Doctoral theses at NTNU, 2020:62

Jonathan D. Kim

# Connections between grammatical gender and occupational gender stereotypes

Doctoral thesis

**NTNU**
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Social and Educational Sciences
Department of Psychology

NTNU
Norwegian University of
Science and Technology

NTNU

Jonathan D. Kim

# Connections between grammatical gender and occupational gender stereotypes

Thesis for the Degree of Philosophiae Doctor

Trondheim, April 2020

Norwegian University of Science and Technology
Faculty of Social and Educational Sciences
Department of Psychology

NTNU
Norwegian University of
Science and Technology

**NTNU**
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Social and Educational Sciences
Department of Psychology



ISBN 978-82-326-4332-5 (printed ver.)
ISBN 978-82-326-4333-2 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2020:62

Printed by NTNU Grafisk senter

**Contents**

**Acknowledgements**

I would firstly like to thank both of my supervisors, Ute Gabriel and Pascal Gygax, for their invaluable guidance, insight, and support throughout this thesis, and for being genuinely lovely and caring people. Their support for my ideas, even when they were a little off the wall, helped me to grow a huge amount as an academic, and I am deeply grateful for that. I will carry forward many fond memories of this time.

I would like to thank Anna Siyanova-Chanturia for accepting the strange request to bring a Kiwi lad back to New Zealand as a 'foreign researcher', for helping me get settled in to the research environment there quickly, and for collaborating in regards to the second paper in this thesis.

I would like to thank everyone involved in the Speech, Cognition, and Language group, especially Dawn Behne, Anton Øttl, and Marzieh Sorati, for the discussions about a myriad of interesting topics, and for the friendliness and support that the group has for those involved. The SCaLA group really made me feel like I belonged here, something that was very nice when I'm so far from my country of birth.

I would like to thank all my other colleagues in the NTNU psychology department, past and present, for being friendly and approachable, and for many late nights and interesting discussions.

I would also like to thank my friends and family for their support and love throughout this long process; those I have back in New Zealand, those I made here in Norway, and those I have met along this journey who live out all around the world. My love and thanks especially to Beena and Jacinta, my super supportive best friends right from the start; wouldn't have been the same without you in my life.

Finally, I would like to thank all my participants, without whom my research would never have been possible.

**List of papers included in this thesis**

**Paper I:** Kim, J. D., Gabriel, U., & Gygax, P., (2019) Testing the effectiveness of the Internet-based instrument PsyToolkit: a comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task. *PLOS One, 14*(9): e0221802.

**Paper II:** Kim, J. D., Gabriel, U., Gygax, P., & Siyanova, A., (Unpub.) Investigating the link between gender stereotypicality and occupational stereotype content through a bottom-up approach.

**Paper III:** Kim, J. D., Gabriel, U., & Gygax, P., (Unpub.). Language structures and gender stereotyped perception: The effect of differences in the level of grammatical gender between fully, semi-, and non-gendered languages.

**Summary**

The intention behind this thesis was to expand knowledge of how language structures affect stereotyped beliefs. More specifically, it examined how linguistic differences between languages affected the interactions between different stereotype categories (i.e., gender and occupation stereotypes). The research conducted in this thesis expands knowledge relating to the importance and content of occupational and gender stereotypes. In order to do so, this thesis examines the interaction between occupational and gender stereotypes in isolation from, and interacting with, grammatical gender. Four aims were addressed throughout this thesis. The first aim was to determine the attributes associated with occupational roles, use these to determine occupational stereotypes, and to examine whether the gender stereotypes related to certain occupations affect perceptions of the importance of each occupational stereotype. The second aim was to examine whether gender ratio can be considered a representative measure of gender stereotypicality. Most central to the thesis, aim three and four were to examine the importance of occupational and gender stereotypes in isolation from (aim three), and interacting with (aim four), grammatical gender.

Four central findings can be extracted from this thesis. Firstly, five occupational stereotypes were determined, with gender stereotypicality found to affect the perceived importance of each stereotype to at least some degree. Along with traditional gender stereotype categories (i.e., feminine and masculine), the categories of unfeminine and unmasculine also emerged. Secondly, gender ratio was indeed found to be representative of gender stereotypicality. Thirdly, there is a strong interaction between occupational stereotypes and gender stereotypes, with masculine stereotyped occupations perceived as more suitable for men than unsuitable for women, feminine stereotyped occupations perceived as more suitable for women than unsuitable for men, and non-stereotyped occupations perceived as equally suitable for both women and men. Fourthly, grammatical gender was found to interact with occupational and gender stereotypes, with semi-gendered language speakers relying significantly less on occupational gender stereotypes to guide social perception than fully and non-gendered language speakers, and with non-gendered language speakers relying less on occupational gender stereotypes to guide social perception than fully gendered language speakers.

Globally, the results of this thesis suggest that occupational and gender stereotypes are rich, interesting constructions that they interact strongly with the level of grammatical gender within a language. Suggestions are also provided in this thesis for applying what was found in wider examinations related to the interplay between linguistic properties and stereotypes.

In exploring the interplay between grammatical gender and occupational gender stereotypes, this thesis successfully reached its aim of expanding knowledge related to the interplay between linguistic factors and stereotype beliefs. The results found in this thesis have wider implications for examinations of differences between stereotypical categories. Most prominently, the finding of unfeminine and unmasculine stereotype categories suggest that the methods used in this thesis (specifically attribute naming and rating tasks) may be useful in identifying and exploring counter-stereotypical categories (i.e., unmasculine and unfeminine categories). Secondly, these results support the use of the experimental methods utilised in this thesis (word association, attribute naming, attribute rating) in exploring the interactions between gender stereotypical categories in a more general sense.

# 1. Introduction

Human perceptions of reality are, in part, created and maintained through narratives; stories that we tell ourselves, and stories that we are told by others. We seek to define and understand our experiences in relation to these narratives, encoding them as key aspects of our social schema and stereotypes, in turn shaping social perception. The narratives an individual learns depends strongly upon cultural aspects, especially upon the specific language(s) that they speak throughout their childhoods. These differences lead to fundamental differences in social schema and social stereotypes between languages (e.g., Little, 1968; Rodríguez-Arauz et al., 2017; Wigboldus & Douglas, 2007). During stereotype activation, it is common for multiple stereotype categories to be activated simultaneously (e.g., Kang et al., 2014), but much research focuses specifically on individual stereotype categories, and in single languages. The intention behind this thesis is to expand knowledge of stereotype beliefs, and how linguistic factors may affect them. Specifically, this thesis aims to examine interactions between stereotypical categories that are activated by a single social stimulus. To improve flow, this will be referred to within this thesis as a 'stereotype interaction'. For the purposes of this thesis, the stereotype interaction under examination is that between occupational stereotypes and gender stereotypes, while the linguistic factor under examination is grammatical gender. Stereotype beliefs arising from the interaction between gender-based and occupational stereotypes are herein referred to as occupational gender stereotypes.

Three studies were conducted over the course of this thesis. These studies were designed to build towards a confident examination of how the interplay between grammatical gender level and occupational gender stereotypes affects social stereotype beliefs. Paper I tests modifications to an existing word association paradigm, to determine whether it is suitable for the purposes of this thesis. It also tests the replicability of a specific internet-based instrument for psycholinguistic research, to determine whether the instrument is suitable for the same. Paper II examines occupational gender stereotype through a ground-up process, to allow for easier and more thorough access to occupational gender stereotype content. Paper III examines the effect of differences in level of grammatical gender between fully, semi-, and non-gendered languages on occupational gender stereotyped information through a complex choice response time task. Consequently, this introduction section starts by introducing the topic of language and social perception (1.1), then turns to a discussion of social stereotypes and stereotype interactions, especially relating to occupational gender stereotypes (1.2), and finishes with a

discussion of grammatical gender, and how it interplays with occupational gender stereotypes (1.3).

### 1.1. Language and Social Perception

Language is the primary medium through which we communicate meaning on both a personal and an interpersonal basis. On the personal level, language acts to frame our thought processes, allowing us to rationally consider complicated matters, and thus guiding our pre-meditated actions. On the interpersonal level, it allows us to share cultural and personal knowledge between individuals. This interpersonal sharing affects many aspects of our cognition, including social cognition (e.g., Chen & Bond, 2010; Chen, 2015; Fausey & Boroditsky, 2010; Güngör et al., 2012; Krauss & Chiu, 1997).

Languages can be understood as structures that are each composed of specific sets of linguistic factors. These factors are phonetic, syntactic, and grammatical 'rules' that exist within a language, and can differ greatly between languages. Grammatical gender is an example of a grammatical linguistic factor. The linguistic structure within a language provides a relatively rigid guideline by which meaning is communicated on both the personal and interpersonal level (Krauss & Chiu, 1997). These guidelines are solid, if mostly invisible, structures that define what, how, and to what level specific ideas and concepts can be communicated within each language (Krauss & Chiu, 1997). These guidelines have an incredibly strong effect on social perception, to the point where even subtle differences in the exact words used in explaining a concept, event, or idea can significantly impact upon meaning interpretation (Fausey & Boroditsky, 2010). Linguistic structures differ between languages, often to a large degree. These structural differences strongly impact upon how meaning is both communicated and interpreted (Boroditsky, 2011; Fausey & Boroditsky, 2010). Slobin (2002) states that language production requires taking a specific perspective based upon these guidelines, and that, even when languages seem so share these guidelines, they may differ in what 'feels natural' for individuals speaking each language, with possible consequences for meaning activation. Slobin (2002) points to the example of newspaper reports relating to an incident between Greenpeace and the French military, with a British newspaper using dynamic terms such as 'troops stormed the Greenpeace flagship' while a French newspaper used less dynamic terms such as 'troops took control of the flagship'. As such, it is reasonable to assume that British and French people who read those papers are likely to have seen the confrontation as more (British) or less (French) violent. Even small changes in word use are known to greatly impact upon communicated meaning. For example, when shown the same video of a car crash, people who are asked 'what

speed did those two cars bump into each other' indicate a much lower speed than those asked, 'what speed did those two cars smash into each other' (Loftus & Palmer, 1974). These differences between languages have such strong effects on our perceptions that personality has been found to shift depending on the language a multilingual individual is speaking at a given time (e.g., Chen & Bond, 2010). Even when two languages appear to share cultural beliefs on the surface level, there may be differences in how the beliefs are understood between languages. These differences in contextual structure even when surface-level structures are identical also can lead to misunderstandings in communicated meaning.

#### 1.1.1. The human processing approach, and social schema

The process by which language affects human social perception can be understood through the human processing approach. This approach states that humans can be characterised as information processing systems (e.g., Proctor & Vu, 2012). Our biological processes require us to selectively interpret sensory input, to encode relevant information into our memories, to retrieve this information from memory when appropriate, and to make decisions and undertake actions based upon both memory and active sensory information (Proctor & Vu, 2012). We receive more than 10 million bits per second of sensory information but, due to inherent biological limits, we only actively perceive, interpret, and store approximately 50 bits per second of this into our conscious memory (Riener, 2017). The remaining information is not entirely lost, as our subconscious utilises some of this information to inform us on the unconscious level (Riener, 2017). When an individual attempting to interpret meaning from a conversation or a text, the human information processing approach holds that activation of schema occurs within the context of linguistic processing (An, 2013; Carreli & Eisterhold, 1983). Linguistic processing is directly affected by the linguistic structures inherent within the language in which the information was spoken or written (e.g., Paap, 1975; Phillips, 2018), acting to limit the ways in which linguistic information can be encoded. This encoding forms and alters our cognitive schema (Axelrod. 1973; Tse et al., 2011).

Schema are complex, complete, yet fundamentally abstract cognitive knowledge structures that allow for the mental representation of concepts stored in memory (Anderson & Pearson, 1984; Medin & Ross, 1992; Rumelhart, 1980). These structures allow us to mentally organize and interpret information based upon both current external stimuli and memory, guiding our interpretations of reality, and impacting upon our thoughts and actions (e.g., Medin & Ross, 1992), and include knowledge of how all appropriate characteristics are associated with both the schema and with each other. Social schema specifically hold conceptual

knowledge or information about social frameworks and concepts, as well as how they relate to each other to form a coherent complete image of the topic of the schema (An, 2013). This allows us to interpret aspects of our social environments correctly. Activation of both social and non-social schema happens very quickly. This quickness is important, as it allows for the activation of schema prior to cognitive processes requiring information contained within the schema, such as attribute selection during meaning activation in language comprehension. Development of social schema occurs through the internalization of successfully communicated cultural beliefs that begin during infancy. This development starts with practice play, where initial knowledge about objects that they are surrounded with forms sensorimotor schema, which develops into symbolic play, where sensorimotor schema are chained together to form beliefs about higher-level actions and interactions, allowing for the identification of chains of actions that need to occur to achieve certain aims, which finally develop into play activities with socially defined rules, where these chains of schema are concretely tied into culturally held beliefs (Kumar et al., 2018). Even before the 'play with rules' phase, cultural factors impact upon the kinds of objects with which infants are surrounded, and thus upon their development of, and formed connections between, sensorimotor schema.

Linguistic processing is a difficult cognitive task, as the slow and linear nature of information gained through listening or reading is much slower than schematic activation in other areas of information processing. As such, a feedback loop occurs in which existing social schema are activated to inform the processing of linguistic information, which again informs the social schema. This reliance on social schema to guide linguistic processing is so strong that it occurs even when an individual is faced with a stimulus that they have never encountered before. When this occurs, beliefs are formed based on previously encountered stimuli that are perceived to be similar to the novel stimuli. This attribution is so strong that new information going against these rapidly formed beliefs produces a negative reaction of the same degree as if it had contradicted long-held beliefs (Jaeger & Weatherholtz, 2016; Rácz, 2012; Rácz, 2013). This intensity in reaction means that even new beliefs are relatively difficult to change once they have been formed. Even when everyone participating in a conversation are native speakers of the same language, it is possible for a sentence to be interpreted as meaning something different than what the writer/speaker of the sentence intended. When communicated meaning is interpreted, regardless of whether it was communicated correctly, certain mental processes occur that can have lasting impacts upon social schema and stereotypes. These processes can be understood through meaning activation theory.

### 1.1.2. Meaning activation theory, and the activation selection model

Meaning activation theory is based on the idea that cognition and the encoding of linguistic information are fundamentally informed by the salience of every word within a sentence. Globally, this theory states that the processing of individual words within a sentence activate the semantic, morphological, and phonological features of those words (e.g., Lévy, Gygax, & Gabriel, 2014). Semantic features are the different elements that provide meaning for a specific word. For example, the word 'chocolate' might activate the semantic features of 'food', 'sweet', and 'cocoa'. Morphological features are those which add contextual information relating to how words relate to each other, aiding with the interpretation of each word within a sentence, both within itself (derivational morphemes) and in relation to the other words in the sentence (inflectional morphemes). Phonological features are the specific distinctive sounds (referred to as phonemes; e.g., 'ga', 'ba', 'da') that exist within a language, and, when used in the context of meaning activation theory, relate to the sound patterns of phonemes associated with each individual word in a sentence. The activation of these features allows for each individual word to be interpreted within the specific sentence that was used, as well as within the wider societal context within which the writer/speaker of the sentence exists (e.g., Lévy, Gygax, & Gabriel, 2014). This allows our brains to provide us with as much salient information as possible to inform memory and perception. This theory is built upon through the activation selection model, which defines a mechanism by which meaning becomes associated with words.

The activation selection model states that weighted attributes allow for the determination of meaning for words used within a given context. This occurs through a two-stage process of 1) transient activation, and 2) long-term changes in meaning representations (Gorfein & Bubka, 1989; Gorfein, 2001; Gorfein, Brown, & DeBiasi, 2008). When a word is used in text or speech, the reader or listener automatically selects a series of attributes to help with the interpretation of the word in both societal and sentence-level contexts. For example, the word 'water' is likely to activate the attributes 'cold' and 'wet'. This process is quick and automatic, leading to a reliance on brief activations of social schema to provide societal context for attribute selection. Every time an attribute is activated it gains weight in relation to the word it was activated by, and the heavier an attribute is the more likely it is to be selected even when doing so is incorrect (e.g., the activation of 'cold' when you see boiling water). The number of attributes selected depends on how difficult it is to contextualise the word, with heavier attributes selected before lighter attributes. Every activation of a word is dynamic, meaning that often very different attributes, and different numbers of attributes, are activated by the same

word. Using water as an example again, it could be activated due to personal experiences (e.g., drinking it, washing with it, being rained on), through external elements such as geography (e.g., a stream or the sea), or through a myriad of other possible reasons. For words that have not previously been encountered, or are used in an ambiguous manner, attributes activated by other words in the sentence, or which would be activated by words perceived as being similar, are utilised within the context of the sentence to help to determine the new/ambiguous word's meaning (Gorfein, Brown, & DeBiasi, 2008). The larger the perceived overlap between the new/ambiguous word and the related word(s), the more the attributes activated by these related words are relied upon to determine the meaning of the new/ambiguous word.

The process of attribute selection is referred to as transient activation as, while the attributes guide perception within a sentence and gain weight in relation to a specific word, their activation decays rapidly. Attributes activated within one sentence have little to no effect on meaning activation within subsequent sentences. As attributes gain weight in relation to a word, the individual's meaning representations shift. Attributes with heavier weightings become more likely to be activated even when this is inappropriate given the social and/or sentence-level context in which the word was used. Simultaneously, attributes with lighter weightings become less likely to be activated even when it would be appropriate (Gorfein, Brown, & DeBiasi, 2008). The increased likelihood of selecting 'heavy' attributes and not selecting 'light' attributes has a dampening effect on the dynamic nature of attribute selection, and can lead to certain attributes being nearly always selected (if sufficiently heavy) or almost never selected (if sufficiently light), regardless of context. This shift from a more dynamic interpretation to a more static interpretation of specific words is an explicit shift in meaning representation. For example, if you lived your entire life by a crystal-clear spring from which you could draw safe drinking water, the attribute 'potable' (i.e., safe to drink) might become very heavily weighted. If you were then lost at sea and were thirsty, you may incorrectly activate the 'potable' attribute in relation to the sea water and try to drink it, as activating 'potable' suppressed the correct attribute 'non-potable'. It follows that this change in understanding also affects any existing schema that utilised the words whose common meanings have been altered in this manner.

### 1.1.3. Summary of this section

In summary, language is the primary medium through which we communicate meaning. It is composed of linguistic factor structures that provide rigid guidelines for social perception. The human processing approach states that the encoding and retrieving relevant social

information occurs automatically through linguistic processing, which acts to create and continuously inform social schema. These schemas are complex and complete beliefs relating to specific social structures, allowing for the quick activation of information relating to the content of the schema. After social schema have been created a feedback loop is created with linguistic processing where they both inform each other to a very high degree. The activation of social schema during human processing can be understood through the activation selection model of meaning activation theory, which states that meaning is interpreted at the word level, with each word in a sentence activating associated weighted attributes to inform understanding both within the sentence and within the culture to which the interpreter belongs. Increased frequency of activation of a specific attribute with a specific word leads the attribute to be activated even when it is not a correct activation, leading to both short-term and long-term changes in social schema content.

The next section focuses on social stereotypes, a form of smaller social schema that are more precise, yet less flexible, than most schema (e.g., Seta, Seta, & McElroy, 2003; Stangor & Schaller, 2000).

### 1.2. Social stereotypes and stereotype interactions, especially relating to occupational gender stereotypes.

Social stereotypes focus specifically on the characteristics and/or qualities perceived as common among individuals within a specific social group (e.g., Locksley, Hepburn, & Ortiz, 1982), allowing us to form quick opinions about individuals based upon the groups to which they are perceived as belonging to. Social stereotypes are composed of both descriptive and prescriptive beliefs (e.g., Fiske & Stevens, 1993; Koenig, 2018). In this context, descriptive beliefs focus on the attributes, roles, and behaviours that are perceived as *characterising* members of a specific group, such as swimmers all eat raw eggs every day, while prescriptive beliefs focus on behaviours to which members of a specific group are *expected to conform*, such as swimmers *should* eat raw eggs every day (Burgess & Borgida, 1999).

Social stereotypes, as with other social schema, are created through communication with those around us (Aboud & Doyle, 1996), and through engagement with media (Brown, 1995). Social processes act to teach these stereotypes cross-generationally. This cross-generational learning makes them difficult to alter, with fundamental aspects of them likely to remain as societal beliefs for very long periods. However, they are somewhat flexible in that they can be

altered on the individual level through the same processes (e.g., activation selection) that impact upon larger social schema. Stereotype beliefs are generally useful, as they are a low cognitive cost method of accessing information about an individual based on the group(s) they belong to (Lee, Jussim, & McCauley, 1995). These beliefs are not infallible, being subject to prejudices and overgeneralisation effects wherein differences between members of a group are minimised while similarities are maximised. As such, these beliefs are only informative to the extent that they are accurate at the group level. This accuracy is measured through the *proportion* of a given population about which a stereotype is true (e.g., Swim, 1994). The higher the proportion of a population that the stereotype is about, the more accurate it is. For example, for a hypothetical stereotype of swimmers' eye colour that states 'all professional swimmers have green eyes; all amateur swimmers have blue eyes', research examining the accuracy of these stereotypes could examine the eye colours of swimmers in each of these categories. If all professional swimmers were found to have green eyes, then the stereotype is rated as completely accurate. If they were all found to have eyes of any colour except green, then the stereotype is rated as completely inaccurate. When individuals are perceived as acting in a counter-stereotypical manner (e.g., being a professional swimmer with blue eyes), they risk being seen as socially undesirable. Depending on the stereotype category that was broken, this can lead to social punishment (e.g., verbal harassment; Harrison, Welch, & Adler, 2012) or economic punishment (e.g., being fined or fired; Shaw & Hoeber, 2003).

### 1.2.1. Stereotype interactions

Stereotypes do not exist in isolation, and the group(s) an individual belongs to is likely to activate multiple stereotypes in an observer's mind simultaneously. When this occurs, these stereotypes interact, leading to the possible creation of new stereotype beliefs. For example, if the hypothetical stereotype of eye and hair colour (everyone with green eyes has red hair; everyone with blue eyes has black hair) was activated simultaneously with the stereotype about swimmers eye colour in a manner that reinforced both stereotypes, it is likely that a stereotype relating to swimmers hair colour would develop over time (i.e., professional swimmers have red hair; amateur swimmers have black hair). Once these interaction-based stereotypes are formed, those acting in a manner counter to the stereotype are again viewed as socially undesirable.

When multiple stereotype categories are activated simultaneously, those who are perceived as breaking multiple stereotypes are seen as more socially undesirable than those who break only one. In other words, a professional swimmer with black hair and blue eyes would be

seen as less desirable than a professional swimmer with black hair and green eyes, who in turn is seen as less desirable than a professional swimmer who, in keeping with the stereotype, has both red hair and green eyes.

It is also possible that beliefs relating to one stereotype inhibit beliefs related to a second stereotype. For example, Rydell, McConnell, and Beilock (2009) found that the stereotype 'women are worse at maths than men' activated in isolation led to reduced working memory in women attempting to answer maths questions, but activation concomitantly with a positive self-reliant stereotype (e.g., 'everyone who works in your specific occupation is great at math') inhibited gender identity, largely eliminating the working memory deficits. This richness of content, where beliefs from stereotype interactions can both expand upon and inhibit more general stereotype beliefs, means that there is a wealth of information to explore. The focus of this thesis is specifically on the interaction between gender stereotypes and occupational stereotypes.

### 1.2.2. Gender stereotypes

Gender stereotypes can be classified as being the cultural beliefs around what is expected of an individual based solely upon their gender. Gender congruent behaviour (i.e., individuals exhibiting the attributes associated with their gender) is seen as highly socially desirable, while gender incongruent behaviour (i.e., individual exhibiting the attributes specifically not associated with their gender) is seen as highly socially undesirable (Prentice & Carranza, 2002), leading to social punishment. For example, a man displaying feminine traits is likely to face verbal abuse such as being called a "sissy" (Harrison, Welch, & Adler, 2012). Under Social Role Theory, gender stereotypes are held to arise due to the roles that women and men are observed as fulfilling (Eagly, 1987). The attributes displayed by individuals in a role lead to the creation of role-based stereotypes over time, including about distributions of women and men within the roles. This perceived gender ratio then acts to inform the creation of gender stereotypes, which are then taught through social processes to younger generations. Under this theory, women are perceived as more likely to hold caretaking roles, which have become associated with communal traits, while men are more likely to hold provider roles, which have become associated with agentic traits (Eagly & Wood, 2016). This association has led to the belief that agentic traits are masculine, while communal traits are feminine (e.g., Abele, 2003; Bakan, 1966; Deaux & LaFrance, 1998). This approach is inherently predicated on the assumption that judgements of gender ratio are most commonly based on gender stereotyped attributes perceived as inherently connected to the role, and that therefore perceived gender

ratio is a true measurement of gender stereotypicality. Specifically, agentic traits (and other masculine stereotyped attributes) lead to occupations being perceived as having a high male to female ratio, while communal traits (and other feminine stereotyped attributes) lead to occupations being perceived as having a high female to male ratio (e.g., Adachi, 2013; Misersky et al., 2014). This assumption underlies a lot of research into gender stereotypicality (e.g., Adachi, 2013; Carreiras, Garnham, Oakhill, & Cain, 1996; Garnham, Doehren, & Gygax, 2015; Kennison & Trofe, 2003; Kulik, 1999; Misersky et al., 2014), but has only been examined on a preliminary level. Research into this topic has offered some preliminary support for this assumption. For example, Adachi (2013) compared occupational gender ratios determined by the Japanese government to Japanese participants' independent ratings of the level to which occupations are feminine or masculine stereotyped, and found that occupations with a high male to female ratio tend to be perceived as highly masculine, while occupations with a high female to male ratio tend to be perceived as highly feminine.

Gender ratio has also been conceptualised as a measurement of either conceptual gender (i.e. the level to which an objects'/roles' perceived gender is defined by lexical semantics/stereotypical knowledge without reference to linguistic or natural gender categories; e.g., Irmen, 2007; Sera, Berge, & Pintado, 1994) or gender typicality (i.e., the level of in-group belonging individuals' feel towards their own gender group (Egan & Perry, 2001). Gygax et al. (2016) point out that these conceptualisations are not inherently contradictory, with researchers discussing research conducted in one conceptualization as if it had been conducted in a second. For example, Irmen (2007) presents research on gender stereotypicality as having been on conceptual gender (e.g., Carreiras, Garnham, Oakhill, & Cain, 1996), while Wolfram & Mohr (2010) present research on gender stereotypicality as having been on gender typicality (e.g., Glick, Wilk, & Perreault, 1995).

### 1.2.3. Occupational stereotypes

Occupational Stereotypes can be classified as being the cultural beliefs around what is expected of those who hold specific occupational roles. Occupationally congruent behaviour is seen as essential for an individual to truly 'belong' in an occupation, and occupationally incongruent behaviour seen as evidence that the individual is unprofessional, not truly belonging in the occupation. Judgements of an individual as unprofessional leads to economic punishments, such as being denied promotion, forced into unpaid overtime, and being more likely to be fired or made redundant (Shaw & Hoeber, 2003). Under Social Role Theory (Eagly, 1987) and the Theory of Vocational Choice (Holland, 1997), occupational stereotypes are

intrinsically tied to occupational 'themes', based on the attributes that those who both currently do and previously have held the role within a culture. Each occupation is held to be based around a single 'theme', attracting individuals whose personalities and beliefs match the 'theme'. The close match between the attributes stereotyped as important and the attributes displayed by those entering the occupational role acts to reinforce others' stereotypes relating to that occupation.

### 1.2.4 Occupational gender stereotypes

Occupational gender stereotypes can be defined as cultural beliefs around what is expected of an individual based upon both their gender and upon the specific occupational role that they hold. Under social role theory, feminine stereotyped occupations are perceived as requiring a high level of feminine stereotyped attributes and attract a high ratio of women compared to men, while masculine stereotyped occupations are perceived as requiring a high level of masculine stereotyped attributes and attract a low ratio of women compared to men. Behaviour that is fully congruent (i.e., in keeping with both occupational and gender stereotypes) is seen as highly socially desirable, while partially incongruent behaviour (i.e., in keeping with either occupational or gender stereotypes but not both) is seen as not socially desirable, and fully incongruent behaviour (i.e., not in keeping with either occupational or gender stereotypes) is seen as completely socially undesirable. Of these categories, only fully congruent behaviour is rewarded, while incongruent behaviour is likely to lead to both social and economic punishment (Shaw & Hoeber, 2003). These punishments are unique from those arising from purely gender or occupational stereotypes alone, as they act to inform and encourage each other. When taking leave to look after sick family members (a stereotypically feminine task), men are judged to be fundamentally less compliant and altruistic towards their co-workers (thus less professional) than if they had remained at work, whereas women are perceived as equally professional regardless (Wayne & Cordeiro, 2003). Further, in a management setting where 'masculine' traits (e.g., agency) are seen as essential, *anyone* displaying 'feminine' traits (e.g., communality) are likely to be labelled as chaotic and/or irrational (Putnam & Mumby, 1997). The labels of unprofessional, chaotic, and irrational serve an official basis for economic punishments such as denial of promotion or being fired. This interaction helps to obfuscate that the economic punishment is due to bias on the behalf of those in positions of authority, as they provide reasons by which an individual is perceived as unprofessional that does not outwardly appear to be due to their gender or to their display of agentic/communal attributes. Descriptive occupational gender stereotype beliefs also affect

perceptions of competency; a set of credentials perceived as adequate proof of competency for men is likely to be perceived as inadequate proof of competency for women (Steinpreis et al., 1999).

### 1.2.5. Stereotype strength, importance, and contents

If social role theory approach is accepted, three key approaches to understanding stereotypes can be seen. Firstly, shared knowledge of what the stereotypes are (e.g., Adachi, 2013; Misersky et al., 2014). Secondly, the level to which these beliefs actually guide social perception (e.g., Siyanova-Chanturia et al., 2015). Thirdly, the attributes that these stereotypes are composed of (e.g., Glick, Wilk, & Perreault, 1995). For the purposes of this thesis, the first is defined as stereotype *strength*, the second as stereotype *importance*, and the third as stereotype *contents*.

Stereotype strength can be seen to refer to shared knowledge not only within a specific cultural context, but also between cultural contexts. For example, Misersky et al. (2014) found that the perceived ratio of men and women within specific occupations, utilised as a measure of gender stereotypicality, remains relatively constant between languages. This suggests that, at least for occupational gender stereotypes, stereotype strength is stable between different cultural contexts. Stereotype strength can be conceptualised as the more rigid aspect of social stereotypes.

Stereotype importance refers to the level to which the social perceptions of individuals within specific cultural contexts are affected by stereotyped beliefs. This can be examined through comparing differences in social perception between congruent and incongruent examples (e.g., perception of the ability for nurses to be women [congruent] or men [incongruent]). This is especially useful for examining *comparative* stereotype importance between two groups for which stereotype *strength* is relatively constant. For example, Siyanova-Chanturia et al. (2015) utilised an audial word association paradigm to examine differences in the importance of occupational gender stereotype information for guiding native Italian speakers' social perceptions based on life stage (childhood vs. adulthood). Participants were presented with audio that stated a gender stereotyped occupation followed by a gendered familial role, and were tasked with stating whether both words described the same individual. Stereotype importance was measured across four categories; masculine congruent (masculine occupation, male familial role), feminine gender (feminine occupation, female familial role), masculine incongruent (masculine occupation, female familial role), and feminine incongruent (feminine occupation, male familial role).Their results indicated that stereotype importance was

not significantly different between both age groups, with participants responding more positively and more quickly to the congruent compared to the incongruent pairings. Stereotype importance can be conceptualised as the more flexible aspect of social stereotypes.

Stereotype contents refers to the *specific* attributes that compose *each* of our social stereotypes. Research in this field typically follows a two-study approach, where attributes related to a topic of interest are first established (study 1), are possibly transformed into 'stereotype images' (e.g., Glick et al., 1995), and then a rating task is conducted on the attributes (e.g., Koivula, 2001), the 'images' (e.g., Glick et al., 1995), or upon the role titles (e.g., Imhoff, Koch, & Flade, 2018). These tasks can be used to examine both singular stereotype categories (e.g., Imhoff, Koch, & Flade, 2018 occupational stereotypes) and stereotype interactions (e.g., Glick et al., 1995, occupational gender stereotypes; Koivula, 2001, sports-based gender stereotypes). For example, Glick et al. (1995) utilised an attribute naming task (experiment 1) and an occupational rating task (experiment 2) to explore occupational gender stereotype 'images'. Participants in experiment 1 were presented with a selection of occupations in a random order and, for each occupation, were asked to list attributes of 'the typical person' in the occupation as quickly as possible. The resulting attributes were then classified by two external judges into mutually exclusive general categories suggested by the data (e.g., female vs. male). The judges had an inter-rater reliability of 91%, and the items that they disagreed with were allocated as half-points to both categories selected. The results of this experiment indicated that, in line with social role theory, sex and gender-stereotyped personality traits were commonly listed components for occupational stereotypes. Participants in experiment two were presented with occupational titles and were asked to rate the level to which they perceived the general categories identified in experiment 1 as being important for each occupation. The results of this experiment indicated that prestige and gender stereotypicality were the most important factors for determining the exact nature of stereotypes relating to individual occupations. Research into stereotype contents relating to occupational gender stereotypes has not included measures by which the gender stereotypicality of the occupations selected was balanced, relying upon large numbers of occupations involved in the experiment to ensure that the attributes and stereotype components (i.e., thematically grouped attributes that are used to define stereotype beliefs; e.g., agentic or communal traits, useable to define beliefs related to feminine and masculine stereotypes) are universally applicable. While it is possible that this is the case, the number of masculine compared to feminine stereotyped occupations is much higher (Misersky, 2014). It may be that, without balancing by gender stereotypicality, attributes relating to stereotypically masculine roles are more likely to be named during naming tasks, and more

likely to be higher rated during rating tasks, than attributes relating to stereotypically feminine roles. If this is the case, then balancing by gender stereotypicality may lead to differences in both the specific attributes associated with each stereotype component, or even to differences in what stereotype components are even identified. Importantly, the term stereotype contents is specifically separate from the Stereotype Content Model (Fiske et al., 2002), which states that intergroup stereotypes are based on two primary dimensions; warmth (perceived predisposition of individuals within a social group to attack outgroup members) and competence (perceived likelihood of individuals within a social group successfully attacking outgroup members). This separation from the Stereotype Content Model is driven by a desire to allow stereotypes to be built in a truly ground up manner.

### 1.2.6. Summary of this section

In summary, social stereotypes are generalized beliefs that allow us to form quick opinions about others. They are composed of both descriptive and prescriptive beliefs, with descriptive beliefs focusing on characterizing group members, and prescriptive beliefs focusing on expectations of conformity for group members. Stereotypes are generally useful, as they are low cognitive cost methods of accessing social information, but need to be explored for accuracy if they are to be relied upon. When an individual is perceived to be displaying counter-stereotypical attributes, they are subject to social shaming and to social and/or economic punishment. When multiple social stereotypes are activated concurrently, stereotype beliefs relevant to each stereotyped group separately, and to interactions between those groups, are activated simultaneously. The interactions between stereotypes results in stereotype beliefs both inhibiting each other and leading to the formation of new stereotype beliefs based on shared aspects. In relation to occupational gender stereotypes, under social role theory, feminine stereotyped occupations are perceived as requiring a high level of feminine stereotyped attributes (e.g., communal traits) and attract a high ratio of women compared to men, while masculine stereotyped occupations are perceived as requiring a high level of masculine stereotyped attributes (e.g., agentic traits) and attract a low ratio of women compared to men. Further, if we accept gender ratio as a measurement of gender stereotypicality, three approaches to understanding stereotypes open to us; stereotype *strength*, stereotype *importance*, and stereotype *contents*. This leads to two clear research focuses; examining the specific attributes associated with occupational gender stereotypes (e.g., Glick, Wilk, & Perreault, 1995), and examining the importance of stereotypes in guiding social perception relating to occupational gender stereotypes (e.g., Siyanova-Chanturia et al., 2015). Two more minor research focuses

arises relating to the first focus; whether gender ratio is indeed a measurement of gender stereotypicality, and whether balancing for gender stereotypicality would affect the composition of attribute groupings.

Although the strength of occupational gender stereotypes seems to remain relatively constant between languages (e.g., Misersky et al., 2014), the importance of gender stereotypes in guiding social perception does differ. Preliminary research has suggested that grammatical gender is a significant manner in which this occurs (Gabriel & Gygax, 2016).

### 1.3. Grammatical gender and its interplay with occupational gender stereotypes.

The term grammatical gender refers to a noun-class system where nouns are *grammatically* assigned specific 'genders'. Gender assignation is not necessarily on a biological basis, with the term being used to refer to divisions of language into any grammatically distinct class, such as animacy, innate humanness, and, indeed, the physical gender of the person to whom the noun refers. For the purposes of this thesis, grammatical gender is defined as explicitly referring to nominal classes associated with physical gender. The assignation of grammatical gender within a language differs depending on the inherent nature of the noun being referred to. For inanimate objects, the assignation of grammatical gender is seemingly arbitrary, and often differs between languages, whereas for nouns referring to humans, grammatical gender is normally based on physical gender (Gabriel & Gygax, 2016). This assignation of grammatical gender to nouns referring to humans has been found to be relatively consistent between languages (Misersky et al., 2014).

#### 1.3.1. Levels of grammatical gender within languages

The extent to which the grammatical gender inherent within a language makes, or does not make, a distinction between females and males is known as the level of genderisation within the language. Gygax et al. (2019) state that there are five distinct categories under which languages can be assigned on a continuum from fully gendered to fully ungendered. These are fully gendered (e.g., French, German; also referred to as 'grammatically gendered' or 'gendered'), combination grammatical/natural gendered (e.g., Norwegian), natural gendered (e.g., English), genderless with traces of grammatical gender (e.g., Basque), and genderless (e.g., Finnish). Languages are considered fully gendered if all nouns and pronouns, both animate and inanimate, are assigned a gender and given associated grammatical gender markers (normally masculine or feminine, but sometimes epicene; e.g., French, German, Spanish). In a fully gendered language, grammatical gender is a key aspect of referential gender, and normally

shares the same gender form as other aspects. For fully gendered languages, the noun used to refer generically to someone in a given role (i.e., to refer to an individual where one is not aware of their gender) is essentially the same as the noun used to refer specifically to men (or very rarely women) in that role. This is referred to as grammatical gender asymmetry (e.g., Beatty-Martínez & Dussias, 2019). Under the activation selection model, gender specific attributes will become increasingly associated over time with the generic form of the role noun. This eventually leads to the role being perceived as inherently masculine (or, for those rare roles, feminine) even when intended generically. Languages are considered to be combination grammatical and naturally gendered when grammatical gender distinctions are provided for inanimate as well as personal nouns, but where human-related nouns do not distinguish between feminine and masculine forms. This allows them to refer to female and male referents equally without linguistic differentiation, meaning that they are much closer to natural gender languages than to fully gendered languages. Languages are considered to be natural gendered if personal and object-based nouns are primarily ungendered, and where referential gender is primarily based upon gendered personal pronouns with associated gender markings (e.g., 'he' or 'she'). Languages are considered to be genderless with traces of grammatical gender if most personal nouns *and* pronouns are used in the exact same linguistic form to refer to any individual regardless of gender, but where grammatical gendered forms rarely appear in the form of gender suffixes, gendered adjectives, or gendered verbal forms. Languages are considered to be genderless if most personal nouns *and* pronouns are used in the exact same linguistic form to refer to any individual regardless of gender, and where grammatical gendered forms do not appear on *modern* words. Lexical gender markings are still used (e.g., the Turkish *erkek* [male]), and *historic* grammatical gender markers may appear specifically on human-related nouns (e.g., the Finnish suffix '-tar' can be used to mark old words as feminine; e.g., näyttelijä for actor, näyttelijätär for actress). Along with these five specific categories, two more general categories can be defined. These are semi-gendered languages, composed of natural gendered and combination grammatical and naturally gendered languages (Braun, Oakhill, and Garnham, 2011; Gygax et al., 2019), and non-gendered, composed of genderless and genderless with traces of grammatical gender (Braun, Oakhill, and Garnham, 2011). Semi-gendered languages are grouped based on the fact that gender is distinguished through pronouns, with (most) nouns having no grammatical markings of gender. Grammatical gender markers can be observed to still be utilised, but in a manner decoupled from referential gender. The exact nature of this decoupling differs by language; for example, in Norwegian most role nouns, even those referring to women or to female-stereotyped occupations, only exist in the masculine form.

Non-gendered languages are grouped based on the fact that personal nouns and pronouns are used in the exact same linguistic form regardless of gender, and where other grammatical gender markers are rare or non-existent. In this thesis, we will be utilising the categories fully gendered, semi-gendered, and non-gendered in exploring the interplay between grammatical gender and social perception relating to occupational gender stereotypes.

### 1.3.2. Knowledge on the interplay between grammatical gender and occupational gender stereotypes

The effects of both grammatical gender and occupational gender stereotypes for guiding social perception have both been examined separately in detail. However, research into the interplay between grammatical gender and occupational gender stereotypes has not been explored very deeply. Research into this interplay has found evidence that grammatical gender impacts upon the importance of occupational gender stereotypes within a language (e.g., Gygax et. al., 2008; Gygax et al., 2012; Lévy, Gygax, & Gabriel, 2014), but this evidence is still preliminary (Gabriel & Gygax, 2016). Research has been conducted on the importance of occupational gender stereotypes for guiding social perception between fully and semi-gendered languages (e.g., Gabriel et al., 2008; Gygax et al., 2008), and has examined this interplay for non-gendered languages in isolation (e.g., Pyykkönen, Hyönä, & van Gompel, 2010), but, central to this thesis, no research has examined this interplay between non-gendered languages and either fully or semi-gendered languages, let alone between fully, semi-, and non-gendered languages. This is important as, without this knowledge, the accurate communication of meaning across countries with varying levels of grammatical gender is far more difficult, especially for official agreements between multiple countries with varying levels of grammatical gender within their native languages.

Research that has examined this interplay between fully and semi-gendered languages suggests that social beliefs about the *strength* of stereotype beliefs remain relatively constant across languages regardless of level of grammatical gender (e.g., the role 'doctor' being perceived as strongly masculine regardless of language; Misersky et al., 2014), but that the importance of these stereotype beliefs differs between languages with differing levels of grammatical gender, with social perception more strongly controlled by the grammatically gendered form in fully gendered languages but by gender stereotypicality in semi-gendered languages. For example, Gabriel et al. (2008), examining stereotype strength, conducted an experiment into this interplay between English (semi-gendered), French (fully gendered), and German (fully gendered). Over the course of two studies, they presented native speakers of each

language with a broad range of social roles (primarily occupational) and asked participants to rate each role presented on a 10-point Likert scale based upon their perceptions of gender ratios within the roles. This scale ran from 1 (100% male, 0% female) to 10 (100% female, 0% male). In the first study, all roles were presented in the specifically masculine or feminine form by the number one, and in the other specific form by the number ten. For English, since not all roles had masculine or feminine specific forms, the words 'male' and 'female' were added to the generic form instead. In the second study, following the concept of gender asymmetry, the grammatically masculine form was used in its generic form, and was placed by both numbers one and ten. The results indicated that when the feminine-specific form of a role noun was activated first, participants indicated higher numbers of women within those roles than when the masculine-specific was activated first, and compared to when the masculine-generic was the only form activated. No difference was seen between first activation of the masculine-specific and activation of the masculine-generic. This suggests that even when activated in the generic form, gender asymmetric role nouns activate gender-specific attributes. The results also indicated that stereotypicality ratings were highly reliable between languages, suggesting that differences in grammatical gender between fully and semi-gendered languages does not affect the strength of gender stereotypes. Further, Gygax et al. (2008), examining stereotype importance, conducted an experiment into this interplay with native English, French, and German speakers. Participants were presented with two sentences and were tasked with deciding, as fast as possible, whether the second sentence was a sensible continuation of the first. The results indicated that speakers' responses were guided by occupational gender stereotypes for English, but relied upon grammatical gender markers for French and German. This reliance was to the degree that a clear masculine bias (i.e., participants perceiving each occupation as more suited for men compared to women) was observed across both French and German results regardless of the gender stereotype associated with an occupation. These results suggest that, when gender is attributed to an unknown individual on the basis of an occupational role noun in the generic form (i.e., in a form that has no other associated lexical or syntactic gender markers), this is primarily done based upon gender stereotypes associated with the role for semi-gendered language speakers, and based upon the grammatical form of the role for fully gendered language speakers.

Research into non-gendered languages has found that, similarly to semi-gendered languages, there is a strong reliance on gender stereotypes to guide social perception. For example, Pyykkönen et al. (2010) examined the level to which native Finnish speakers' occupational gender stereotype perception relies upon the activation of general world

knowledge. Participants heard a series of short three sentence stories while looking directly at a screen, upon which images related to the story were shown. Each story followed the same pattern; an introductory sentence where the narrator introduces the scenario (e.g., on the screen you see Robert and Sarah, who live in Trondheim) is followed by a second sentence that gives information about a previously-held conversation about a gender-stereotyped role that specifically ends in a reference to an ungendered object displayed in the scene (e.g., yesterday while talking on the phone, Robert asked Sarah about dangers that nurses face in relation to used needles), with the final sentence using the Finnish non-gender-specific pronoun *hän* (similar to the singular 'they' in English) to ambiguously refer to either person on the screen (e.g., once they became a Nurse, they found that the chances of getting hurt by used needles is small). Analysis was based on gaze location, duration, and movement. Their results indicated that listeners looked at the individual that was congruent with the occupational stereotype more often, and for longer, than they looked at the individual who was incongruent regardless of whether gender stereotype information was salient for understanding the context of the story. From this, they conclude that occupational gender stereotypes guide social perception in Finnish even when it is not salient.

### 1.3.3. Theoretical models of the interplay between grammatical gender and occupational gender stereotypes

Based on research examining fully and semi-gendered languages, it appears that there is an important interplay between how grammatical gender and occupational gender stereotypes affect social perception. As this interplay has not been examined between fully, semi-, and non-gendered languages, this presents an opportunity for greatly expanding knowledge in this area. Based on this previous research, and on the activation selection model, meaning activation theory, and social role model, two theoretical models were designed for this thesis; the grammatical gender bias model, and the stereotype salience model. Both hypotheses suggest that speakers of fully gendered languages are especially sensitive to all linguistic elements relating to gender, with the constant reinforcement of grammatical gender leading to other factors relating to gender (e.g., semantic gender and gender stereotype) being perceived as highly relevant for informing social perception.

The grammatical gender bias model argues that, as grammatical gender markers reduce in prevalence from fully to semi- to non-gendered languages, attributes related to gender in general (and which therefore might lead to the activation of gender stereotypes) are activated most frequently in fully gendered languages and least often in non-gendered languages. This

more frequent activation is held to increase stereotype importance, meaning that occupational gender stereotypes should have the largest effect on social perception for fully gendered language speakers, and the smallest effect for non-gendered language speakers.

Core to the stereotype salience model is the concept of correction of incorrectly activated attributes. This process occurs when new knowledge alerts us to the fact that we have activated certain attributes incorrectly. For example, consider a situation in which you are talking to a close friend about their health. They say to you "I saw my new doctor yesterday". Based on occupational gender stereotypes, you activate a mental representation that incorrectly includes the attribute 'male'. Your friend then says, "Her name is Breanna". This explicit statement of gender 'corrects' your mental representation, deactivating the 'male' attribute and activating a 'female' attribute instead. When an attribute is corrected, it has still been activated for some period of time. While this does not lead to as strong an increase as if it was never corrected, the weight of the attribute still increases. This is thought to happen in relation to the amount of time that passed before the correction, with longer periods of time correlated with larger increases in weight. Based on this idea, the stereotype salience model argues that the lack of gender markers in non-gendered languages can lead to an inability to correct gender attributes that are incorrectly activated, with the effect that occupational gender stereotypes should have the largest effect on social perception for fully gendered language speakers, and the smallest effect for non-gendered language speakers.

### 1.3.4. Summary of this section

In summary, grammatical gender is a noun-class system where nouns are grammatically assigned genders. The level to which this occurs differs based on the level of grammatical gender within a language. For the purposes of this thesis, the important classifications of languages based on grammatical gender are fully gendered languages, where all nouns and pronouns used to refer to humans are grammatically gender marked, semi-gendered languages, where some but not all nouns and pronouns used to refer to humans are grammatically gender marked, and non-gendered languages, where no nouns and pronouns used to refer to humans are grammatically gender marked. Research into the interplay between grammatical gender and occupational gender stereotypes is still preliminary, and no research has yet examined many facets of it. This led to the expansion of the first research focus of this thesis; increasing knowledge of the interplay between grammatical gender and occupational gender through examining fully, semi-, and non-gendered languages. Based on theory and on previous research, two models were identified. The first, the grammatical gender bias model, holds that

occupational gender stereotypes will have the greatest impact upon social perception in fully gendered languages, and the least in non-gendered languages. The second, the stereotype salience model, holds that occupational gender stereotypes will have the greatest impact upon social perception in fully gendered languages, and the least in semi-gendered languages. Importantly, under current understandings of the interplay between grammatical gender and occupational gender stereotypes, both models are equally possible.

## 2. Aims

This thesis aims to explore the complex interactions between two common but separate forms of social stereotype, in order to expand understanding about how social perception is affected when multiple salient schema are activated concurrently and in different cultural contexts. To this end, this thesis focuses on the interaction between gender stereotypes and occupational stereotypes (Paper II, Paper III), and on the interplay between grammatical gender level, gender stereotypes, and occupational stereotypes (Paper III). Three papers are included in this thesis, consisting of one methodological paper (Paper I, two experiments) and two experimental papers (Paper II, two experiments; Paper III, three experiments).

The specific aims of this thesis are as follows:

1. To determine stereotype contents for occupational gender stereotypes;
2. To evaluate whether gender ratio is representative of gender stereotypicality;
3. To examine the importance of occupational gender stereotypes in informing social perception in isolation from grammatical gender; and
4. To examine the importance of the interplay between grammatical gender and occupational gender stereotypes in informing social perception.

# 3. Methodology

This section focuses on the specific experimental approaches utilised across the course of the papers presented in this thesis. The use of different methodology throughout this thesis was intended to ensure that the examination of the interplay between grammatical gender and occupational gender stereotypes covered a broad area to enhance the knowledge able to be gained.

## 3.1. Experimental approaches by paper

### 3.1.1. Paper I

This study was designed to compare responses and response times measured through the internet-based implementation of PsyToolkit (Experiment 1) with those measured by the laboratory-based implementation of E-Prime 3.0 (Experiment 2) using a two-alternative forced choice design. This allowed for examination of the interaction between gender stereotype and occupational stereotype, as well as for determining whether PsyToolkit could be relied upon for complex choice response tasks in psycholinguistic research.

Participants all self-identified as native Norwegian speakers who were current students at NTNU, Norway. To avoid cognitive costs associated with modality switching, all experimental elements were presented in Norwegian.

### 3.1.2. Paper II

This study was designed to examine the interaction between gender stereotypicality and occupational stereotypes, through a ground-up method. Two studies were conducted over the course of this paper. In the first, participants undertook a spontaneous attribute naming task. In the second, participants undertook an attribute rating task, which utilised a Likert scale. Study 1 was conducted using hardcopy questionnaires, while Study 2 was conducted through the internet-based implementation of PsyToolkit.

Participants all self-identified as native English speakers. While many participants were university-level students, this was not a requirement for inclusion in this paper.

### 3.1.3. Paper III

This study was designed to examine the interplay between grammatical gender, gender stereotypes, and occupational stereotypes through comparing responses and response times between native speakers of French (Experiment 1), Norwegian (Experiment 2), and Finnish (Experiment 3). This was achieved using a two-alternative forced choice design very similar to that used in Paper I. Responses to all experiments were measured through the internet-based

implementation of PsyToolkit.

Participants in each experiment all self-identified as native speakers of the language examined in the experiment that they took part in, and as university-level students. To avoid cognitive costs associated with modality switching, all experimental elements for each experiment was presented entirely in the language under examination in that experiment.

### 3.2. Two-alternative forced choice tasks

The two-alternative forced choice tasks utilised in Paper I and Paper III of this thesis were very similar in style, and built upon the same psycholinguistic paradigm (e.g., Gygax & Gabriel, 2008; Siyanova-Chanturia et al., 2015). In this task, participants were presented, in the language of which they were a native speaker, pairs of terms composed of a first name (e.g., David) and a role noun in the plural form (e.g., Architects). They were then required to indicate, as quickly as possible, whether they believed that an individual called [name] could be a member of the group of [noun]. These pairings were always presented in the form '[name] – [noun]' (e.g., David – Architects), and presentation order was randomised by participant. For all experiments in both Paper I and Paper III, responses were given via keyboard. Participants were instructed to press 'e' if they did not agree that the individual could be a member of the group indicated, and to press 'i' if they did agree. Participants had five seconds to respond to each item. Failure to respond to a specific item within that time meant that the instrument would record that it had not received an answer, and the experiment would move on. After an answer was given, either through the participant pressing a key or the item timing out, the pairing was replaced with a fixation cross for 100ms, and then the next pairing was displayed. Participants undertook a five-item training phase before the main experimental phase, which was composed of 360 name-noun pairings. The stimuli for all experiments in both Paper I and Paper III were composed of six first names which were paired with 36 role nouns and with 36 filler items. The exact stimuli differ between Paper I and Paper III.

For Paper I, stimuli were identical for both experiments conducted over the course of the study. The six names used were composed of three male and three female names, selected based on the findings of Öttl (2018). The specific role nouns selected were chosen based upon the findings of Misersky et al. (2014), who examined perceived gender ratios as a measure of occupational gender stereotypicality across many European languages, including French, Norwegian, Finnish, and English. Each occupation was given a ranking between 0 and 1 in each language, with 1 representing a completely stereotypically feminine occupation, 0 representing a completely stereotypically masculine occupation, and 0.5 indicating a non-stereotyped

occupation Their findings for Norwegian were utilised as a basis for the selection of 12 stereotypically feminine occupations (e.g., beautician), 12 stereotypically masculine occupations (e.g., roofer), and 12 non-stereotyped occupations (e.g., artist). The ratings determined by Misersky et al. (2014) allowed for balancing the stereotypicality levels of the selected occupations. Each stereotypically feminine occupation selected was paired with a stereotypically masculine occupation of a similar strength (e.g., a stereotypically feminine role with a rating of 0.9 paired with a stereotypically masculine role with a rating of 0.1). The stereotypicality of the non-stereotyped roles was kept as close to the center, 0.5, as possible. The filler items selected were gender-marked kinship terms. Half were female gender marked (e.g., 'sister'), and the other half were male gender marked (e.g., 'brother'). No other experimental tasks were undertaken in this paper.

For Paper III, stimuli for names and filler items did differ to some degree, but role nouns were kept constant across all three experiments. Names were selected based on different criteria per language, but with an overarching attempt to keep the names of equal lengths between languages to avoid confounds due to reading time differences. The specific role nouns selected were again chosen based upon the findings of Misersky et al. (2014), with gender stereotypicality balancing occurring between languages as well as between gender stereotype levels. In other words, the occupational roles selected were chosen due to having relatively stable levels of stereotypicality across all three languages examined as well as due to the factors guiding selection in Paper I. In this way, errors that may have been caused due to different occupations being used or by different levels of stereotypicality in the examinations across languages are removed. The filler items selected were, as much as possible, gender-marked kinship terms that were the same across all languages, with half being female gender marked and half being male gender marked. However, due to some words not existing in certain languages, explicitly gender-marked occupational roles (e.g., king) were also used to replace these roles, to again ensure uniformity in presented roles between languages. This was the first experimental task participants undertook in this paper.

Data preparation and analysis for both Paper I and Paper III followed the same protocol. Prior to data analysis, item-by-participant deselection and by-participant data screening was used. In keeping with Schubert, Murteira, Collins, and Lopes (2013), responses faster than 300ms or not occurring within 5000ms were removed from the data. By-participant data screening was composed of removing participants outside of the target demographic of Norwegian speaking university students, and of removing participants who had an error rate above 50% as calculated from the percentage of incorrect answers to all filler items. This

calculation was based upon the assumption that correct answers for congruent filler pairings is 'yes' and for incongruent filler pairings is 'no'. Following data screening and deselection, participants' yes/no responses were analysed through generalized linear mixed-effect regression, while participants' response times for 'yes' responses were analysed through linear mixed-effect regression. For both analyses, initial models composed of all experimental factors, their interactions, and random intercepts were defined, and then underwent refinement. Refinement occurred in keeping with Baayen (2008) and Baayen and Milin (2010); that is, the model of best fit was determined through back-fitting the fixed effects structure, forward-fitting the random effects structure, then re-back-fitting the fixed effect structure. This model fitting was done automatically through the *bfFixefLMER_F*, *ffRanefLMER*, and *fitLMER.fnc* functions of the lme4 package (Version 1.1-12; Bates et al., 2015) in R. Following identification of the models of best fit, the general linear mixed-effect regression and linear mixed-effect regression were conducted through the *glmer* and *lmer* functions of the lme4 package in R respectively, and post-hoc analysis for main effects and interaction effects found was conducted through the effects function of the effects package (version 4.1-0; Fox, 2003) in R. Local effect size estimates were also determined, through the *omega_sq* function of the sjstats package (Version 0.17.5; Lüdecke, 2018) in R.

### 3.3. Spontaneous attribute naming task

The spontaneous attribute naming task utilised in Paper II built upon an existing experimental paradigm (Koivula, 2001; Glick, Wilk, & Perreault, 1995). In this experiment, 36 occupational roles were selected as experimental items based on the findings of Misersky (2014) in relation to English. As with the two-alternative forced choice experiments, the ratings provided by Misersky were utilised for balancing the stereotypicality levels of the selected occupations, with each stereotypically feminine occupation selected being paired with a stereotypically masculine occupation of a similar strength, while non-stereotyped occupations were selected for being as close to the center, 0.5, as possible. Over the course of this task, participants selected the feminine, masculine, and non-stereotyped occupation that they were most familiar with, and the one occupation that they were the least familiar with, from the list. For each occupation they were instructed to, as quickly as possible, list up to five essential attributes that an individual working in that occupation should have. All responses received were then collated into a single document which listed participant number, occupation under discussion, and named attribute. This was the first experiment in this study. Due to how this

study was designed, two sets of data preparation and analysis occurred.

For the first set, data preparation involved by-item deselection for attributes named by only one participant (step one), were occupation-specific (step two), were specifically gendered (step three), or were named only once (step four). Steps one and four were in keeping with Koivula (2001) and with Glick et al. (1995). After deselection, the remaining dataset was handled in two separate manners. Firstly, the results as they were, with information relating to participant number and occupation under discussion (dataset 1), were stored as the basis for later reanalysis. Secondly, a document was created in which the uniquely named attributes were recorded once each, and with no accompanying information (dataset 2). Dataset 2 was utilised in the attribute rating task, discussed in detail below, and allowed for the identification of occupational stereotypes based on attribute groupings. This in turn allowed for reanalysis of the results of this experiment.

The results of the Likert scale task did not identify specific agentic nor communal occupational stereotypes. Since the idea that agentic is masculine and communal is feminine is so common in stereotype research, a new dataset (dataset 3) was created through assigning the attributes in dataset 2 as belonging to one of two groups, 'agentic' and 'communal', based upon previous research into agentic and communal traits in occupational settings (Abele et al., 2008; Bem, 1974; Spence, Helmreich, & Stapp, 1975). The aim of this dataset was not to replace the occupational stereotypes determined through the Likert scale task, but rather to explore possible reasons why these were not found.

The reanalysis of the results of this task occurred in two parts. Two columns were added to dataset 1; the first listed, next to each attribute, which occupational stereotype, if any, it had significantly loaded on, while the second listed, again next to each attribute, whether the attribute belonged to the 'agentic' or 'communal' grouping, or whether it belonged to neither. For the first part, attributes were not associated with any occupational stereotype were not included in the reanalysis, and for the second part, attributes that were not associated with either the 'agentic' or the 'communal' grouping were not included in the analysis. For each part, frequency analysis was conducted through the *CrossTable* function of the gmodels library (Version 2.18.1; Warnes, Bolker, Lumley, & Johnson, 2018) in R. Analysis of the first part examined the frequency at which attributes within each occupational stereotype were named for feminine, masculine, and non-stereotyped occupations, as a measure of stereotype salience during spontaneous naming tasks. Analysis of the second part examined the frequency at which attributes within each grouping were named for the feminine, masculine, and non-stereotyped

occupations, as a measure of the perceived importance of agentic and communal traits for occupations with varying levels of occupational gender stereotype.

### 3.4. Likert scale tasks

One Likert scale tasks were utilized over the course of the papers in this thesis. This was an attribute rating task in Paper II; the second experimental task in this paper. Participants in this task did not take part in the previous task, spontaneous attribute naming. Participants who undertook this experiment were presented with one feminine, one masculine, and one non-stereotyped occupation. These occupations were selected randomly by participant from the same 36 occupations that were included in the spontaneous attribute naming task. The order of stereotypicality for the occupations (i.e., when in the experiment they were presented with the feminine, the masculine, and the non-stereotyped occupation) was also randomised by participant. For each occupation, participants were presented with all the attributes within Dataset 2 from the spontaneous attribute naming task discussed above. For each attribute, participants were instructed to indicate the level to which they thought it was important or an individual working in that occupation to either *be* or to *have*, on a scale ranging from 1 (not at all important) to 7 (vitally important). The order in which attributes were shown was randomised both by occupation and by participant. To ensure balanced data by gender ratio category and occupation, the experiment was run until a minimum of ten participants had responded to each occupation.

Initial analysis was conducted through Parallel Analysis, Scree-Testing, Principal Components Analysis, and Rotated Components Analysis. Parallel Analysis, Scree-Testing, and Principal Components Analysis are commonly used to determine the number of components a dataset includes, as well as their cumulative variance. This is then used to inform Rotated Components Analysis, so that the correct form of rotation and number of components are utilized in modelling components to fit the data. The identified components were then named as occupational stereotypes based upon the attributes of which they were each composed. After this, a new dataset was created (dataset 4). The basis for this dataset was the raw data obtained from the Likert scale task, but with an added column which stated, next to each attribute, which occupational stereotype, if any, that the attribute had significantly loaded on. Attributes not loading significantly on any occupational stereotype were removed from the dataset. It was also planned to remove attributes which had loaded significantly on multiple occupational stereotypes, but this proved unnecessary. As mentioned above, no specific agentic

nor communal occupational stereotypes were identified. In order to explore the possible reasons for why these were not found, it was decided to explore this. As such, a new dataset was created (dataset 5), in which a column was added to the raw data from the Likert scale task which stated, next to each attribute, which group, if any, the attribute belonged to. Attributes not belonging to either the 'agentic' or 'communal' group were removed from the dataset.

Linear mixed-effect regression was used to examine both dataset 4 and dataset 5. For dataset 4, this examination focused on the level to which participants viewed each of the occupational stereotypes identified as important for feminine, masculine, and non-stereotyped occupations. For dataset 5, this examination focused on the level to which participants viewed agentic and communal traits as important for feminine, masculine, and non-stereotyped occupations. Analysis of both datasets was done in the same manner. As with the analyses of the two-alternative forced choice tasks above, initial models composed of all experimental factors, their interactions, and random intercepts were defined, and then underwent refinement. Refinement occurred through determining the model of best fit through back-fitting the fixed effects structure, forward-fitting the random effects structure, then re-back-fitting the fixed effect structure automatically.

### 3.5. Methodological rationales

The word association paradigm utilised in this thesis takes the form of a complex choice response task. Choice response tasks involve presenting participants with multiple different stimuli, and requiring the participant to respond in a variety of manners depending on the specific stimuli presented (e.g., Zajdel & Nowak, 2007). For example, participants are instructed to look at a screen upon which the letters 'a', 'g', and 'd' appear, one at a time, and in a random order. Using a regular keyboard, they are instructed that, when a letter is shown on the screen, they should press the corresponding letter on the keyboard as quickly as possible. This contrasts with simple choice response tasks where a single stimulus, and only that stimulus, is presented repeatedly at a set location, with participants responding to every presentation of the stimulus in the same manner (e.g., Zajdel & Nowak, 2007). An example of this is participants watching an LED and pressing a button every time it flashes. Choice response tasks differ in difficulty considerably. Simple choice response tasks, such as the keyboard example above, require participants to recognise a stimulus and indicate which stimulus it was. Complex choice response tasks also require participants to make specific judgements as to the nature of the stimuli. The word association paradigm falls under this category of complex choice response

task, as it requires participants to make judgements as to whether individuals 'belong' within specific social roles. This paradigm utilises the 'yes-no' two-alternative forced choice task specifically. This task provides a simple measure for examining participants' subjective experiences, allowing for examination of both choice (amount of times they responded 'yes' or 'no') and of response time. In the context of this thesis, this specific task focuses on occupational gender stereotypes, with grammatical gender examined through comparing responses gained to this task between languages. As this task is focused on occupational gender stereotypes, and naturalistically it is possible for anyone, regardless of gender, to hold any occupational role, analysis of the results for choice are based on examinations of the percentage of times participants responded 'yes' for each of the experimental conditions. Analysis of response times was conducted only for those items for which participants responded 'yes'. Response time analysis is based on the theory of cognitive load (Paas, Renkl, & Sweller, 2003). This theory holds that cognitive tasks require set amounts of cognitive resources, out of a limited pool, from working memory to be able to be completed successfully. (King & Bruner, 2000). Further, it assumes that working memory has limited capacity for novel information, but near limitless capacity for familiar material (Paas, Renkl, & Sweller, 2004). Activation of familiar information in the form of schema or stereotypes is an automated process, essentially bypassing working memory. Activation of novel information, however, requires active processing in working memory, therefore taking longer to complete. In experimental terms, increasing levels of cognitive load are associated with increased reaction times, as participants have less cognitive resources available for responding to the task. Under this approach, longer response times can be seen to indicate novel information, while shorter response times can be seen to indicate automated information. As such, if an individual was presented with a word association task involving the pairings 'David – Nurse' and 'Sarah – Nurse' and, while responding positively to both, took longer to respond positively to 'David – Nurse', the concept of David as a Nurse could be seen as novel information lying outside of the participants' stereotype beliefs. As such, choice can be understood to measure participants' *explicit* stereotype beliefs, while response time can be understood to measure their *implicit* stereotype beliefs.

The exact word association paradigm utilised in this thesis is based upon similar paradigms used in previous research into the interplay between grammatical gender and occupational stereotypes (e.g., Gygax & Gabriel, 2008; Siyanova-Chanturia et al., 2015). This was decided upon to ensure that there was a high level of comparability between the results presented in this thesis and this previous research. One key change was made from this previous

work. As discussed above, the word association paradigm we use compared gender stereotyped occupational roles with explicitly gendered first names. Previous research compared gender stereotyped occupational roles with gender marked kinship terms (e.g., son, daughter). This shift away from kinship terms was done based on the fact that kinship terms inherently include age related information; for example, you are likely to perceive the term 'grandmother' as referring to an older individual, and to perceive the term 'son' as referring to a younger individual. Since this previous research did not correct or control for age, it is very likely that this age-related information adds additional unexplained noise to the results obtained through the paradigm, making it more difficult to determine the exact interaction effect between occupational stereotypes and gender stereotypes. While it is possible that the explicitly gendered first names utilised in this thesis may also be subject to extraneous associations, (e.g., 'all the nurses I know are called Sarah'), it is likely that this occurs on a more individualistic level than is the case for kinship terms. As such, any noise that occurs is likely to be individualistic rather than systematic, and therefore easier to address during analysis.

The attribute naming and attribute rating paradigms utilised in this thesis were based upon similar paradigms used in previous research that has examined stereotype contents relating to different stereotypes in a variety of manners (Glick et al., 1995; Imhoff et al., 2018; Koivula, 2001). The paradigms used in this thesis are closest to those used by Koivula (2001). The attribute naming paradigm utilised by Koivula (2001) consisted of participants naming an unlimited number of attributes for two sports of their own choosing (one that they were very familiar with, one that they were not at all familiar with), while the attribute rating paradigm consisted of participants being randomly presented with one of 41 sports (19 non-stereotyped, 15 masculine stereotyped, 7 feminine stereotyped; Koivula, 1995) and asked to rate the level to which all attributes determined from the results of the attribute naming task were important for that sport, with occupation selection and attribute presentation order randomised by participant. Since Koivula (2001) did not control for gender stereotypicality in sport distribution, it is possible that some of the results obtained by them, especially around richer masculine representations, are due to more people answering on stereotypically masculine sports during both experiments. As such, the attribute naming and rating paradigms used in this thesis were designed to avoid this as a possible explanation. Occupations were selected and balanced for stereotypicality from the results of Misersky et al. (2014), with equal numbers of feminine, masculine, and non-stereotyped roles, and with the feminine roles selected matched with masculine roles perceived as being as equally masculine as the feminine roles are perceived as feminine. These occupations were identical during both the attribute naming and rating tasks.

For the attribute naming paradigm, participants spontaneously naming a maximum of five attributes each for four occupations. Selecting from the predefined list, these occupations were the feminine, masculine, and non-stereotyped occupation that they were most familiar with, and the occupation regardless of gender stereotypicality that they were least familiar with. Under the attribute rating paradigm used in this thesis, participants responded to one feminine, one masculine, and one non-stereotyped occupation, each of which are paired with all attributes, and with occupation, occupation presentation order, and attribute presentation order randomised by participant and by occupation.

While this thesis does not examine stereotype strength, the results obtained by Misersky et al. (2014) across multiple languages including English, French, Finnish, and Norwegian are utilised throughout this thesis as a basis for determining occupational gender stereotypes. This included an examination of whether the underlying assumption of gender ratio as a measurement of gender stereotypicality is valid (Paper II).

As this thesis includes examinations of multiple languages, participants in all experiments undertaken over the course of this thesis were presented with all experimental information in their native language to prevent any issues due to modality switching costs and priming effects between languages. This was not only the materials presented during the experiments, but also the briefing information, demographic questionnaire, and informed consent forms prior to each experiment, the debriefing information subsequent to each experiment, and the user interface (when present) throughout each experiment.

# 4. Results

## 4.1. Paper I

The first aim of this paper was to begin to explore the importance of occupational gender stereotype information for Norwegian speakers' social perceptions, in order to determine whether altering the experimental paradigm to utilise first names instead of kinship terms significantly impacted upon the results obtained. The results for both choice and response time were in keeping with previous research utilising similar paradigms for semi-gendered languages (e.g., Gygax & Gabriel, 2008; Siyanova-Chanturia et al., 2015), and indicated, although not to a significant degree, that the perceived gender of an individual serves as a basis for judgements as to their ability to 'belong' within occupational roles. For choice, participants responded more positively to congruent pairings (i.e., pairings where the gender of the first name matched the stereotype of the role noun) than to incongruent pairings (i.e. pairings where the gender of the first name went against the stereotype of the role noun) for both female and male names, and generally responded 'yes' more frequently for both female and male names paired with non-stereotyped roles than with incongruent roles. For response time, participants generally responded faster to congruent pairings than to incongruent pairings for both female and male names. These results indicate that congruent pairings were more in keeping with participants' existing stereotype beliefs than incongruent pairings.

The second aim of this paper was to explore whether PsyToolkit was a good instrument to utilise over the course of this thesis, to aid with international experimentation. The results of both choice and response time indicated that it was indeed a good instrument for this purpose. The results suggested a main effect of experimental form for response time, with participants in Experiment 1 (PsyToolkit) responded faster than those in Experiment 2 (E-Prime 3.0), but no main effect of experimental form for choice. As the aim of the paper was to explore whether PsyToolkit produced replicable results, the three-way interplay between experimental form, gender stereotypes, and occupational stereotypes was also examined. This was found to be non-significant for both choice and response time, but, as the non-significance is interesting in and of itself, post-hoc analysis was still conducted on this interaction. For choice, the interaction indicated no significant differences in mean responses to each condition between Experiment 1 and Experiment 2, and supported the finding above that participants responded more positively to congruent pairings than incongruent pairings. For response time, this interaction indicated that there were no significant differences in mean responses to each condition between

Experiment 1 and Experiment 2, with participants in both experiments responding faster to congruent pairings compared to incongruent pairings.

### 4.2. Paper II

The results of Study 1 (spontaneous naming task) indicated five occupational stereotypes. These were named based upon the attributes which loaded significantly upon each rotated component. They were thus defined as *Interpersonal Skills*, *Precision*, *Creativity*, *Physicality*, and *Work Identity*. As these groupings did not include explicitly agentic nor communal groupings, groupings for these were determined based upon previous research (Abele et al., 2008; Bem, 1974; Spence, Helmreich, & Stapp, 1975), and analysis based upon these artificial groupings was also conducted.

The results for the interaction between gender stereotypicality and occupational stereotypes indicated not only the existence of inherently feminine (Interpersonal Skills) and masculine (Physicality, Precision) occupational stereotypes, but also the existence of inherently unfeminine (Work Identity) and unmasculine (Creativity) occupational stereotypes. Interestingly, for the results of Study 1, participants only spontaneously named attributes related to Physicality for masculine stereotyped occupations, but, for the results of Study 2 (attribute rating task), rated Physicality as unimportant but significantly above 'completely unimportant' for both feminine and non-stereotyped occupations.

The results for the interaction between gender stereotypicality and agentic/communal attributes indicated that participants spontaneously named communal attributes more often, and perceived them as more important, for feminine compared to masculine occupations. The results also indicated that participants viewed agentic attributes as being equally important across all occupations, but were most likely to spontaneously name agentic attributes when responding about non-stereotyped occupations. The finding that communal attributes were most strongly associated with feminine stereotyped occupations is in keeping with previous research (e.g., Eagly & Wood, 2016), but the finding that agentic attributes were most strongly associated with non-stereotyped occupations is not.

### 4.3. Paper III

Multiple analyses were conducted over the course of this experiment, but they can be generally classified as examining the interplay between grammatical gender, gender stereotype, and occupational stereotype.

The primary aim of this paper was to explore the interplay between grammatical gender,

gender stereotypes, and occupational stereotypes and, in doing so, determine whether the "grammatical gender bias" theory (i.e., that gender stereotypes will be activated least in non-gendered languages) or the "salience of stereotype" theory (i.e., that gender stereotypes will be activated least in semi-gendered languages) were supported by the results. As with Paper I, this paper utilised explicitly gendered first names to activate participants' occupational gender stereotype beliefs. The results for both choice and response time were in line with the "salience of stereotype" theory, indicating that semi-gendered language speakers were the least affected by gender stereotypes, while fully gendered language speakers were the most affected by gender stereotypes. For choice participants across all languages responded relatively similarly to congruent pairings, but, for incongruent pairings, semi-gendered language speakers were likely to respond more positively than either fully or non-gendered language speakers, while fully gendered language speakers were more likely to respond negatively to incongruent pairings than either semi- or non-gendered language speakers.

## 5. General Discussion

The intention behind this thesis was to expand knowledge of how language structures affect stereotyped beliefs through the examination of how linguistic differences between languages affected social perception relating to the interactions between different stereotype categories. The research conducted in the course of this thesis can be seen to expand knowledge relating to the importance and content of occupational gender stereotypes. The aims of this research were to determine stereotype content for occupational gender stereotypes (Aim 1); to evaluate whether gender ratio is representative of gender stereotypicality, as had previously been assumed (Aim 2); and to examine the importance of occupational gender stereotypes both in isolation from (Aim 3) and interacting with (Aim 4) grammatical gender.

### 5.1. Overarching methodological implications

In this thesis, Paper I represented a methodological examination, while Papers II and III represented theoretical examinations. The focus of Paper I on methodological issues provides a good basis for the examinations conducted over the course of this thesis, both in terms of exploring occupational gender stereotypes, and in terms of determining whether the internet-based instrument PsyToolkit was adequate for the needs of this thesis.

In relation to exploring occupational gender stereotypes (Aim 3), the results of Paper I and Paper III showed that participants were more likely to respond positively to gender congruent than to gender incongruent pairings. Further, when responding positively, participants were more likely to respond faster to gender congruent pairings. The exact results obtained were largely in keeping with previous research (Garnham et al., 2012; Gabriel et al., 2008), and supports the idea that occupational stereotypes are strongly affected by gender stereotypes, as incongruent pairings are associated with higher cognitive load. The slight variations observed between the results obtained in this thesis and those obtained in past research is in keeping with the theory that kinship terms, as utilised by Garnham et al. (2012) and Gabriel et al. (2008) inherently hold age-related information that introduces experimental noise.

The finding that PsyToolkit has a high level of replicability of both response choice and response time, and low level of excess noise, when compared to the laboratory-based implementation of E-Prime 3.0 strongly supports the use of PsyToolkit for complex choice experiments not only within the confines of this specific thesis, but in terms of cognitive research (and especially psycholinguistic research) in general. It is interesting to note that mean

response time (and associated standard deviations) found for the results of both PsyToolkit and E-Prime are much higher than that found in previous research using choice response tasks to examine the replicability of internet-based instruments (Reimers & Stewart, 2007; Schubert et al., 2013). It is likely that this is due to the increased complexity of the experimental paradigm utilised in this paper, as neither Reimers and Stewart (2007) or Schubert et al. (2013) utilised tasks requiring participant judgement prior to responding. Reimers and Stewart (2007) presented participants with green and red coloured blocks, and instructed them to press buttons matching the colour of the block. Participants in Schubert et al. (2013) were presented with a version of the Stroop task consisting of the words 'red', 'blue', or a neutral letter string, in red or blue on a white background, and were instructed to press keys corresponding to the colour of the word/letter string shown.

In the wider context of stereotype interactions, the papers in this thesis support the use of the word association, attribute naming, and attribute rating paradigms for use in exploring stereotype interaction effects. This is perhaps harder for the word association task; the specific paradigm we utilised in Papers I and III used explicitly gendered first name to activate occupational gender stereotype beliefs held within gender stereotyped occupational roles. While explicitly gendered first names could again be utilised in an examination of sports-based gender stereotypes, other stereotype interactions, for example age-based occupational stereotypes, would require determining an adequate external source of information for activating stereotype beliefs. This requirement is not a fundamental aspect of the attribute naming and attribute rating tasks, which simply require identification of the components of interest for both stereotype categories under examination. It is worth noting that these approaches are best when they are informed by research that properly provides a basis from which to accurately define levels of the stereotype interaction of interest, such as Misersky et al. (2014) provide for the occupational gender stereotypes examined in this thesis.

### 5.2. Interplay between grammatical gender and occupational gender stereotypes provide support for the 'stereotype salience' hypothesis.

The focus of Paper III was on determining the interplay between grammatical gender and occupational gender stereotype beliefs. As mentioned above, when averaged across all languages, the examination of occupational gender stereotypes (Aim 3) indicated that participants were more likely to respond positively, and, when responding positively, to respond more quickly to gender congruent name/occupational noun pairings than to gender incongruent

pairings. The evaluation of the interplay between grammatical gender and occupational gender stereotypes (Aim 4) strongly supported the 'stereotype salience' hypothesis. Semi-gendered language speakers were found to be less likely to activate occupational gender stereotype beliefs than either fully or non-gendered language speakers, while fully gendered language speakers were found to be more likely to activate these beliefs than non-gendered language speakers. Under the 'stereotype salience' hypothesis, it is theorised that this is caused by a lack of potential sources of correction when gender stereotyped attributes are incorrectly activated in non-gendered languages. This lack of correction increases the weight of gendered attributes to the point where gender stereotyped attributes are always perceived as salient, while the increased opportunities for correction in semi-gendered languages reduces the comparative weight of these attributes and allows for less reliance on occupational gender stereotypes, and the general increase in salience of gendered information in fully gendered languages leads to the highest level of reliance across any level of grammatical gender.

For fully gendered language speakers, grammatical gender was found to have a modulating effect on social perception. However, this effect was relatively weak, and did not counteract the congruent/incongruent effect of occupational gender stereotypes discussed above. This finding is not in keeping with previous research (e.g., Gygax & Gabriel, 2008; Gygax et al., 2008), which found the inverse. This may have been due to differences in experimental paradigm. First name gender was utilised in this paper as the implicit indicator of physical gender, whereas both Gygax and Gabriel (2008) and Siyanova-Chanturia et al. (2015) utilised kinship terms as this indicator instead. If this is true, then it suggests that the kinship terms did include other attributes (such as age-related information) that impacted upon the results obtained by Gygax and Gabriel (2008) and Gygax et al. (2008). This would be in keeping with what was discussed in the methodological rationale section (3.5), suggesting that the paradigm used for this thesis produces less experimental noise. It is also possible that social perception has shifted over time. Over a decade has passed between Gygax and Gabriel (2008) and Gygax et al. (2008), and it is possible that fully gendered language speakers' social perceptions have fundamentally shifted to the point where, regardless of paradigm, gender stereotypes are more salient than grammatical gender in guiding social perception.

A slight 'feminine bias' was found for semi- (Norwegian) and non-gendered (Finnish) languages, consisting of an increased difficulty in accommodating male names paired with feminine stereotyped role nouns. This finding was unexpected. A possible explanation for this bias is that, historically, women have been 'allowed' to belong to far fewer professions than men. As societal beliefs changed, women were able to enter more historically masculine roles

than men were able to enter more historically feminine roles. The slow shift in cultural beliefs would therefore be more heavily focused on women entering masculine roles. It follows that, with increased focus and thus awareness of the ability for women to hold masculine roles, the ability for anyone to hold any role is especially salient for semi- and non-gendered language speakers when considering women in counter-stereotypical occupations. Interestingly, this finding suggests that the flexibility of stereotypes within an individual is stronger than the rigid transmission of information cross-generationally. This finding is not in keeping with social role theory, which holds that the cross-generational transmission of cultural beliefs should form the core for stereotype beliefs. It is possible that this finding is due to the purposeful attempt within Norwegian to degender the language through masculinisation of all role nouns (Gabriel & Gygax, 2008). As this includes a government-level mandate for change in social perception, including changes in word selection in both official documentation and in media, it may be that the increased speed with which cultural beliefs have shifted, encouraging flexibility and discouraging traditional stereotype communication, has led to heightened perception of women, but not for men, to hold occupational roles regardless of stereotypicality.

### 5.3. Occupational stereotyped attributes and their connection to feminine, masculine, unfeminine, and unmasculine gender stereotypes

The focus of Paper II was on determining occupational gender stereotypes through a bottom-up process. This exploration of whether gender ratio is representative of gender stereotypicality (Aim 1) found strong support for the idea, with all occupational stereotypes found to interact to at least some degree with gender stereotypicality. Five occupational stereotype categories were identified; creativity, interpersonal skills, physicality, precision, and work identity. One was seen as feminine (*interpersonal skills*), one as unfeminine (*work identity*), two as masculine (*physicality* and *precision*), and one as unmasculine (*creativity*). In relation to the exploration of whether gender ratio balancing affects occupational stereotype groupings , the specific groupings found were not in keeping with previous research that did *not* control for gender stereotypicality of occupations (e.g., Imhoff et al., 2018, who identified four stereotype groupings; *agentic/competent*, *progressive*, *social*, and *communal*). As occupations are more often masculine than feminine stereotyped (Kennison & Trofe, 2003; Gabriel et al. 2008), one explanation is that attributes associated with masculine stereotyped occupations were included at a much higher rate than attributes associated with feminine stereotyped occupations in the results obtained by Imhoff et al. (2018), serving to overrepresent

masculine-associated attributes and underrepresent feminine-associated attributes. It is also possible that these differences are due to fundamental cultural differences between languages, as Paper II focuses on native New Zealand English speakers while Imhoff et al. (2018) examined native US English and native German speakers.

The higher number of masculine-associated compared to feminine-associated occupational stereotypes in the results of Paper II suggest that masculine occupations generate richer cognitive representations than feminine occupations. This is in line with what Koivula (2001) found in relation to sports-based gender stereotypes, and suggests that feminine stereotypes are more narrowly defined than masculine stereotypes. The results also indicated differences in the salience of attributes based on the specific task; for example, participants rated attributes in the *physicality* grouping as being unimportant yet above 'completely unimportant' for the feminine and non-stereotyped occupations, but only named attributes in the *physicality* grouping for masculine stereotyped occupations.

Regarding Aim 2, the attribute groupings identified in the course of this paper were not in keeping with previous research (e.g., Eagly & Wood, 2016), with no specifically agentic or communal stereotype components identified. When agentic and communal attributes were assigned based on previous research (Abele, Uchronski, Suitner, & Wojciszke, 2008; Bem, 1974; Spence, Helmreich, & Stapp, 1975), agentic traits were not found to significantly relate to masculine stereotyped occupations, although communal traits were found to significantly relate to feminine stereotyped occupations. This suggests that, while communal traits are seen as feminine, agentic traits are seen as generically important. This finding is in keeping with Eagly, Nater, Miller, Kaufmann, & Sczesny (2019), who state that *positive* agentic traits have become non-stereotyped, although negative agentic traits are still perceived as masculine. One possibility is that, due to the slow shift of social perception under social role theory, the high number of occupations that are generally held to be feminine or non-stereotyped today but which historically have only been able to be held by men are still perceived as intrinsically 'needing' masculine stereotyped traits. As such, non-males within these occupations may still be perceived to some degree as 'not belonging', although this is likely to be counterbalanced for feminine-stereotyped occupations due to higher associations with feminine stereotyped attributes.

The finding that occupations can be stereotyped as unfeminine and unmasculine is not in keeping with previous research (e.g., Glick et al., 2015). It is possible, as Glick et al. (2015) do not utilise non-stereotyped occupations in their research, that non-stereotyped occupations allow for more nuanced interpretations of results obtained as it serves as a secondary benchmark

against which perceptions of feminine and masculine occupations can be measured. The existence of unfeminine and unmasculine occupational stereotypes is supported by research examining individual occupations; examples of unfeminine occupations are computing (Berki & Payton, 2015) and science (Kessels, Rau, & Hannover, 2006), while examples of unmasculine occupations are teacher and nurse (Allen & Smith, 2011). Of interest, lesbianism is perceived as being 'unfeminine', while male homosexuality is perceived as being 'unmasculine' (Adams, 2005; Oakenfull, 2013). This identification of inherently gendered personal attributes as being intrinsically incongruent with the gender of the individual who holds them strongly supports the idea of unfeminine stereotypes being intrinsically based upon femininity, and unmasculine stereotypes being intrinsically based upon masculinity. This is an important distinction to make, as it indicates a fundamental difference between unfeminine and masculine stereotypes, and between unmasculine and feminine stereotypes.

The existence of unfeminine and unmasculine stereotype categories suggests that we form concrete stereotypes for purely counter-stereotypical information, and that these counter-stereotypes are explicitly separate from stereotypes about secondary groups. It is not clear, however, whether these counter-stereotypical beliefs are stereotype categories in their own right, or whether they form part of the overarching stereotype category in relation to specific groups. Regardless of whether it is a part of or an entire stereotype category, this finding suggests that stereotype interactions may be a key manner by which these counter-stereotypes can be properly examined.

### 5.5. Methodological concerns

Aside from Experiment I in Paper II, all experiments conducted over the course of this thesis were done through the internet-based instrument PsyToolkit. While most of the issues relating to internet-based experimentation were addressed through testing the replicability of the instrument and through the exact methods used to structure the experiments and analyse the data obtained, one unavoidable issue is that of demographic information. The experiments by necessity assume that all participants were truthful about their demographic information, such as age, gender, and student status, but, as suggested by Reips (2002) and Reips et al. (2015), it is impossible to be completely sure that the demographic information participants provide through internet-based instruments is correct. This was addressed through avoiding direct discussions of these factors, to avoid making concrete claims about factors that may be influenced by incorrectly reported information. However, if this did occur, the high level of

replicability between PsyToolkit and E-Prime 3.0 found in the results of Paper I suggest that any effects from this are likely to be minor.

As the examination of the effects of grammatical gender on occupational gender stereotyped perception relied on only one language per level of grammatical gender. The decision to examine only one language per level of grammatical gender was taken due to time and resource constraints related to the PhD thesis process. The results obtained in this thesis are largely in keeping with previous research, with minor differences being attributed to changes in the exact experimental paradigm used, but as only one language per level is examined it is possible that other cultural aspects that differ between the countries examined (such as predominant religious or political beliefs with associated social attitudes) may at least partially impact upon the results obtained.

It should also be noted that, for the examination of agentic and communal traits in Paper II, only a small proportion of the attributes identified in the paper were able to be utilised. This was primarily due to previous research that has examined communal and agentic attributes focusing on individuals' personalities in their home and everyday lives. As such, many of the attributes identified in this previous research do not translate directly into attributes perceived as important in occupational settings. However, this is mitigated to some degree by the careful selection of occupations used in this thesis. Care was taken to ensure not only that there were equal numbers of feminine and masculine stereotyped occupations, but that each feminine stereotyped occupation was matched with a masculine stereotyped occupation based on perceived gender ratios to ensure that feminine stereotyped occupations were as feminine as the masculine stereotyped occupations were masculine. The results obtained in Paper II should therefore have a high external validity in relation to representing gender stereotyped information, and, as such, the gender associations naturally found in this paper are of more importance to this thesis than the artificially derived categories of agentic and communal traits.

### 5.6. Future directions for research

As the aim of this thesis was to explore the interplay between grammatical gender and occupational gender stereotypes as a method of exploring the effect of linguistic factors on stereotype interactions, the clearest direction for future research would be to continue to use the experimental paradigms utilised in this thesis for exploration in the area of stereotype interactions. An example that easily presents itself would be in relation to sports-based gender stereotypes. Building on the work of Koivula (1995; 2001), the utilisation of an in-depth cross-

linguistic rating task in the style of Misersky et al. (2014) could easily form the required basis for examination, allowing for both the word association and attribute naming/rating approaches to be utilised. Grammatical gender again suggests itself as a linguistic factor of interest if this approach is utilised, although another strong factor in this vein would be the level to which a language activates agentic information during autobiographical memory recall, as speakers of languages where agentic information is activated to a high degree (e.g., English) are far more likely to attribute actions to individuals (e.g., 'Sarah broke the table') than speakers of languages where agentic information is activated to a small degree (e.g., Spanish), where actions might not even be attributed to anyone (e.g., 'the table was broken').

In relation to the attribute naming/rating approach, an interesting question arises as to the effect of explicitly gendered information on the results obtained. If the generic perception of agentic traits is due to a hangover in historic attributes, then it would hold that individuals perceived as inherently displaying communal traits in those roles would be socially and/or economically punished (Shaw & Hoeber, 2003). One avenue for examination would be to explore the level to which individuals identified through explicitly gendered first names are, when paired with specific occupations, perceived as displaying each of the identified attributes. If the generic perception of agentic traits is indeed due to historic attributes, we would expect that women would be rated lower on displaying attributes required for 'belonging' within an occupation, especially for non-stereotyped occupations where other gendered information might influence participants' responses.

The existence of unfeminine and unmasculine occupational stereotypes poses an interesting question as well. As demonstrated throughout this thesis, and in research into gender stereotypes and their interactions on a wider scale, the framing is always in terms of a scale that goes from feminine through non-stereotyped to masculine. Using the results of Paper II as a basis for examination, one approach would be to explore the question of whether a framework that natively includes unfeminine and unmasculine categories might better explain ambiguities and overlaps than the current framing.

The approach to determining stereotype content utilised in this thesis was predicated on the idea that, in keeping with the concepts of stereotype strength and stereotype importance, occupational gender stereotyped attributes should remain relatively constant between languages even when grammatical gender levels differ. This is supported by the findings of Imhoff et al. (2018), who found very similar results in both English, a semi-gendered language, and German, a fully gendered language. However, since this has not been directly examined, an interesting area for future examination would be to examine the level to which occupational gender

stereotypes created from the ground up differ between languages with differing levels of grammatical gender.

In order to address the concern that only one language was utilised in the course of this thesis for the examination of the effect of grammatical gender on occupational gender stereotype perception, future research into the topic might examine differences in bilingual and multilingual speakers' occupational gender stereotype perception between languages with differing levels of grammatical gender that exist natively within their home country for example, examining bilingual Norwegian (semi-gendered) and Sami (non-gendered) speakers, or bilingual New Zealand English (semi-gendered) and Māori (non-gendered) speakers.

As differences found over the course of this thesis suggest that gendered first names and kinship terms fundamentally activate different attributes, it may be interesting to see what attributes are activated by kinship terms and gendered first names both in isolation from, and in connection, with other stimuli in order to determine whether there is a true fundamental difference between them. While it seems likely that the difference found in this thesis was due to noise from age-related information, it is also possible that gendered first names are more difficult to project upon more generally, leading to them activating less attributes (including gendered attributes) than occur for kinship terms.

As only a minority of previous studies into communal and agentic attributes has focused on the occupational context, one area for future research would be to examine this. It may be especially useful to undertake this in relation to a wide variety of possible contexts and languages, to provide a core guiding document for the examination of the 'communal as feminine' and the 'agentic as masculine' across a wide variety of stereotype interactions.

### 5.4. Conclusion

When considered as a whole, the findings of this thesis indicate that occupational gender stereotypes are rich, interesting constructions that strongly interact with the level of grammatical gender within a language. In keeping with previous research, occupational stereotypes and gender stereotypes interact in a clear manner. Occupational stereotypes – determined through a ground-up process that utilised attribute naming and rating tasks – were found to be significantly linked to gender stereotypes, and offered strong evidence for the existence of occupational gender stereotypes best categorised as unfeminine and unmasculine. They also indicated that masculine occupations generate richer mental representations than feminine ones. The results of word association tasks were in keeping with these findings,

indicating that the activation of mental representations of women and men is easier in relation to congruently gender stereotyped occupations than to incongruently gender stereotyped occupations. In relation to the possible effects of grammatical gender on occupational gender stereotyped beliefs, the results offered support for the 'stereotype salience' hypothesis, with semi-gendered language speakers relying the least, and fully gendered language speakers relying the most, on occupational gender stereotypes to guide social perception.

In terms of implications for research examining stereotype interactions on a more general level, the results of this thesis support the use of word association, attribute naming, and attribute rating paradigms for exploratory purposes. Further, the results strongly offer support for the use of the internet-based instrument PsyToolkit for complex psycholinguistic research. The findings also offer support for the idea of counter-stereotype categories, whether they be independent of or a small part of the stereotype categories to which they are related. The clarity with which they appeared in the results of Paper II suggest that, when properly designed, attribute naming and rating experiments examining stereotype interactions may be a key manner in which these categories can be defined and explored. The slight 'feminine bias' found in the results of Paper III has some very interesting ramifications, as it suggest that, when external forces cause shared cultural beliefs to shift rapidly, the increased effect of flexibility and decreased effect of generational knowledge transmission can lead to violations of the expectations underpinning social role theory.

In conclusion, in exploring the interplay between grammatical gender and occupational gender stereotypes, this thesis has successfully reached its aim of expanding knowledge related to the interplay between linguistic factors and stereotype beliefs.

## 6. References


Abele, A. E., (2003) The dynamics of masculine-agentic and feminine-communal traits: findings from a prospective study. *Journal of Personality and Social Psychology, 85*(4): 768-776

Abele, A. E., Uchronksi, M., Suitner, C., & Wojciszke, B., (2008) Towards an operationalisation of the fundamental dimensions of agency and communion: trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology, 38*: 1202-1217

Adachi, T., (2013) Occupational gender stereotypes: is the ratio of women to men a powerful determinant? *Psychological Reports: Sociocultural Issues in Psychology, 112*(2): 640-650

Aboud, F. E., & Doyle, A. -B., (1996) Parental and peer influences on children's racial attitudes. *International Journal of Intercultural Relations, 20*: 371-383

An, S., (2013) Schema theory in reading. *Theory and Practice in Language Studies, 3*(1): 130-134

Anderson, R. C., & Pearson, P. D., (1984) A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 255-291). New York, NY: Longman

Axelrod, R., (1973) Schema theory: an information processing model of perception and cognition. *The American Political Science Review, 67*(4): 1248-1266

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*: 12–28.

Bakan, D., (1966) *The duality of human existence: an essay on psychology and religion*. Chicago, IL: Rand McNally

Bates, D., Maechler, M., Bolker, B., & Walker, S., (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1): 1-48

Beatty-Martinez, A. L., & Dussias, P. E., (2019) Revisiting masculine and feminine grammatical gender in Spanish: linguistic, psycholinguistic, and neurolinguistic evidence. *Frontiers in Psychology, 10*: 751

Bem, S. L., (1974) The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*(2): 155-162

Berki, E., Payton, F., & Pinto, N., (2015) Bind and double bind: media images of women in computing. *Twenty-first Americas Conference on Information Systems*. Fajardo, Puerto Rico.

Boroditsky, L., (2011) How language shapes thought: The languages we speak affect our perceptions of the world. *Scientific American, 2*: 63-65

Braun, F., Oakhill, J., & Garnham, A., (2011) *The language gender index.* Paper presented at the Language, Social Roles, and Behaviour LCG-ITN Summer School, Berlin, Germany.

Brown, R., (1995) *Prejudice: its social psychology.* Cambridge, MA: Blackwell.

Burgess, D. J., & Borgida, E., (1999) who women are, who women should be: descriptive and prescriptive gender stereotyping in sex discrimination. Psychology, Public Policy, and Law, 5(3): 665-692

Carreiras, M., Garnham, A., Oakhill, J., & Cain, K., (1996) The use of stereotypical gender information in constructing a mental model: evidence from English and Spanish. *The Quarterly Journal of Experimental Psychology, 49A*(3): 639-663

Carrel, P., & Eisterhold, J. C., (1983) Schema theory and ESL reading pedagogy. *TESOL Quarterly, 17*: 553-573

Chen, S. X., & Bond, M. H., (2010) Two languages, two personalities? Examining language effects on the expression of personality in a bilingual context. *Personality and Social Psychology Bulletin, 36*(11): 1514-1528

Chen, S. X., (2015) Towards a social psychology of bilingualism and biculturalism. *Asian Journal of Social Psychology, 18*(1): 1-11

Coyle, E. F., & Liben, L. S., (2015) Affecting girls' activity and job interests through play: the moderating roles of personal gender salience and game characteristics. *Child Development, 87*(2): 414-428

Deaux, K., & LaFrance, M., (1998) Gender. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol.1, pp.788-827). New York, NY: McGraw-Hill

Eagly, A. H., (1987) *Sex differences in social behavior: a social-role interpretation*. Hillsdale, NJ: Lawrence Eribaum.

Eagly, A. H., Nater, C., Miller D. I., Kaufmann, M., & Sczesny, S., (2019) Gender stereotypes have changed: a cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist,* Advance online publication. DOI: 10.1037/amp0000494

Eagly, A. H., & Wood, W., (2016) Social role theory of sex differences. In N. Naples et al. (Eds.) *The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies* (pp. 458-476). Singapore: John Wiley and Sons

Egan, S. K., & Perry, D. G., (2001) Gender identity: a multidimensional analysis with implications for psychosocial adjustment. *Development Psychology, 37*: 451-463

Fausey, C. M., & Boroditsky, L., (2010) Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic Bulletin & Review, 17*(5): 644-650

Fausey, C. M., & Boroditsky, L., (2011) Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review, 18*(1): 150-157

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J., (2002) A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6): 878-902

Fiske, S. T., & Stevens, L. E. (1993). What's so special about sex? Gender stereotyping and discrimination. In S. Oskamp & M. Costanzo (Eds.), *Claremont Symposium on Applied*

*Social Psychology, Vol. 6. Gender issues in contemporary society* (pp. 173-196). Thousand Oaks, CA: Sage Publications, Inc.

Fox, J., (2003) Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*(15): 1-27

Gabriel, U., & Gygax, P., (2008) Can societal language amendments change gender representation? The case of Norway. *Scandinavian Journal of Psychology, 49*(5): 451-457

Gabriel, U. & Gygax, P. (2016). Gender and linguistic sexism. In H. Giles & A. Maas (Eds). *Advances in Intergroup Communication* (pp. 177-192). New York, NY: Peter Lang.

Gabriel, U., Gygax, P., Sarrasin, O., Garnham, A., & Oakhill, J., (2008) Au-pairs are rarely male: norms on the gender perception of role names across English, French, and German. *Behaviour Research Methods, 40*: 206-212

Garnham, A., Doehren, S. D., & Gygax, P. M., (2015) True gender ratios and stereotype rating norms. *Frontiers in Psychology – Cognition, 6*(1023). http://doi.org/10.3389/fpsyg.2015.01023

Garnham, A., Gabriel, U., Sarrasin, O., Gygax, P., & Oakhill, J., (2012) Gender representation in different languages and grammatical marking on pronouns: when beauticians, musicians, and mechanics remain men. *Discourse Processes, 49*: 481-500

Glick, P., Wilk, K., & Perreault, M., (1995) Images of occupations: components of gender and status in occupational stereotypes. *Sex Roles, 32*(9/10): 565-582

Goclowska, M. A., Baas, M., Crisp, R. J., & De Dreu, C. K., W., (2014) Whether social schema violations help or hurt creativity depends on need for structure. *Personality and Social Psychology Bulletin, 40*(8), 959-971

Gorfein, D. S., (Ed.) (2001) *On the consequences of meaning selection: perspectives on resolving lexical ambiguity*. Washington, DC: American Psychological Association.

Gorfein, D. S., & Bubka, A., (1989) *Resolving semantic ambiguity*. New York, NY: Springer.

Gorfein, D. S., Brown, V. R., & DeBiasi, C., (2007) The activation-selection model of meaning: explaining why the son comes out after the sun. *Memory & Cognition, 35*(8): 1986-2000

Gygax, P., Elmiger, D., Zufferey, S., Garnham, A., Sczesny, S., von Stockhausen, L., Braun, F., & Oakhill, J., (2019) A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in Psychology, 10*:1604

Gygax, P., & Gabriel, U., (2008) Can a group of musicians be composed of women? Generic interpretation of French masculine role names in the absence and presence of feminine forms. *Swiss Journal of Psychology, 67*: 143-151

Gygax, P., Gabriel, U., Sarrasin, O., Garnham, A., & Oakhill, J., (2008) Some grammatical rules are more difficult than others: the case of the generic interpretation of the masculine. European Journal of Psychology of Education, 24: 235-246

Gygax, P., Gabriel, U., Lévy, A., Pool, E., Grivel, M., & Pedrazzini, E., (2012) The masculine form and its competing interpretations in French: when linking grammatically masculine role names to female referents is difficult. *Journal of Cognitive Psychology, 24*(4): 395-408

Güngör, D., Bornstein, M. H., De Leersnyder, J., Cote, L., Ceulemans, E., & Mesquita, B., (2012) Acculturation of personality: a three-culture study of japanese, japanese americans, and european americans. *Journal of Cross-Cultural Psychology, 44*(5): 701-718

Hamilton, D. L., (1981) Illusory correlation as a basis for stereotyping. In D. L. Hamilton (Ed.) *Cognitive processes in stereotyping and intergroup behavior*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Harrison, S. D., Welch, G. F., & Adler, A., (2012) *Perspectives on males and singing.* New York, NY: Springer

Haslam, S. A., Reicher, S. D., & Platow, M. J., (2011) *The new psychology of leadership: identity, influence, and power*. New York, NY: Psychology Press.

Hogg, M. A., & Reid, S. A., (2006) Social identity, self-categorisation, and the communication of group norms. *Communication Theory, 16*: 7-30

Imhoff, R., Koch, A., & Flade, F., (2018) (Pre)occupations: a data-driven model of jobs and its consequences for categorization and evaluation. *Journal of Experimental Social Psychology, 77*: 76-88

Irmen, L., (2007) What's in a (role) name? formal and conceptual aspects of comprehending personal nouns. *Journal of Psycholinguistic Research, 36*: 431-456

Jaeger, T. F., & Weatherholtz, K., (2016) What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology, 7*: 1115

Kang, S. K., Chasteen, A. L., Cadieux, J., Cary, L. A., & Syeda, M. (2014). Comparing young and older adults' perceptions of conflicting stereotypes and multiply-categorizable individuals. *Psychology and Aging, 29*(3), 469-48

Kennison, S. M., & Trofe, J. L., (2003) Comprehending pronouns: a role for word-specific gender stereotype information. *Journal of Psycholinguistic Research, 32*(3): 355-378

Kessels, U., Rau, M., & Hannover, B., (2006) What goes well with physics? measuring and altering the image of science. *British Journal of Educational Psychology, 76*(4): 761-780

Krauss, R. M., & Chiu, C.-y., (1997) Language and social behavior. In Gilbert, D., Fiske, S., & Lindsey, G., (Eds.) *Handbook of social psychology, vol. 2.* (pp. 41-88) Boston: McGraw-Hill

Koenig, A. M., (2018) Comparing prescriptive and descriptive gender stereotypes about children, adults, and the elderly. *Frontiers in Psychology, 9*: 1086

Koivula, R., (1995) Ratings of gender appropriateness of sports participation: effects of gender-based schematic processing. *Sex Roles, 33*: 543-557

Koivula, R., (2001) Perceived characteristics of sports categorized as gender-neutral, feminine, and masculine. *Journal of Sport Behaviour*, 24(4): 377-393

Kulik, L. (1999) The impact of evaluation procedure on occupational sex-typing at different educational levels. *Journal of Career Assessment, 7*: 415-427.

Kumar, S., Shaw, P., Giagkos, A., Braud, R., Lee, M., & Shen, Q., (2018) Developing hierarchical schemas and building schema chains through practice play behavior. *Frontiers in Neurorobotics, 12*: 33

Lee, Y. T., Jussim, L. J., & McCauley, C. R., (1995) *Stereotype accuracy: toward appreciating group differences.* Washington, DC: American Psychological Association.

Lévy, A., Gygax, P., & Gabriel, U., (2014) Fostering the generic interpretation of grammatically masculine forms: when my aunt could be one of the mechanics. *Journal of Cognitive Psychology, 26*: 27-38

Lindsay, P. H., & Norman, D. A., (1972) *Human Information Processing: and introduction to psychology.* New York, NY: Academic Press Inc.

Little, K. B. (1968). Cultural variations in social schemata. *Journal of Personality and Social Psychology, 10*(1): 1-7.

Locksley, A., Hepburn, C., & Ortiz, V., (1982) Social stereotypes and judgements of individuals: an instance of the base-rate fallacy. *Journal of Experimental Social Psychology, 18*(1): 23-42

Loftus, E. F., & Palmer, J. C., (1974) Reconstruction of automobile destruction: an example of the interaction between language and memory. *Journal of Verbal Learning and Behavior, 13*: 585-589

Lüdecke, D., (2018) *Sjstats: statistical functions for regression models*. Retrieved from https://CRAN.R-project.org/package=sjstats

Martin, P. Y., (1996) Gendering and evaluating dynamics: men, masculinities and management. In D. L. Collinson & J. Hearn (Eds.), *Men as manangers, managers as men: critical perspectives on men, masculinities, and management* (pp.197-214). Toronto: Oxford University Press.

Medin, D. L., & Ross, B. H., (1992) *Cognitive Psychology.* San Diega, CA: Harcourt Brace Jovanovich

Misersky, J., Gygax, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., Chiarini, T., Englund, K., Hanulikova, A., Øttl, A., Valdrova, J., Von Stockhausen, L., & Sczesny, S., (2014) Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavioural Research Methods 46*(3): 841-871

Oakenfull, G., (2013) Unraveling the movement from the marketplace: lesbian responses to gay-oriented advertising. *Journal of Marketing Development and Competitiveness, 7*(2): 57-71

Paap, K. R., (1975) Theories of speech perception. In D. W. Massaro (Ed.) *Understanding Language* (pp. 151-206). New York, NY: Academic Press

Paas, F., Renkl, A., & Sweller, J., (2003) Cognitive load theory and instructional design: recent developments. *Educational Psychologist, 38*(1): 1-4

Phillips, I., (2018) Cross-linguistic structural priming in heritage Spanish speakers: the effects of exposure to English on the processing of preposition stranding in English. In A. B. Bertolini & M. J. Kaplan (Eds.) *Proceedings of the 42nd annual Boston University Conference on Language Development* (pp. 618-631). Somerville, MA: Cascadilla Press

Prentice, D. A., & Carranza, E., (2002) What women and men should be, shouldn't be, are allowed to be, and don't have to be: the contents of prescriptive gender stereotypes. *Psychology of Women Quarterly, 26*(4): 269-281

Proctor, R. W., & Vu, K-P. L., (2012) Human Information Processing. In N. M. Seel (Ed.) *Encyclopedia of the Sciences of Learning*. Boston, MA: Springer

Pyykkönen, P., Hyönä, J., & van Gompel, R. P. G., (2010) Activating gender stereotypes during online spoken language processing: evidence from visual world eye tracking. *Experimental Psychology, 57*(2): 126-133

Rácz, P. (2012). Operationalising salience: Definite article reduction in the North of England. *English Language & Linguistics*, *16*: 57–79

Rácz, P. (2013). *Salience in Sociolinguistics.* New York, NY: De Gruyter Mouton.

Reimers, S. & Stewart, N., (2007) Adobe flash as a medium for online experimentation: a test of reaction time measurement capabilities. *Behaviour Research Methods, 39*(3): 365-370

Riener, A., (2017) Subliminal perception or "can we perceive and be influenced by stimuli that does not reach us on a conscious level". In Myounghoon, J. (Ed.) *Emotionas and affect in human factors and human-computer interaction.* (pp. 503-538). Cambridge, MA: Academic Press

Rodríguez-Arauz, G., Ramírez-Esparza, N., Pérez-Brena, N., & Boyd, R. L., (2017) Hablo Inglés y Español: Cultural Self-Schemas as a Function of Language. *Frontiers in Psychology*, 8: 885

Rumelhart, D. E., (1980) Schemata: the building blocks of cognition. In R. J. Spiro et al. (Eds.) *Theoretical Issues in Reading Comprehension.* Hillsdale, NJ: Lawrence Erlbaum

Rydell, R.J., McConnel, A.R., & Beilock, S.L., (2009) Multiple social identities and stereotype threat: imbalance, accessibility, and working memory. *Journal of Personality and Social Psychology, 96*(5): 949-966

Schubert, T. W., Murteira, C., Collins, E. C., & Lopes, D., (2013) ScriptingRT : a software library for collecting response latencies in online studies of cognition. *PLoS One, 8*(6): e67769.

Seta, J. J., Seta, C. E., & McElroy, T. (2003). Attributional biases in the service of stereotype maintenance: A schema-maintenance through compensation analysis. *Personality and Social Psychology Bulletin, 29*(2), 151–163.

Shaw, S., & Hoeber, L., (2003) A strong man is direct and a direct woman is a bitch: gendered discourses and their influence on employment roles in sport organisations. *Journal of Sport Management, 17*: 347-375

Siyanova-Chanturia, A., Warren, P., Pesciarelli, F., & Cacciari, C., (2015) Gender stereotypes across the ages: on-line processing in school-age children, young and older adults. *Frontiers in Psychology, 6*: 1388

Slobin, D. I., (2002) Cognitive and communicative consequences of linguistic diversity. In S. Strömqvist (Ed.), *The diversity of languages and language learning* (pp. 7-23). Lund, Sweden: Lund University, Centre for Languages and Literature

Spence, J. T., Helmreich, R., & Stapp, J., (1975) Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology, 32*: 29-39

Stangor, C., & Schaller, M., (2000) Stereotypes as individual and collective representations. In C. Stangor (Ed.), *Stereotypes and prejudice* (pp. 64-82). Philadelphia, PA: Psychology Press

Swim, J. K., (1994) Perceived versus meta-analytic effect sizes: an assessment of the accuracy of gender stereotypes. *Journal of Personality and Social Psychology, 66*(1): 21-36

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., Bito, H., & Morris, R. G. M., (2011) Schema-dependent gene activation and memory encoding in neocortex. *Science, 333*(6044): 891-895

Wayne, J. H., & Cordeiro, B. L., (2003) Who is a good organizational citizen? Social perception of male and female employees who use family leave. *Sex Roles, 49*(5/6): 233-246

Wigboldus, D., & Douglas, K., (2007) Language, stereotypes, and intergroup relations. In K. Fielder (Ed.) *Social Communication* (pp. 79-107). New York, NY: Psychology Press

Wolfram, H-J., & Mohr, G., (2010) Gender-typicality of economic sectors and gender-composition of working groups as moderating variables in leadership research. *Gender in Management: An International Journal, 25*(4): 320-339

Zajdel, R., & Nowak, D., (2007) Simple and complex reaction time measurement. A preliminary evaluation of new approach and diagnostic tool. *Computers in Biology and Medicine, 37*: 1724-1730

## 7. Compilation of Papers

### 7.1. Paper I: Testing the effectiveness of the Internet-based instrument PsyToolkit: A comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task.

**Authors:** Jonathan D. Kim, Ute Gabriel, and Pascal Gygax.

**Status:** Published in PLoS One, Volume 14, Number 9; DOI: e0221802

**Testing the effectiveness of the Internet-based instrument PsyToolkit: A comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task**

*Jonathan D. Kim, Ute Gabriel, & Pascal Gygax*

**Abstract**

To test the effectiveness of the Internet-based instrument PsyToolkit for use with complex choice tasks, a replicability study was conducted wherein an existing psycholinguistic paradigm was utilised to compare results obtained through the Internet-based implementation of PsyToolkit with those obtained through the laboratory-based implementation of E-Prime 3.0. The results indicated that PsyToolkit is a viable method for conducting both general and psycholinguistic specific experiments that utilise complex response time tasks, with effects found to replicate for both response choice *and* response time.

## Introduction

The advent of the Internet opened new avenues of exploration for us as psychological researchers. Internet-based experimental instruments allow us to conduct experiments with demographically and culturally diverse samples, to recruit large subject pools in less time, to avoid organisational issues such as scheduling conflicts, to save costs related to laboratory space, equipment, personnel hours, and administration, and to increase our ability to conduct international experiments (Krantz & Reeshad, 2000; Reips, 2000; Reips, 2002). For these benefits to be worthwhile we must be able to trust Internet-based instruments to accurately record participants' responses, in terms of both the actual responses as well as their intrinsic characteristics, such as response times. The current study investigates this issue by testing the replicability of the Internet-based implementation of PsyToolkit for use with paradigms requiring complex Choice Response Time (CRT) tasks.

It has been argued that, for instruments found to reliably record participants' responses, Internet-based experimentation has three main advantages over laboratory-based experimentation (Reips, 2002); increased generalisability, increased voluntariness, and increased ecological validity. Increased generalisability refers to participants being able to be recruited from much broader demographic and/or geographic backgrounds, meaning that the sample is more likely to be truly representative of society. Increased voluntariness refers to participants having fewer constraints on their decisions to participate and to continue to participate as, for example, there is no researcher whose presence might socially pressure a participant to continue. Further, responses may be more authentic when participants are more comfortable in their ability to stop the experiment (Reips, 2002). Ecological validity is a measure of the level to which participant behaviour in an experiment resembles their behaviour in a naturalistic setting. The closer to reality an experiment can be, the higher the level of ecological validity the experiment is said to have, and the more we can be confident that the results obtained reflect the participant's real-world behaviours. As an example, driving simulators attempt to simulate, to different degrees, the feeling of driving a real car. The closer the simulator is to the experience of naturalistically driving a car, the higher the level of ecological validity. As such, an experiment in which you sit inside an actual car, observe a scene projected on the wall in front and to the sides of you, and respond using the car's steering wheel, accelerator, and break is likely to have a higher level of ecological validity than an experiment in which you sit in front of a computer screen, observe a scene shown on the screen, and respond using controllers shaped like a steering wheel, accelerator, and break, which in

turn is likely to have a higher level of ecological validity than an experiment in which you sit in front of a computer screen, observe a scene shown on the screen, and respond by moving the mouse on the screen to control direction and speed. With reference to internet-based studies, it has been argued that the ability for participants to take part in experiments in environments (and using equipment) that they are familiar with, and the ability for participants to undertake experiments without the presence of a researcher in the room, lead to increased ecological validity (Reips, 2002).

The ability to undertake experiments in familiar environments, and with familiar equipment, has the potential to enhance ecological validity in at least two manners; increased familiarity and reduced cognitive load. Increased familiarity refers to the fact that participants can choose the time, place, and surroundings in which to undertake the experiment, ensuring that any effects found cannot be attributed to being in an unfamiliar setting (Reips, 2002). Cognitive load refers to the amount of cognitive resources required, out of a limited pool, to fulfil the requirements of mentally demanding tasks (King & Bruner, 2000). In experimental terms, increasing levels of cognitive load are associated with increased reaction times, as participants have less cognitive resources available for undertaking experimental tasks. Unfamiliar environmental factors are known to increase cognitive load, as the level to which the brain actively monitors the environment is higher, which in turn reduces the cognitive resources available for other tasks. As such, the more familiar an individual is with their surroundings, the less cognitive resources are utilised in monitoring the environment, meaning that there are more cognitive resources available for focusing on the experimental task with which they are presented.

The lack of a researcher present has the potential to enhance ecological validity through reduced social desirability bias and reduced cognitive load. Social desirability bias refers to a cognitive bias in which individuals act to increase the level to which answers they give are in line with social norms in order to present themselves in the best possible light (Fisher, 1993; King & Bruner, 2000). The level to which this bias occurs is, among other factors, heightened in the presence of others (Holbrook, Green, & Krosnick). As such, responses given in the absence of researchers are more likely indicative of how an individual truly feels about the subject, leading to higher ecological validity. Cognitive load is also reduced in the absence of a researcher, as the presence of others when undertaking a task divides attention, at least to some degree, between the experimental task and anyone else present (Nicholson, Parboteeah, Nicholson, & Valacich, 2005; Sanders, Baron, & Moore, 1978).

While increasing ecological validity is an important factor for experimental design,

laboratory-based experiments also have advantages over internet-based experiments. Firstly, laboratory-based experiments have a higher range of possible research approaches. This is primarily due to equipment requirements. It is not reasonable, for example, to expect participants recruited from the general populace to all own eye tracking equipment; as such, it is more logical to undertake experiments in which eye tracking is included in laboratory conditions. Further, hardware and software related issues have historically introduced a high level of error noise into results obtained through internet-based instruments compared to those obtained through laboratory-based instruments, primarily observable as response time noise. A wide variety of factors can affect response time recording, such as hardware timing features, device driver issues and interactions, script errors, operating system variability, interactions with other software, tools to construct the paradigm, interactions with other hardware, and configuration of settings and levels (Krantz & Reeshad, 2000; Plant, 2009). In laboratory-based experiments these sources of noise are less likely to affect the final results of the experiment, as all participants undertake the experiment with the same hardware, software, device drivers, operating system, and system configuration. In internet-based experiments, however, there are large potential differences in these elements between participants' computers, which can lead to a higher level of noise within the results obtained. Further, responses given via the internet are also affected by the amount of time it takes for the website hosting the experiment to successfully send an image to the participants' computer, and then, after responding, by the amount of time it takes for the response to be sent from the participants' computer to the website hosting the experiment (Høiland-Jørgensen, Ahlgren, Hurtig, & Brunstrom, 2016). As high noise levels can obscure small effects and give the illusion of heterogeneous responses, care must be taken when analysing results obtained through internet-based instruments to ensure that an increase in heterogeneous responses are due to ecological validity improving rather than noise level increasing. However, technology continues to evolve, and recent advances in the design of internet-based experimental tools – such as has occurred with PsyToolkit, the instrument we present next – may have significantly reduced error noise compared to older internet-based instruments, even to the point of bringing them fully in line with laboratory-based instruments. As such, for instruments with which there is minimal Internet-related noise, if ecological validity was indeed increased we could (for example) expect participants to respond to items in a less self-monitored and/or socially accepted manner, with participants displaying wider response choice variability and overall faster response times.

PsyToolkit is an open-access psychological instrument developed to allow researchers, including student researchers, to easily program and run experimental psychological

experiments and surveys for both laboratory and Internet settings (Stoet, 2010; Stoet, 2017). Two versions of PsyToolkit are available; a laboratory-based version that runs on Linux, and an Internet-based version that is Javascript based and can run on modern browsers without participants needing to download any programs. The Internet-based version of the instrument is specifically aimed at addressing financial and technical limitations commonly faced by students, as it is free software that has specifically been designed for running online questionnaires, Simple Response Time (SRT) tasks, and Choice Response Time tasks (CRT) (Stoet, 2017). An SRT is an experimental task in which a single stimulus, and only that stimulus, is presented repeatedly at the same on-screen location, with participants tasked with responding to every presentation of the stimulus in the exact same manner and quickly as possible (Zajdel & Nowak, 2007). An example of this is participants being instructed to watch an LED and to press a specific button as quickly as possible whenever the LED lights up. A CRT is an experimental task in which instead multiple stimuli are shown, and/or stimuli are presented on different areas of the screen, and the participant is tasked with responding in different manners depending on the nature of each presentation (e.g., Zajdel and Nowak, 2007). An example of this is participants being instructed to look at a screen on which letters will appear, with the task of pressing the corresponding letter on a keyboard. CRTs can also differ in complexity. Simple CRTs, such as in the above example, require participants to recognise the stimuli and respond accordingly. More complex CRTs require participants to also make judgements about the nature of the stimuli.

In the present experiment, participants were instructed to look at a screen on which first names paired with role nouns appeared, with the task of pressing one of two buttons depending on whether they believed that it made logical sense for an individual with the name shown to hold the role shown. Stoet (2017) states that PsyToolkit is designed for a teaching environment, with minimal technical barriers and free web-based hosting of their studies. A library of existing psychological scales and experiments is available for students to examine and adapt, and extensive online documentation and tutorials are available to assist if students face any issues. Further, Stoet (2017) states that PsyToolkit is designed to allow for students to randomise item order in both questionnaires and in cognitive experiments, to allow for a convenient way of scoring, and to give feedback to participants about their test scores; options not available in all Internet-based instruments. All users of the Internet-based version must register an account to be able to create experiments, but accounts are free. Randomisation is possible in both the survey and the experiment, and partial randomisation is also possible for if one wishes for only certain portions of the survey and/or experiment to be randomised. Further, alternate versions

of the experiment can be created, with participants randomly assigned between versions. In terms of reliability, Stoet (2018) states that both the Internet and Linux versions of PsyToolkit can reliably measure small effects of less than 50ms, with the Linux version being more precise. However, to our knowledge, currently no research has been published examining the replicability of the Internet-based version of PsyToolkit.

As PsyToolkit is intended to be a student-focused instrument, and many universities do not set up experimental computers with Linux for their students, it was decided to compare results obtained through the Internet-based implementation of PsyToolkit to results obtained through E-Prime 3.0 in a laboratory setting. E-Prime was chosen as it is a commonly used psychological research tool in university settings, including in teaching environments, and, like PsyToolkit, it has a low barrier to entry and has an experiment library. Further, Stoet (2018) states that the Linux-based version of PsyToolkit is on par with E-Prime, so, while there is likely to be noise due to differences in software, this is expected to be minimal.

While the replicability of PsyToolkit has not been examined, the replicability of other Internet-based instruments has been tested through CRT tasks (e.g., Reimers and Stewart, 2007; Schubert, Murteira, Collins, & Lopes 2013). Reimers and Stewart (2007) used a CRT task to test the replicability of an experiment in the Internet-based version of Adobe Flash compared to the same experiment in a laboratory-based version of Adobe Flash, with the same experiment coded in C used as a baseline. Participants were shown green and red rectangles and were required to press buttons corresponding to the colour of the rectangle on the screen. They found that, compared to the baseline, (a) response times of the laboratory-based version of Flash were 10ms longer, (b) response times of the Internet-based version of Flash were 30-40ms longer, and (c) there were no significant differences in Response Time standard errors across conditions. Schubert et al. (2013) used both SRT and CRT experiments in a study testing the replicability of ScriptingRT to Flash. Six experiments were conducted over the course of their study. The first three studies used SRT tasks but were automated to test specific aspects of ScriptingRT. The last three studies used CRT tasks, specifically a version of the Stroop task, where participants were presented with either the words “red” or “blue”, or a neutral letter string, in either red or blue on a white background. Participants were instructed to press keys corresponding to the colour of the word or neutral letter string shown. Experiment 4 tested the Internet-based version of ScriptingRT by itself, while Experiment 5 compared ScriptingRT to the same experiment coded in DMDX (a laboratory-based instrument [Forster & Forster, 2003]) with participants undertaking both tasks on the same computer, and Experiment 6 compared ScriptingRT to Inquisit Web Edition, both running via the Internet. In Experiment 5, the

experiment of most interest to the present experiment as it compares an Internet-based implementation of the instrument to a laboratory-based one, Schubert et al. (2013) found that the size of the Stroop effect was not affected by which software was used.

Historically, psycholinguistic research has not relied upon Internet based testing, as it often relies upon small differences in response times in CRT tasks to detect effects and is strongly affected by response time noise. Recent research (e.g., Enochson & Culbertson, 2015) has found that some modern Internet-based instruments are reliably able to test these small differences, meaning that modern psycholinguistic research may safely utilise Internet based tools that have been properly validated. Some researchers have suggested that PsyToolkit may be a delicate enough tool for psycholinguistic experimentation (e.g., Sampaio, 2017). An opportunity arises therefore to test both general replicability and psycholinguistic specific replicability of PsyToolkit through a psycholinguistic experimental paradigm.

The present study was designed to compare responses and Response Times measured by the Internet-based implementation of PsyToolkit with those measured by the laboratory-based implementation of E-Prime 3.0 using a complex CRT task composed of an existing and published psycholinguistic paradigm (i.e., Gygax & Gabriel, 2008) to test replicability between the Internet-based implementation of PsyToolkit (Version 2.4.3) and E-Prime (Version 3.0.3.31). The paradigm uses a between-subjects two-alternative forced choice design, with a CRT task in which participants are shown pairs of terms (in the present experiment a first name and a role noun; e.g., 'Kate – Chefs') and are then required to, as quickly as possible, make a judgement as to whether the pairing makes logical sense (i.e., could an individual named Kate be a member of a group of chefs). Experimental item pairings were composed of first names paired with professional roles that vary in gender stereotypicality. As logically any individual can hold any professional role, filler item pairings were included to prevent participants of developing a strategy of always answering positively to all roles seen. The filler items were first names paired with gender-marked kinship terms, with both congruent (e.g., 'Kate – Mothers') and incongruent (e.g., 'Kate – Fathers') pairings shown to prevent participants from developing a strategy of answering positively to professional roles and negatively to familial roles.

The paradigm we utilise is more complex than those used by Reimers and Stewart (2007) and by Schubert et al. (2013), as the paradigm used in the current study requires participants to make subjective judgements of the items presented before responding, while the paradigms used by Reimers and Stewart (2007) and Schubert et al. (2013) required that participants responded based on the colour, an objective quality, of the items presented to them. One can therefore expect that overall response times will be longer for this study than those

found by Reimers and Stewart (2007) and Schubert et al. (2013), and, compared to Reimers and Stewart (2007), it is likely that Response Time standard errors will be larger. Further, if the results indicate that there is a high level of replicability between PsyToolkit and E-Prime, then it may be possible to determine whether the results offer any support for the concept of increased ecological validity in Internet-based experiments. If the results obtained in PsyToolkit do have a higher level of environmental validity than the results obtained in E-Prime, we would expect that participants who undertake the PsyToolkit version of the experiment would be more likely to respond negatively, and would overall respond more quickly (i.e., more spontaneously), than those who undertake the E-Prime version of the experiment.

It is worth noting that Norwegian is considered a semi-gendered language. This is because some, but not all, nouns have associated gender markers. Specifically, only nouns that refer to living beings, especially humans, are gendered in Norwegian. Further, most role nouns in the plural form are the same as the masculine-specific singular form. This is due in part to a linguistic policy of gender neutralisation (Gabriel & Gygax, 2008), under which the masculine grammatically marked form of role nouns are actively encouraged to become the main linguistic device to refer to most roles (Swan, 1992).

## Method

### Participants

A total of 81 participants took part in this study (39 [18 female, 20 male, 1 nonbinary] through PsyToolkit, 42 [20 female] through E-Prime). Across both versions of the experiment participants were between 19 and 31 years old (M = 23.4; SD = 2.3), were self-reported Norwegian first language speakers, and were currently studying at NTNU, Norway. Participants in both PsyToolkit (Version Web) and the E-Prime (Version Lab) were recruited through posters and flyers placed around the Dragvoll campus at NTNU, and through direct recruitment (i.e., the researchers involved approaching people directly and asking whether they would be willing to take part in the experiment). Those who responded to the advertisements or to the direct recruitment were then either asked to undertake Version Lab at the Dragvoll campus of NTNU or were sent a link to undertake Version Web. Recruitment into both versions occurred concurrently. All participants were compensated through coffee vouchers. Informed consent was obtained from all participants prior to the experiment. This study received approval from the Norwegian Centre for Research Data.

**Materials and research design**

A two-alternative forced choice design was used for both versions of the experiment. All experimental elements were translated into Norwegian. Participants gave informed consent, answered questions on age, gender, and handedness, and stated whether they were currently enrolled university students, before experimental onset. For Version Web this was done through a form on the website hosting the experiment, while for Version Lab this was done in hard copy.

Participants were presented with pairs of terms composed of a first name (e.g., Daniel) and a role noun in the plural form (e.g., Astronauts). Participants were then required to indicate, as quickly as possible, whether an individual named [name] could be a member of the group of [noun]. These pairings were always presented in the form '[name] – [noun]' (e.g., Daniel – Astronauts), with presentation order randomised by participant. Participants in both versions of the experiment responded via a keyboard, and were instructed to press 'e' if they did not agree that the individual could be a member of the group indicated, or 'i' if they did agree. After each answer was given, the pairing was replaced with a fixation cross of 100ms, after which the next pairing was displayed. The lack of a response within 5000ms was recorded as a non-response, after which the experiment would continue. Participants undertook a five-item training phase before undertaking the main experimental phase. Both versions of the experiment took between 20 and 30 minutes to complete.

**Stimuli.**

The stimuli were composed of six first names paired with 36 role nouns and 36 filler items. In total, participants were presented with 360 noun-name pairings, composed of 216 experimental pairings and 144 filler pairings.

The 36 role nouns (12 female stereotyped roles, 12 male stereotyped roles, and 12 non-stereotyped roles; Tables 1-3) were selected based on Misersky et al. (2014). Misersky et al. produced stereotypicality rankings between 0 and 1, with 0 representing male stereotyped roles, 0.5 representing non-stereotyped roles, and 1 representing female stereotyped roles. For this study, the masculine roles selected had a mean rating of .20 (SD = .03), while the feminine roles had a mean rating of .81 (SD = .04), and the non-stereotyped roles had a mean rating of .53 (SD = .06).

Three female (Ida, Nina, Sandra) and three male (Espen, Geir, Robert) names were used to maintain gender balance. These were selected based on the findings of Öttl (Unpub.), who tested typicality of names through a response time experiment. In Öttl's experiment participants were presented with names and were instructed to press a button marked 'female' if they thought the name was female, and 'male' if they thought the name was male. The names used were taken from Statistics Norway, and were selected to represent the most frequent Norwegian names among people born between 1976 and 1996. Lower response times were interpreted as indicating a higher level of gender typicality associated with that name. The typicality of the names selected for the current study was balanced by gender (Table 4). Each name was paired with all role nouns, for a total of 216 experimental pairings.

The 36 filler items were gender-marked kinship terms (e.g., Father, Sister; 18 female gender marked, 18 male gender marked) that were selected to prevent participants developing a strategy of always answering positively. Kinship terms were paired with both incongruent and congruent names so that participants would be unlikely to adopt a strategy of responding positively to all items, but would also be unlikely to adopt a strategy of responding positively to professional roles but negatively to kinship terms. Each name was paired with all the incongruent filler items, for a total of 108 first name – incongruent filler item pairings, and was paired with six of the congruent filler items, for a total of 36 first name – congruent filler item pairings.

**Procedure for Version Web.**

Participants undertook this version of the experiment on their home computers, for which we do not have the specifications. The experiment was run through the PsyToolkit website. Before starting the survey, participants were required to give informed consent through a checkbox on the website. Failure to check this box meant that the survey would not begin. During the first part of the survey, participants answered the demographic questions stated above. After this, a black box was shown on the screen, and participants were instructed to click a button underneath it to start the experiment when they were ready. When this button was pressed, the black box expanded to full-screen mode, and the experiment began. Responses were only saved by PsyToolkit if participants completed the survey and experiment in entirety, with all survey questions needing to be answered before participants could move on. After completing the experiment, participants were presented with a code and were instructed to email the code to the researchers to arrange a time to receive their compensation. As emails constitute

identifying information, the emails were deleted after participants received their compensation. PsyToolkit created two files per participant. The first contained information relating to when they started and ended the experiment, their IP address, and their responses to the demographic questions. The second contained their responses to each of the experimental pairings.

**Procedure for Version Lab.**

Participants undertook this version of the experiment in a laboratory setting in the Psychology Department at NTNU. Before starting the experiment, participants were required to give informed consent, and then to answer demographic questions, through hard-copy forms. After this, participants undertook the experiment. This was presented to them on a screen (1920 x 1080), which was attached to an air-gated Dell Latitude E5470 laptop with an Intel core i7-6820HQ CPU and 16gb RAM, running Windows 10 Education in 64-bit, with a screen refresh rate of 60Hz. The laptop sat facing the researcher, while the screen sat facing the participant. The participant was seated directly opposite the researcher, so that they faced each other but direct line of sight during the experiment was blocked by the screen. The display was mirrored between the laptop and the connected screen. A USB keyboard was attached to the laptop, and placed in front of the participant. After the participant had given informed consent and filled in the demographics questionnaire, the researcher present initiated the experiment. Participants received compensation directly after completing the experiment. E-Prime created two files per participant. These both contained the participant's responses to the experimental pairings, with one being in the .edat3 format, and the other in the .txt format.

**Data Preparation**

For the analysis, demographic information for all participants in each version of the study was compiled into two .txt files, while the experimental data was kept in its uncollated raw form as .txt files for each participant. As IP addresses are identifying information, in order to anonymise the data, they were removed from the demographic information files and deleted prior to data analysis.

Prior to data analysis, both item-by-participant deselection and by-participant data screening were used. Item-by-participant deselection was conducted based on response times. In keeping with standard procedures, such as in Schubert et al. (2013), responses faster than

300ms or not occurring within 5000ms were removed from the data. This represented 0.75% of the data. By-participant data screening was composed of removing participants who (a) were outside of our target demographic group (native Norwegian speaking university students aged between 18 and 35), and (b) removing participants who were found to have an error rate at or above 50%. Error rate by participant was calculated based on the percentage of incorrect answers to all filler items, with the assumption that the correct answer for congruent name – filler item pairings is 'yes' and for incongruent name – filler item pairings is 'no'. One participant was removed for not being a native Norwegian speaker, and seven were removed because their error rate was above 50%. All the participants deselected in this manner took part in Version Lab. The remaining 37 participants who completed Version Web (18 female, 18 male, 1 nonbinary) and 36 participants who completed Version Lab (18 female, 18 male, 0 nonbinary) were used for analysis (N = 72). After deselection, mean participant age was 23.4 (SD = 2.4).

Mean error rate across the study and by version of the experiment was calculated post data screening and deselection. Mean error rate across the study was 11.56%. Mean error rate for Version Web was 10.76%, while mean error rate for Version Lab was 12.38%.

The results were examined through two forms of linear mixed-effects modelling. First, participants' responses (yes/no) were analysed, and second, response times for positive responses were analysed (as in Gygax et al., 2012), both within and between versions of the experiment. Participants' yes/no responses were modelled through generalised linear mixed-effect regression, while participants' response times for positive responses were modelled through linear mixed-effect regression. Analysis was conducted through the *glmer* and *lmer* functions of the lme4 package (Version 1.1-12; Bates, Maechler, Bolker, & Walker, 2015) in R (version 3.3.3). Initial models were defined for both analyses, composed of all experimental factors (Version [Version Web vs. Version Lab], Name Gender [female vs. male], and Stereotype [female vs. male vs. non-stereotyped roles]), their 2-way and 3-way interactions, and random intercepts (Participants, Role Noun, Researcher, and First Name). Researcher refers to which researcher, if any, was present while participants undertook the experiment. All participants who undertook Version Lab did so in the presence of a researcher, while all participants who undertook Version Web did without a researcher present. The models also included fixed effects of Participant Gender, Handedness, Trial Number and Character Count (i.e., how many characters [specifically letters, symbols, and spaces] were in each name – noun pairing). In keeping with Baayen (2008) and Baayen and Milin (2010), refinement to find the model of best fit occurred through back-fitting the fixed effects structure, forward-fitting the

random effects structure (by-participant random slopes for the experimental variables, trial number, and number of characters), then re-back-fitting the fixed effect structure. This was done automatically through the *bfFixefLMER_F*, *ffRanefLMER*, and *fitLMER.fnc* functions of the lme4 package. Post-hoc analysis for main effects and interaction effects was done through the *effects()* function of the effects package (version 4.1-0; Fox, 2003).

Effect size estimates in linear mixed-effects modelling are complicated to determine, with a large variety of practices being utilised in research (Peugh, 2010). To best fit our data, we have selected two methods which address local effect sizes. For this purpose, we utilise the definition of local effect sizes as the effect of individual fixed effect variables on the dependent variable (Peugh, 2010). In line with previous research (Nakagawa & Schielzeth, 2010; Wagner-Egger & Gygax, 2017), estimation of local effect sizes is done through partial omega squared ($\omega_p^2$), obtained through the *omega_sq()* function of the sjstats package (Version 0.17.5; Lüdecke 2018). Further, in keeping with previous research (Wagner-Egger & Gygax, 2017), we present the slopes of the reported effects for each individual level, along with their 95% confidence intervals. The estimation of the slope for the effects was done through the *summary()* function in R, while 95% confidence intervals were calculated through the equation [CI = slope estimate ± (Critical value * Standard error of the slope coefficient)].

## Results

### Response

The AIC value for the initial model was 4885. Version (Web vs. Lab) was automatically removed from the model of best fit during backfitting. However, as this study aims at exploring the Version's impact, Version was nevertheless kept in the final model. The final model for Response contained random intercepts by Role Noun, First Name, Researcher, and Participant, as well as random slopes of Stereotype by Participant, Name Gender by Participant, and Trial Number by Participant. The AIC value for the final model was 4660. Trial Number was found to have a small significant effect on the model (Wald Chi² = 20.41, $p < 0.001$, $\omega_p^2 = 0.005$), with participants increasingly likely to respond positively over time. There was a small yet significant main effect of Name Gender (Wald Chi² = 36.92, $p < 0.001$, $\omega_p^2 = 0.001$), which was qualified by a medium sized and significant two-way interaction between Stereotype and Name Gender (Wald Chi² = 564.08, $p < 0.001$, $\omega_p^2 = 0.114$). There was no significant main effect of Version (Wald Chi² = 0.01, $p = 0.930$, $\omega_p^2 = 0.000$) or of Stereotypicality (Wald Chi² = 1.76, $p$

= 0.410, $\omega_p^2$ = 0.000). No significant two-way interactions were found between Stereotype and Version (Wald Chi² = 0.57, $p$ = 0.750, $\omega_p^2$ = 0.000), or between Name gender and Version (Wald Chi² < 0.001, $p$ = 0.990, $\omega_p^2$ = 0.000). No significant three-way interaction was found between Stereotype, Name Gender, and Version (Wald Chi² = 4.07, $p$ = 0.13, $\omega_p^2$ = 0). Estimates of the slope sizes and confidence intervals for the final model can be found in Table 5.

The interaction between Stereotype and Name Gender (Fig 1, Table 6) indicated no significant differences between conditions, but that participants were, on average, more likely to respond positively to congruent pairings (i.e., pairings where the gender of the first name matched the stereotype of the role noun) than to the incongruent pairings (i.e., pairings where the gender of the first name does not match the stereotype of the role noun) for both male and female names. Participants also tended to respond more positively to names when paired with non-gender-stereotyped roles than with incongruent roles for both female and male names, and tended to respond more positively to non-gender-stereotyped roles when paired with female names compared to male names.

As it was of importance to this study, we will still discuss the non-significant interaction between Stereotype, Name Gender, and Version (Table 7, Fig 2). No significant differences were observed between conditions, and a visual scan of Fig 2 indicates that participants across both versions of the experiment tended to respond more positively to congruent pairings compared to incongruent pairings. Participants who responded to Version Web showed more variation in the responses they gave, resulting in much lower Lower Bound values compared to Version Lab, with the largest differences observed for incongruent role noun pairings.

As Version was automatically removed from the model of best fit, Bayes factors were calculated to examine whether there was support for accepting or rejecting the null hypothesis (i.e., that there was no difference between Version Web and Version Lab). This was done with both the BF_BIC function of the lme4 package (Bates et al., 2015). The comparison models used for this analysis were the final model (stated above) compared to the model of best fit. The model of best fit was identical to the final model aside from the removal of the main effect and interaction effects of Version. The results indicated a Bayes factor of > 0.001, indicating that we can confidently accept the null hypothesis for Response.

**Response Time**

The REML criterion at convergence for the initial model was 212699. Two- and three-way effects involving Version (i.e., Web vs. Lab) were automatically removed from the model of best fit during backfitting, although, unlike with Response, the main effect of Version was kept in the model of best fit. As with Response, the two- and three-way effects involving Version were kept in the final model due to their importance in this study. In order to correct for outlier responses, the final model excluded responses that were more than 2.5 standard deviations from the mean. The final model for Response Time contained random intercepts by Role Noun, First Name, Researcher, and Participant, as well as random slopes of Character Count by Participant and Name Gender by Participant. The REML criterion at convergence for the final model was 199660. Trial Number was found to have a large and significant effect on the model, $F(1, 13664) = 156.19$, $p < 0.001$, $\omega_p^2 = 0.105$, with participants responding increasingly quickly over the length of the experiment. Character Count was also found to have a small yet significant effect on the model, $F(1, 13664) = 29.10$, $p < 0.001$, $\omega_p^2 = 0.002$, with participants responding increasingly slower as character count increased. A small but significant main effect of Stereotype, $F(1, 13664) = 4.56$, $p = 0.01$, $\omega_p^2 = 0.001$ was found, as well as a very small but significant main effect of Name Gender, $F(1, 13664) = 4.90$, $p = 0.03$, $\omega_p^2 < 0.001$, which were qualified by a small yet significant two-way interaction between Stereotype and Name Gender, $F(1, 13664) = 20.87$, $p < 0.001$, $\omega_p^2 = 0.003$. A significant main effect of Study was also found, $F(1, 13664) = 28.91$, $p < 0.001$, $\omega_p^2 = 0.002$, but no significant two-way interactions were found between Stereotype and Version, $F(1, 13664) = 0.43$, $p = 0.653$, $\omega_p^2 < 0.001$, or between Name Gender and Version, $F(1, 13664) = 0.36$, $p = 0.547$, $\omega_p^2 < 0.001$, and no significant three-way interaction was found between Stereotype, Name Gender, and Version, $F(1, 13664) = 0.14$, $p = 0.867$, $\omega_p^2 < 0.001$. Estimates of the slope sizes and confidence intervals for the final model can be found in Table 8.

The main effect of Version (Fig. 3) indicated no significant differences between conditions, but that participants who responded to Version Web (mean response time = 985ms) responded faster on average than participants who responded to Version Lab (mean response time = 1148ms; mean difference = 163ms, 95%CI [-67ms to 396ms]).

The interaction between Stereotype and Name Gender (Fig 4, Table 9) indicated no significant differences between conditions, but that participants, on average, responded more quickly to the congruent roles compared to incongruent roles for both female and male names. Participants also tended to be slower to answer positively to male names paired with feminine stereotyped roles than to female names paired with masculine stereotyped roles.

Again, as it was of importance to this study, we examine still the non-significant interaction between Stereotype, Name Gender, and Version (Table 10, Fig 5). No significant differences between conditions, but there was an overall tendency for participants in Version Web to respond faster than participants in Version Lab, as well as larger response time differences between congruent and incongruent pairings for participants in Version Lab compared to Version Web There was a decrease in mean standard error between Version Web (mean *SE* = 54) and Version Lab (mean *SE* = 64).

## Discussion

The aim of this study was to test the replicability of the Internet-based instrument PsyToolkit when compared to the laboratory-based implementation of E-Prime 3.0 for use with complex choice experiments through a psycholinguistic paradigm. Both PsyToolkit and E-Prime are psychological testing tools that are designed to be easy to use by students, having a low barrier to entry with coding requirements, and having extensive libraries of experiments. PsyToolkit was run on participants' personal Internet-connected computers outside of laboratory conditions, while E-Prime was run on a single air-gated computer inside of laboratory conditions. The results of this study supported a high level of replicability between PsyToolkit and E-Prime, with Bayes factors indicating that we can accept the null hypothesis of no difference between Versions for Response. A secondary aim of this study was to examine the possibility that Internet-based experimentation might have a higher level of ecological validity than laboratory-based experimentation. It is possible that the ability to undertake experiments in familiar surroundings, and in the absence of researchers, could lead to participants being more comfortable while responding. If so, it would be expected that participants would be less affected by, for example, social desirability bias, meaning that their results should be more in line with how they would react in a naturalistic setting. If this was indeed the case, it would follow that participants who undertook the PsyToolkit version of the experiment would be more comfortable in responding negatively, and would overall respond more quickly, than those who undertook the E-Prime version of the experiment. The results of this study offer partial support for these assumptions, but this was to a very minor level, and as such cannot be generalised outside of this study.

Analyses in this study focused on both response choice and Response Time for positive responses. The automatic removal of Version from the models of best fit for both Response (at all levels) and Response Time (for two- and three-way interactions) indicates that there were no significant overall differences between the results obtained in PsyToolkit and E-Prime. Version was re-added at all levels to the final models that were analysed. The results for Response indicated that there were no significant main or interaction effects involving Version at the 95% confidence level, strongly supporting the idea that results obtained through PsyToolkit are in line with those obtained in a laboratory setting. The results for Response Time indicated that there were no interaction effects involving Version at the 95% confidence level, and, while a significant main effect of Version was found, this was found to indicate a general tendency towards faster responses for those who undertook the PsyToolkit version of the

experiment, but overlapping 95% confidence intervals indicated that we cannot be sure that there is truly a difference in response time between responses received through PsyToolkit and E-Prime. As such, the results of Response Time also support the idea that results obtained through PsyToolkit are in line with those obtained in a laboratory setting.

Although no significant three-way interaction was found for either response or response time, some general variability between the two Versions can be seen. For Response, this variability takes the form of participants who undertook the PsyToolkit version of the experiment tending towards answering less positively, especially for female first name / masculine role nouns pairings, than those who undertook the E-Prime version of the experiment. For Response Time this takes the form of participants who undertook the PsyToolkit version tending to respond faster than those who undertook the PsyToolkit version, while participants who undertook the E-Prime version of the experiment had larger mean response time differences between congruent and incongruent pairings. While this is in line with the expected effects of ecological validity, it is also possible that the differences in mean Response Time found between PsyToolkit and E-Prime are due to differences in the way PsyToolkit and E-Prime measure reaction time. Further, the difference in Response Time standard errors between PsyToolkit and E-Prime (mean difference = 10) is in keeping with the concept of increased ecological validity.

It was expected that Response Time and standard deviations for both PsyToolkit and E-Prime should be higher than those found by Reimers and Stewart (2007) and by Schubert et al. (2013). Mean Response Time in this study was found to be higher than mean Response Time for both Reimers and Stewart (2007) (approximately 900ms) and for Experiment 5 of Schubert et al. (2013) (approximately 800ms), and Response Time standard errors were higher than both Reimers and Stewart (2007) and Schubert et al (2013). The increase in Response Time supports the idea that this is a more complex decision task than those used by Reimers and Stewart (2007) and Schubert et al. (2013).

Since data collection was completed, updates have been released for both PsyToolkit and for E-Prime 3.0. These updates have improved performance and may affect response time measurements for both PsyToolkit and E-Prime. However, as the results presented in this study already show high levels of replicability between the instruments, it is unlikely that these updates would remove this replicability, at least for the task at hand.

In conclusion, the results of the current study indicated that PsyToolkit is a viable method for conducting both general and psycholinguistic specific experiments that utilise CRT tasks, with effects found to replicate for both response choice and response time.

**Acknowledgements**

We would like to thank students in the PSY2900 and PSY2909 courses at the Norwegian University of Science and Technology for assisting with recruitment and data collection.

## References


Baayen, R. H., (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics using R.* Cambridge, UK: Cambridge University Press

Baayen, R. H., & Milin, P., (2010) Analyzing reaction times. *International Journal of Psychological Research, 3*: 12–28.

Bates, D., Maechler, M., Bolker, B., & Walker, S., (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1): 1-48

Enochson, K., & Culbertson, J., (2015) Collecting psycholinguistic response time data using amazon mechanical turk. *PLoS One, 10*(3): e0116946

Fisher, R. J., (1993) Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20*(2): 303-315

Forster, K. I., & Forster, J. C., (2003) DMDX: a windows display program with millisecond accuracy. *Behaviour Research Methods, Instruments, & Computers, 35*: 116-124

Fox, J., (2003) Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*(15): 1–27

Gabriel, U., & Gygax, P., (2008) Can societal language amendments change gender representation? The case of Norway. *Scandinavian Journal of Psychology, 49*(5): 451-457

Gygax, P., & Gabriel, U., (2008) Can a group of musicians be composed of women? Generic interpretation of French masculine role names in the absence and presence of feminine forms. *Swiss Journal of Psychology, 67*: 143-151

Gygax, P., Gabriel, U., Lévy, A., Pool, E., Grivel, M., & Pedrazzini, E., (2012) The masculine form and its competing interpretations in French: when linking grammatically masculine role names to female referents is difficult. *Journal of Cognitive Psychology, 24*(4): 395-408

Holbrook, A. L., Green, M. C., & Krosnick, J. A., (2003) Telephone versus face-to-face interviewing of national probability samples with long questionnaires: comparisons of responding satisficing and social desirability response bias. *Public Opinion Quarterly, 67*: 79-125

Høiland-Jørgensen, T., Ahlgren, B., Hurtig, P., & Brunstrom, A., (2016) Measuring latency variation in the internet. *Proceedings of the 12th international conference on emerging network experiments and technologies* (pp. 473-480). Irvine, CA: ACM

King, M. F., & Bruner, G. C., (2000) Social desirability bias: a neglected aspect of validity testing. *Psychology and Marketing, 17*(2): 79-103

Krantz, J. H., & Reeshad, S. D., (2000) Validity of web-based psychological research. In M H. Birnbaum (Ed.) *Psychological experiments on the Internet* (pp. 35-60). San Diega, CA: Academic Press

Lüdecke, D., (2018) *Sjstats: statistical functions for regression models.* Retrieved from https://CRAN.R-project.org/package=sjstats

Misersky, J., Gygax, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., Chiarini, T., Englund, K., Hanulikova, A., Öttl, A., Valdrova, J., Von Stockhausen, L., & Sczesny, S., (2014) Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavioural Research Methods*, *46*(3): 841-871

Nakagawa, S. & Schielzeth, H., (2010) A general and simple method for obtaining $R^2$ from generalised linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2): 133-142

Nicholson, D. B., Parboteeah, D. V., Nicholson, J. A., Valacich, J. S., (2005) Using distraction-conflict theory to measure the effects of distractions on individual task performance in a wireless mobile environment. *Proceedings from the 38th Annual Hawaii International Conference on System Science.* Piscataway, NJ: Institute of Electrical and Electronic Engineers.

Peugh, J. L., (2010) A practical guide to multilevel modelling. *Journal of School Psychology, 48*(1): 85-112

Plant, R. R., (2009) Millisecond precision psychological research in a world of commodity computers: new hardware, new problems? *Behaviour Research Methods, 41*(3): 598-614

Reimers, S. & Stewart, N., (2007) Adobe flash as a medium for online experimentation: a test of reaction time measurement capabilities. *Behaviour Research Methods, 39*(3): 365-370

Reips, U. -D., (2000) The web experiment method: advantages, disadvantages, and solutions. In M H. Birnbaum (Ed.) *Psychological experiments on the Internet* (pp. 89-114). San Diega, CA: Academic Press

Reips, U. -D., (2002) Standards for Internet-based experimenting. *Experimental Psychology, 49*(4): 243-256

Sampaio, T. O. d.M., (2017) A escolha de software e hardware na psicolinguisticia: revisão e opinião [The choice of software and hardware in psycholinguistics: review and opinion]. *Belo Horizonte, 25*(3): 971-1010

Sanders, G. S., Baron, R. S., & Moore, D. L., (1978) Distraction and social comparison as mediators of social facilitation effects. *Journal of Experimental Social Psychology, 14*(3): 291-303

Schubert, T. W., Murteira, C., Collins, E. C., & Lopes, D., (2013) ScriptingRT: a software library for collecting response latencies in online studies of cognition. *PLoS One, 8*(6): e67769

Stoet, G., (2010) PsyToolkit: a software package for programming psychological experiments using Linux. *Behaviour Research Methods, 42*(4): 1096-1104

Stoet, G., (2017) PsyToolkit: a novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*(1): 24-31

Stoet, G., (2018) *Frequently Asked Questions (FAQ)*. Retrieved from https://www.psytoolkit.org/faq.html

Swan, T., (1992) All about eve: women in Norwegian newspapers in the 20th century. *Working papers on Language, Gender, and Sexism, 2*(2): 37-54

Wagner-Egger, P., & Gygax, P., (2017) Diana was not involved in the 9/11 terrorist attacks! Or was she? Newspaper headings and the boomerang effect. *Swiss Journal of Psychology, 77*(1): 15-22

Zajdel, R., & Nowak, D., (2007) Simple and complex reaction time measurement. A preliminary evaluation of new approach and diagnostic tool. *Computers in Biology and Medicine, 37*: 1724-1730

Öttl, A., (2018) [Study examining stereotypicality ratings associated with job roles in Finnish, French, and Norwegian]. Unpublished raw data.

Table 1

*Stereotypicality score for feminine experimental role nouns as determined from the findings of Misersky et al.*

| **Role noun** | **English translation** | **Score** | **SD** |
|---|---|---|---|
| Manikyrister | Manicurists | .88 | .08 |
| Bryllupsplanleggere | Wedding planners | .85 | .10 |
| Kosmetikere | Beauticians | .85 | .10 |
| Eksotiske dansere | Exotic dancers | .83 | .10 |
| Prostituerte | Prostitutes | .83 | .13 |
| Strippere | Strippers | .81 | .18 |
| Fødselshjelpere | Birth attendants | .80 | .13 |
| Frisører | Hairdressers | .79 | .11 |
| Barnevakter | Childminders | .78 | .13 |
| Groupier | Groupies | .77 | .17 |
| Synske | Clairvoyants | .76 | .12 |
| Sekretærer | Secretaries | .75 | .10 |
| *Mean* | | .81 | .12 |

Table 2

*Stereotypicality score for masculine experimental role nouns as determined from the findings of Misersky et al.*

| **Role noun** | **English translation** | **Score** | **SD** |
|---|---|---|---|
| Fabrikkbestyrere | Factory managers | .25 | .13 |
| Fyrvoktere | Lighthouse keepers | .24 | .15 |
| Guvernører | Governors | .23 | .12 |
| Datateknikere | Computer technicians | .23 | .09 |
| Skogsforvaltere | Forest rangers | .22 | .14 |
| Trommeslagere | Drummers | .21 | .11 |
| Astronauter | Astronauts | .20 | .12 |
| Brytere | Wrestlers | .20 | .18 |
| Søppeltømmere | Rubbish collectors | .17 | .11 |
| Taktekkere | Roofers | .17 | .14 |
| Kranførere | Crane operators | .15 | .10 |
| Soldater | Soldiers | .15 | .11 |
| *Mean* | | .20 | .13 |

Table 3

*Stereotypicality score for non-stereotyped experimental role nouns as determined from the findings of Misersky et al.*

| **Role noun** | **English translation** | **Score** | **SD** |
|---|---|---|---|
| Fysioterapeuter | Physiotherapists | .60 | .11 |
| Miljøaktivister | Environmentalists | .60 | .12 |
| Fiolinister | Violinists | .59 | .14 |
| Arkivarer | Archivists | .57 | .19 |
| Meteorologer | Meteorologists | .55 | .19 |
| Akrobater | Acrobats | .53 | .13 |
| Kunstnere | Artists | .53 | .11 |
| Fagforeningsmedlemmer | Trade unionists | .51 | .10 |
| Fotografer | Photographers | .51 | .13 |
| Biologer | Biologists | .46 | .16 |
| Oceanografer | Oceanographers | .45 | .14 |
| Idrettsutøvere | Athletes | .42 | .11 |
| *Mean* | | .53 | .14 |

Table 4

*Typicality of female and male first names as indicated by response time results from the findings of Öttl.*

| **First Name Gender** | **Name** | **Mean Response Time** |
|---|---|---|
| Male | Espen | 566ms |
| | Geir | 574ms |
| | Robert | 583ms |
| Female | Ida | 584ms |
| | Nina | 565ms |
| | Sandra | 573ms |

Table 5

*Effect sizes for the fixed effects in the model 'effect of Version (Web-Based vs. Laboratory Based) on positive responses. Table shows the estimated effect size and 95% confidence intervals. Intercept included Masculine Roles, Female Names, and Version Web as contrast levels.*

| **Fixed effect** | **Effect size** | **Lower Bound** | **Upper Bound** |
|---|---|---|---|
| Intercept | 4.856 | 2.812 | 6.899 |
| Trial Number | 0.006 | 0.004 | 0.008 |
| Feminine Roles | 0.281 | -0.413 | 0.975 |
| Non-Stereotyped Roles | -0.322 | -0.960 | 0.316 |
| Male Names | 0.029 | -0.172 | 0.230 |
| Version Lab | 0.201 | -1.741 | 2.143 |
| Feminine Roles: Male Names | -1.715 | -1.871 | -1.559 |
| Non-stereotyped Roles: Male Names | 0.288 | 0.128 | 0.448 |
| Feminine Roles: Version Lab | 0.112 | -0.211 | 0.435 |
| Non-stereotyped Roles: Version Lab | -0.137 | -0.420 | 0.146 |
| Male Names: Version Lab | -0.007 | -0.216 | 0.202 |
| Feminine Roles: Male Names: Version Lab | -0.142 | -0.287 | 0.003 |
| Non-stereotyped Roles: Male Names: Version Lab | 0.048 | -0.110 | 0.206 |

Table 6

*The effect of the two-way interaction between Stereotype and Name Gender on response. Table shows mean positive response (%) and 95% confidence interval for each name/role pairing, rounded to the nearest full percentage.*

| **Stereotype** | **Name Gender** | **Mean Response** | **Lower Bound** | **Upper Bound** |
|---|---|---|---|---|
| Feminine | Female | 100 | 99 | 100 |
| | Male | 99 | 91 | 100 |
| Non-Stereotyped | Female | 100 | 98 | 100 |
| | Male | 99 | 96 | 100 |
| Masculine | Female | 99 | 91 | 100 |
| | Male | 100 | 100 | 100 |

Table 7

*The effect of the three-way interaction between Version, Stereotype, and Name Gender on response. Table shows mean positive response (%) and 95% confidence interval for each name/role pairing, rounded to the nearest full percentage.*

| **Version** | **Stereotype** | **Name Gender** | **Mean Response** | **Lower Bound** | **Upper Bound** |
|---|---|---|---|---|---|
| Version Web (PsyToolkit) | Feminine | Female | 100 | 96 | 100 |
| | | Male | 99 | 63 | 100 |
| | Non-Stereotyped Masculine | Female | 100 | 88 | 100 |
| | | Male | 99 | 81 | 100 |
| | | Female | 99 | 62 | 100 |
| | | Male | 100 | 97 | 100 |
| Version Lab (E-Prime) | Feminine | Female | 100 | 100 | 100 |
| | | Male | 99 | 95 | 100 |
| | Non-Stereotyped Masculine | Female | 100 | 99 | 100 |
| | | Male | 99 | 98 | 100 |
| | | Female | 99 | 96 | 100 |
| | | Male | 100 | 100 | 100 |

Table 8

*Effect sizes for the fixed effects in the model 'effect of Version (Web-Based vs. Laboratory Based) on positive responses. Table shows the estimated effect size and 95% confidence intervals. Intercept included Masculine Roles, Female Names, and Version Web as contrast levels*

| **Fixed effect** | **Effect size** | **Lower Bound** | **Upper Bound** |
|---|---|---|---|
| Intercept | 6.993 | 6.859 | 7.127 |
| Trial Number | -0.001 | -0.001 | -0.001 |
| Feminine Roles | 0.027 | -0.002 | 0.055 |
| Non-Stereotyped Roles | -0.027 | -0.055 | 0.000 |
| Male Names | -0.011 | -0.031 | 0.008 |
| Version Lab | -0.077 | -0.149 | -0.005 |
| Number of Characters | 0.009 | 0.003 | 0.014 |
| Feminine Roles: Male Names | -0.026 | -0.033 | -0.018 |
| Non-stereotyped Roles: Male Names | 0.005 | -0.002 | 0.013 |
| Feminine Roles: Version Lab | -0.002 | -0.017 | 0.013 |
| Non-stereotyped Roles: Version Lab | -0.003 | -0.016 | 0.010 |
| Male Names: Version Lab | 0.003 | -0.008 | 0.014 |
| Feminine Roles: Male Names: Version Lab | 0.002 | -0.006 | 0.009 |
| Non-stereotyped Roles: Male Names: Version Lab | 0.001 | -0.007 | 0.008 |

Table 9

*The effect of the two-way interaction between Stereotype and Name Gender on response time. Table shows mean response time (ms), SE, and 95% confidence interval for each name/role pairing.*

| Stereotype | Name Gender | Mean Response | SE | Lower Bound | Upper Bound |
|---|---|---|---|---|---|
| Feminine | Female | 1052 | 44 | 970 | 1141 |
| | Male | 1132 | 48 | 1041 | 1230 |
| Non-Stereotyped | Female | 1028 | 47 | 940 | 1124 |
| | Male | 1040 | 48 | 949 | 1139 |
| Masculine | Female | 1073 | 49 | 980 | 1174 |
| | Male | 1053 | 49 | 961 | 1155 |

Table 10

*The effect of the three-way interaction between Version, Stereotype, and Name Gender on response time. Table shows mean response time (ms), SE, and 95% confidence interval for each name/role pairing.*

| **Version** | **Stereotype** | **Name Gender** | **Mean Response** | **SE** | **Lower Bound** | **Upper Bound** |
|---|---|---|---|---|---|---|
| Version Web (PsyToolkit) | Feminine | Female | 978 | 52 | 881 | 1087 |
| | | Male | 1044 | 58 | 937 | 1164 |
| | Non-Stereotyped | Female | 954 | 57 | 848 | 1072 |
| | | Male | 959 | 58 | 851 | 1081 |
| | Masculine | Female | 1001 | 60 | 890 | 1126 |
| | | Male | 982 | 60 | 870 | 1107 |
| Version Lab (E-Prime) | Feminine | Female | 1135 | 61 | 1021 | 1262 |
| | | Male | 1232 | 69 | 1105 | 1374 |
| | Non-Stereotyped | Female | 1112 | 67 | 988 | 1251 |
| | | Male | 1132 | 70 | 1003 | 1278 |
| | Masculine | Female | 1154 | 70 | 1025 | 1300 |
| | | Male | 1134 | 70 | 1005 | 1281 |


*Figure 1:* The effect of Stereotype and Name Gender on response. Error bars indicate the 95% confidence interval.

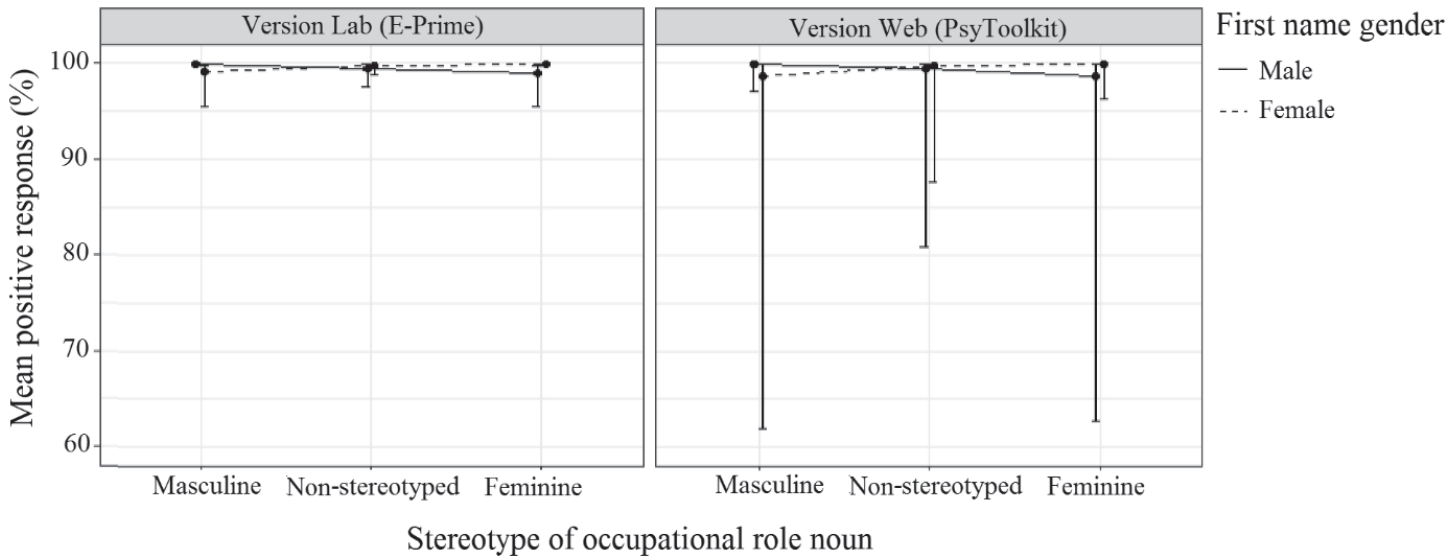


*Figure 2:* The effect of Version, Stereotype, and Name Gender on response. Error bars indicate the 95% confidence interval.

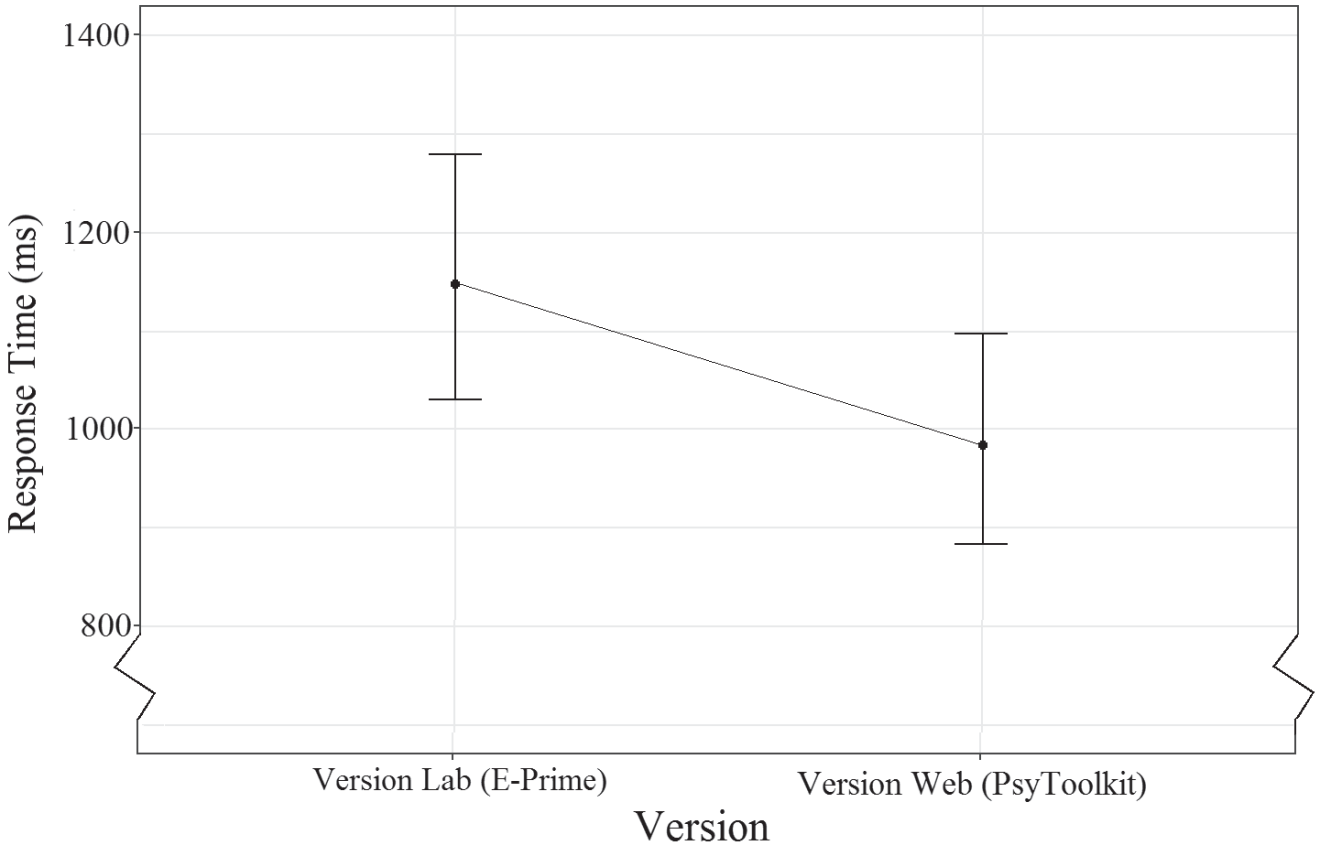


*Figure 3:* The main effect of Version on response time. Error bars indicate the 95% confidence interval.

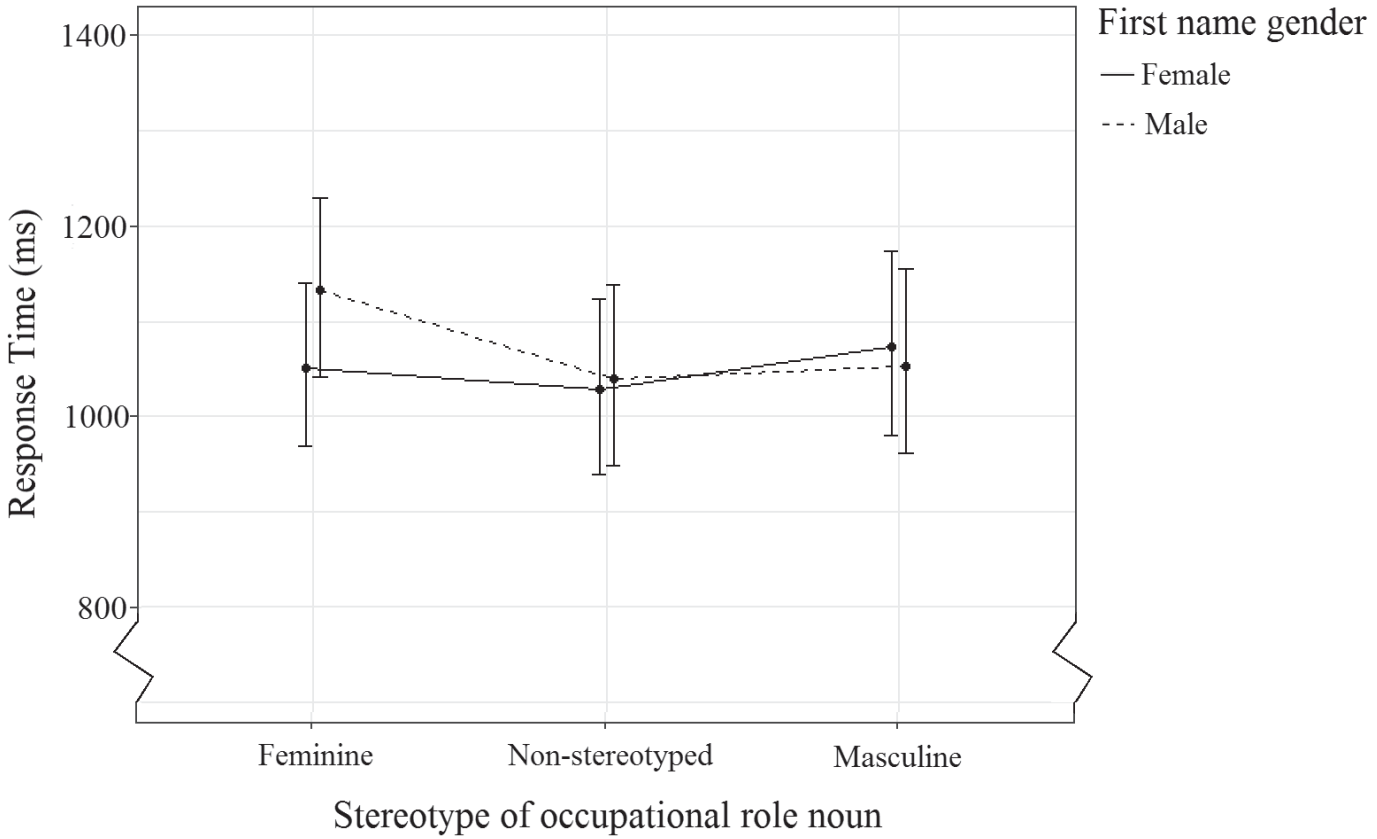


*Figure 4:* The effect of Stereotype and Name Gender on response time. Error bars indicate the 95% confidence interval.

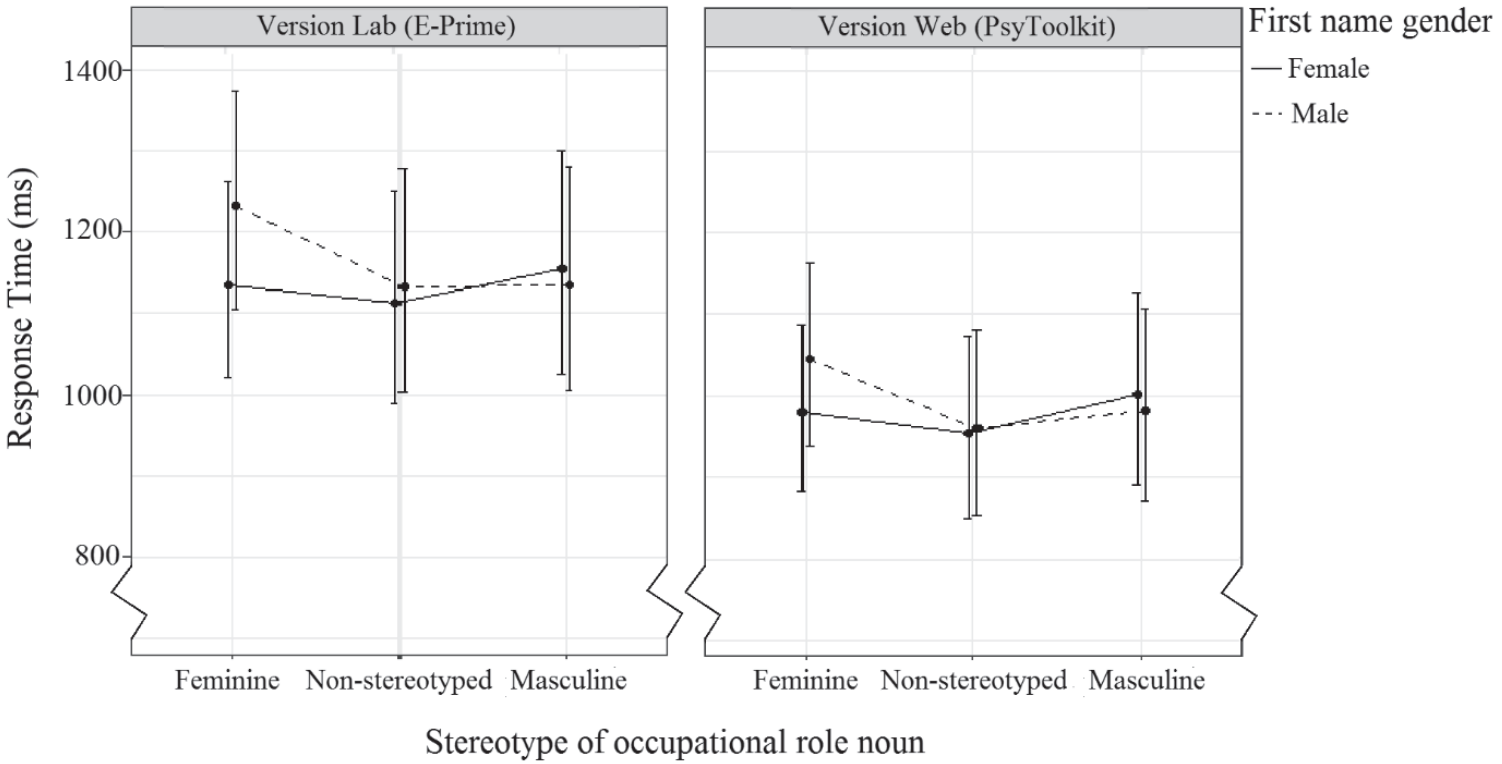


*Figure 5:* The effect of Version, Stereotype, and Name Gender on response time. Error bars indicate the 95% confidence interval.

## 7.2. Paper II: Investigating the link between gender stereotypicality and occupational stereotype content through a bottom-up approach.

**Authors:** Jonathan D. Kim, Ute Gabriel, Pascal Gygax, and Anna Siyanova-Chanturia.

**Status:** Manuscript in Preparation*

** A previous version of this manuscript was submitted to a journal for consideration but was ultimately rejected. The journal in question provided reviewers comments along with the letter of rejection. The version of the manuscript included in this thesis has been updated in line with these suggestions.*

**7.3. Paper III: Language structures and gender stereotyped perception: The effect of differences in the level of grammatical gender between fully, semi-, and non-gendered languages.**

**Authors:** Jonathan D. Kim, Ute Gabriel, and Pascal Gygax.

**Status:** Manuscript in Preparation*

* *A previous version of this manuscript was submitted to a journal for consideration but was ultimately rejected. The journal in question provided reviewers comments along with the letter of rejection. The version of the manuscript included in this thesis has been updated in line with these suggestions.*

ISBN 978-82-326-4332-5 (printed ver.)
ISBN 978-82-326-4333-2 (electronic ver.)
ISSN 1503-8181



NTNU
Norwegian University of
Science and Technology