG Reveals Comprehension Control Mechanisms in Anterior Cingulate Cortex and Domain-General Control of Production in Dorsolateral Prefrontal Cortex. *Journal of Neuroscience*, *36*(2), 290–301. https://doi.org/10.1523/JNEUROSCI.2597-15.2016

Brown, H., Sharma, N. K., & Kirsner, K. (1984). The Role of Script and Phonology in Lexical Representation. *The Quarterly Journal of Experimental Psychology Section A*, *36*(3), 491–505. https://doi.org/10.1080/14640748408402173

Brown, M. B., & Benedetti, J. K. (1977). Sampling Behavior of Tests for Correlation in Two-Way Contingency Tables. *Journal of the American Statistical Association*, *72*(358), 309–315. https://doi.org/10.1080/01621459.1977.10480995

Brysbaert, M. (2020). Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition*, *24*(5), 813–818. https://doi.org/10.1017/S1366728920000437

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1), 1–20. https://doi.org/10.5334/joc.10

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Chen, B., Zhou, H., Gao, Y., & Dunlap, S. (2014). Cross-Language Translation Priming Asymmetry with Chinese-English Bilinguals: A Test of the Sense Model. *Journal of Psycholinguistic Research*, *43*(3), 225–240. https://doi.org/10.1007/s10936-013-9249-3

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Costa, A., & Caramazza, A. (1999). Is lexical selection in bilingual speech production language-specific? Further evidence from Spanish–English and English–Spanish bilinguals. *Bilingualism: Language and Cognition*, *2*(3), 231–244. https://doi.org/10.1017/S1366728999000334

Crinion, J., Turner, R., Grogan, A., Hanakawa, T., Noppeney, U., Devlin, J. T., Aso, T., Urayama, S., Fukuyama, H., Stockton, K., Usui, K., Green, David W., & Price, C. J. (2006). Language Control in the Bilingual Brain. *Science*, *312*(5779), 1537–1540. https://doi.org/10.1126/science.1127761

Dannenbring, G. L., & Briand, K. (1982). Semantic priming and the word repetition effect in a lexical decision task. *Canadian Journal of Psychology / Revue Canadienne de Psychologie*, *36*(3), 435–444. https://doi.org/10.1037/h0080650

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*, *53*(2), 397–428. https://doi.org/10.1080/713755891

Dijikstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175–197.

Dimitropoulou, M., Duñabeitia, J. A., & Carreiras, M. (2011). Two Words, One Meaning: Evidence of Automatic Co-Activation of Translation Equivalents. *Frontiers in Psychology*, *2*, 1–20. https://doi.org/10.3389/fpsyg.2011.00188

Educational Testing Service. (2016). *TOEFL iBT® Test and Score Data Summary 2015* (p. 16). Educational Testing Service. https://www.ets.org/content/dam/ets-org/pdfs/toefl/toefl-ibt-test-score-data-summary-2015.pdf

Educational Testing Service. (2018a). *TOEIC Listening Score Descriptors* (p. 2). Educational Testing Service. https://www.ets.org/pdfs/toeic/toeic-listening-reading-score-descriptors.pdf

Educational Testing Service. (2018b). *TOEIC Speaking Proficiency Level Descriptors* (p. 4). Educational Testing Service. https://www.ets.org/pdfs/toeic/toeic-speaking-writing-score-descriptors.pdf

Educational Testing Service. (2019). *Mapping the TOEIC® Tests on the CEFR* (p. 2). Educational Testing Service. https://www.ets.org/pdfs/toeic/toeic-mapping-cefr-reference.pdf

Educational Testing Service. (2021). *Performance Descriptors for the TOEFL iBT® Test* (p. 4). Educational Testing Service. https://www.ets.org/pdfs/toefl/toefl-ibt-performance-descriptors.pdf

Educational Testing Service. (2023a). *Comparing TOEFL iBT Scores*. TOEFL iBT® Test. https://www.ets.org/toefl/score-users/ibt/compare-scores.html

Educational Testing Service. (2023b). *TOEIC Scores and Score Use*. The TOEIC® Tests. https://www.ets.org/toeic/test-takers/scores.html

Eguchi, M., Suzuki, S., & Suzuki, Y. (2022). Lexical Competence Underlying Second Language Word Association Tasks: Examining the construct validity of response type and response time measures. *Studies in Second Language Acquisition*, *44*(1), 112–142. https://doi.org/10.1017/S0272263121000164

Elston-Güttler, K. E., Gunter, T. C., & Kotz, S. A. (2005). Zooming into L2: Global language context and adjustment affect processing of interlingual homographs in sentences. *Cognitive Brain Research*, *25*(1), 57–70. https://doi.org/10.1016/j.cogbrainres.2005.04.007

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Finkbeiner, M., Forster, K., Nicol, J., & Nakamura, K. (2004). The role of polysemy in masked semantic and translation priming. *Journal of Memory and Language*, *51*(1), 1–22. https://doi.org/10.1016/j.jml.2004.01.004

Forster, K., & Davis, C. (1984). Repetition Priming and Frequency Attenuation in Lexical Access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680–698. https://doi.org/10.1037/0278-7393.10.4.680

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Garcia, G. D. (2021). *Data Visualization and Analysis in Second Language Research* (1st ed.). Routledge.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gollan, T. H., Forster, K. I., & Frost, R. (1997). Translation priming with different scripts: Masked priming with cognates and noncognates in Hebrew-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1122–1139. https://doi.org/10.1037/0278-7393.23.5.1122

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, P., & MacLeod, C. J. (2016). simr: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

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

Hermans, D., Bongaerts, T., de Bot, K., & Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language and Cognition*, *1*(3), 213–229. https://doi.org/10.1017/S1366728998000364

Hoenig, J. M., & Heisey, D. M. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician*, *55*(1), 19–24. https://doi.org/10.1198/000313001300339897

IBM SPSS Statistics for Windows (27.0). (2020). [Windows]. IBM Corporation.

Jiang, N. (1999). Testing processing explanations for the asymmetry in masked cross-language priming. *Bilingualism: Language and Cognition*, *2*(1), 59–75. https://doi.org/10.1017/S1366728999000152

Jiang, N., & Forster, K. I. (2001). Cross-Language Priming Asymmetries in Lexical Decision and Episodic Recognition. *Journal of Memory and Language*, *44*(1), 32–51. https://doi.org/10.1006/jmla.2000.2737

Kerkhofs, R., Dijkstra, T., Chwilla, D. J., & de Bruijn, E. R. A. (2006). Testing a model for bilingual semantic priming with interlingual homographs: RT and N400

effects. *Brain Research*, *1068*(1), 170–183. https://doi.org/10.1016/j.brainres.2005.10.087

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kirby, J., & Sonderegger, M. (2018a). Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics*, *70*, 70–85. https://doi.org/10.1016/j.wocn.2018.05.005

Kirby, J., & Sonderegger, M. (2018b). Model selection and phonological argumentation. In D. Brentari & J. Lee (Eds.), *Shaping Phonology* (pp. 234–252). University of Chicago Press. https://doi.org/10.7208/chicago/9780226562599.001.0001

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kroll, J. F., & Stewart, E. (1994). Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections between Bilingual Memory Representations. *Journal of Memory and Language*, *33*, 149–174. https://doi.org/10.1006/jmla.1994.1008

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

Lee, Y., Jang, E., & Choi, W. (2018). L2-L1 Translation Priming Effects in a Lexical Decision Task: Evidence From Low Proficient Korean-English Bilinguals. *Frontiers in Psychology*, *9*, 267. https://doi.org/10.3389/fpsyg.2018.00267

Li, W. (2000). Dimensions of Bilingualism. In W. Li (Ed.), *The Bilingualism Reader* (pp. 2–21). Routledge.

Lindsay, D. S. (2020). Seven steps toward transparency and replicability in psychological science. *Canadian Psychology / Psychologie Canadienne*, *61*(4), 310–317. https://doi.org/10.1037/cap0000222

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01171

López, L. (2020). Remarks on Separationist Architectures. In *Bilingual Grammar: Toward an Integrated Model* (pp. 12–21). Cambridge University Press.

Luo, X., Cheung, H., Bel, D., Li, L., Chen, L., & Mo, L. (2013). The Roles of Semantic Sense and Form-Meaning Connection in Translation Priming. *The Psychological Record*, *63*(1), 193–208. https://doi.org/10.11133/j.tpr.2013.63.1.015

Lupker, S. J. (1979). The semantic nature of response competition in the picture-word interference task. *Memory & Cognition*, *7*(6), 485–495. https://doi.org/10.3758/BF03198265

Marsden, E., Thompson, S., & Plonsky, L. (2017). Open science in second language acquisition research: The IRIS repository of research materials and data. *SHS Web of Conferences*, *38*, 00013. https://doi.org/10.1051/shsconf/20173800013

Nakayama, M., Ida, K., & Lupker, S. J. (2016). Cross-script L2-L1 noncognate translation priming in lexical decision depends on L2 proficiency: Evidence from Japanese–English bilinguals. *Bilingualism: Language and Cognition*, *19*(5), 1001–1022. https://doi.org/10.1017/S1366728915000462

Nakayama, M., Sears, C. R., Hino, Y., & Lupker, S. J. (2013). Masked translation priming with Japanese–English bilinguals: Interactions between cognate status, target frequency and L2 proficiency. *Journal of Cognitive Psychology*, *25*(8), 949–981. https://doi.org/10.1080/20445911.2013.839560

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas - Part II. *Language and Linguistics Compass*, *10*(11), 591–613. https://doi.org/10.1111/lnc3.12207

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and Confirmatory Analyses in Sentence Processing: A Case Study of Number Interference in German. *Cognitive Science*, *42*(S4), 1075–1100. https://doi.org/10.1111/cogs.12589

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian Revolution in Second Language Research: An Applied Approach: Bayesian Revolution in L2 Research. *Language Learning*, *68*(4), 1032–1075. https://doi.org/10.1111/lang.12310

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., Van 'T Veer, A. E., & Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., … Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour*, *6*(3), 312–318. https://doi.org/10.1038/s41562-021-01269-4

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, *7*(6), 531–536. https://doi.org/10.1177/1745691612463401

Perea, M., Vergara-Martínez, M., & Gomez, P. (2015). Resolving the locus of cAsE aLtErNaTiOn effects in visual word recognition: Evidence from masked priming. *Cognition*, *142*, 39–43. https://doi.org/10.1016/j.cognition.2015.05.007

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, *6*, 1–11. https://doi.org/10.3389/fpsyg.2015.00223

Plonsky, L. (2013). Study Quality in SLA: An Assessment of Designs, Analyses, and Reporting Practices in Quantitative L2 Research. *Studies in Second Language Acquisition*, *35*(4), 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L., & Gass, S. (2011). Quantitative Research Methods, Study Quality, and Outcomes: The Case of Interaction Research. *Language Learning*, *61*(2), 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Potter, M. C., So, K.-F., Eckardt, B. V., & Feldman, L. B. (1984). Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior*, *23*(1), 23–38. https://doi.org/10.1016/S0022-5371(84)90489-4

R Core Team. (2020). *R: A language and environment for statistical computing.* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Roettger, T. B., Winter, B., & Baayen, H. (2019). Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics*, *73*, 1–7. https://doi.org/10.1016/j.wocn.2018.12.001

Rusted, J. (1988). Orthographic effects for Chinese-English bilinguals in a picture-word interference task. *Current Psychology*, *7*(3), 207–220. https://doi.org/10.1007/BF02686669

Scarborough, D. L., Gerard, L., & Cortese, C. (1984). Independence of lexical access in bilingual word recognition. *Journal of Verbal Learning and Verbal Behavior*, *23*(1), 84–99. https://doi.org/10.1016/S0022-5371(84)90519-X

Schoonbaert, S., Duyck, W., Brysbaert, M., & Hartsuiker, R. J. (2009). Semantic and translation priming from a first language to a second and back: Making sense of the findings. *Memory & Cognition*, *37*(5), 569–586. https://doi.org/10.3758/MC.37.5.569

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Sonderegger, M. (2022). *Regression Modelling for Linguistic Data*. https://osf.io/pnumg/

Spivey, M. J., & Marian, V. (1999). Cross Talk Between Native and Second Languages: Partial Activation of an Irrelevant Lexicon. *Psychological Science*, *10*(3), 281–284. https://doi.org/10.1111/1467-9280.00151

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Zandt, T. V., & Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Thierry, G., & Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences*, *104*(30), 12530–12535. https://doi.org/10.1073/pnas.0609927104

Thomson, R. (2015). Fluency. In M. Reed & J. Levis (Eds.), *The Handbook of English Pronunciation* (pp. 209–226). John Wiley & Sons, Inc.

Vasishth, S., & Nicenboim, B. (2016). Statistical Methods for Linguistic Research: Foundational Ideas - Part I: Statistical Methods for Linguistics - Part I. *Language and Linguistics Compass*, *10*(8), 349–369. https://doi.org/10.1111/lnc3.12201

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008

Wang, X. (2013). Language dominance in translation priming: Evidence from balanced and unbalanced Chinese–English bilinguals. *Quarterly Journal of Experimental Psychology*, *66*(4), 727–743. https://doi.org/10.1080/17470218.2012.716072

Wang, X., & Forster, K. (2015). Is translation priming asymmetry due to partial awareness of the prime? *Bilingualism: Language and Cognition*, *18*(4), 657–669. https://doi.org/10.1017/S1366728914000650

Wen, Y., & van Heuven, W. J. B. (2017). Non-cognate translation priming in masked priming lexical decision experiments: A meta-analysis. *Psychonomic Bulletin & Review*, *24*(3), 879–886. https://doi.org/10.3758/s13423-016-1151-1

Wen, Y., & van Heuven, W. J. B. (2018). Limitations of translation activation in masked priming: Behavioural evidence from Chinese-English bilinguals and computational modelling. *Journal of Memory and Language*, *101*, 84–96. https://doi.org/10.1016/j.jml.2018.03.004

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Winter, B. (2011). Pseudoreplication in Phonetic Research. *Proceedings of the International Conference in Phonetic Sciences*, *17*(1), 2137–2140.

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R* (1st ed.). Routledge. https://doi.org/10.4324/9781315165547

Witzel, N. O., & Forster, K. I. (2012). How L2 words are stored: The episodic L2 hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1608–1621. https://doi.org/10.1037/a0028072

Wu, Y. J., Cristino, F., Leek, C., & Thierry, G. (2013). Non-selective lexical access in bilinguals is spontaneous and independent of input monitoring: Evidence from eye tracking. *Cognition*, *129*(2), 418–425. https://doi.org/10.1016/j.cognition.2013.08.005

Wu, Y. J., & Thierry, G. (2017). Brain potentials predict language selection before speech onset in bilinguals. *Brain and Language*, *171*, 23–30. https://doi.org/10.1016/j.bandl.2017.04.002

Xia, V., & Andrews, S. (2015). Masked translation priming asymmetry in Chinese-English bilinguals: Making sense of the Sense Model. *Quarterly Journal of Experimental Psychology*, *68*(2), 294–325. https://doi.org/10.1080/17470218.2014.944195

Zhang, T., van Heuven, W. J. B., & Conklin, K. (2011). Fast Automatic Translation and Morphological Decomposition in Chinese-English Bilinguals. *Psychological Science*, *22*(10), 1237–1242. https://doi.org/10.1177/0956797611421492