Strategies for overcoming gender stereotypes in cognitive representations

Eimear Finnegan

Thesis submitted for the degree of Doctor of Philosophy

School of Psychology

University of Sussex

Declaration

I hereby declare that this thesis has not been, and will not be, submitted in whole or in part to another University for the award of any other degree.

Eimear Finnegan

January, 2014.

Acknowledgements

First and foremost, thank you to my supervisors Jane Oakhill and Alan Garnham for their advice and guidance throughout my PhD. I have learned a huge amount from you both and really appreciate the consistent support you offered along the way - thank you so much.

I am hugely indebted to Sabine Sczesny and Lisa von Stockhausen as organisers of the fantastic Initial Training Network on Language, Cognition and Gender and would like to thank them for the tireless effort they put in to ensuring its success. Thanks also to all the other partners and fellows of ITN LCG. Our workshops, summer schools and conferences across Europe have left me with many fond memories and I look forward to our paths crossing again. A special shout-out to Dries, President of the Chubby Club.

Thank you to my colleagues at Sussex, both old and new, who shared the journey and helped make this PhD so enjoyable. Special thanks to Natalie who has been there every step of the way as a fantastic friend and organiser of my social life, to Paolo for his help and guidance as I began my studies and helped me to find my research feet and to Clio, a fellow April-starter and Jedward enthusiast who was always there to swap stories along the way.

To my other friends in Ireland and abroad, thanks for listening to me talk about little other than my thesis over the past few years - I am very much looking forward to spending more time with you as I get my life back post-thesis. Mairead, Mary, Oonagh, Ger and Doireann – you can now look forward to many many more visits!

A huge thank you to my fantastic family, especially my parents who have got me where I am today. I appreciate your continuous efforts to feign interest in my studies despite having no idea what I do. You can look forward to a shiny copy of this thesis to figure it out!

Finally, to Eoin – I can't thank you enough. I'll do the commuting wherever we go next!

The fact that I have made it through this PhD without any tears speaks volumes for you all!

UNIVERSITY OF SUSSEX

Eimear Finnegan

Thesis submitted for the degree of Doctor of Philosophy

STRATEGIES FOR OVERCOMING GENDER STEREOTYPES IN COGNITIVE REPRESENTATIONS

SUMMARY

Gender stereotypes are activated spontaneously and unintentionally when certain role nouns are read. For example, people expect a builder to be male and a beautician to be female. Such gender inferences lead to processing difficulties when violations of stereotypical gender occur. The aim of this thesis was to devise strategies aimed at overcoming the activation of gender stereotype biases in English.

Across nine studies, a variety of stereotype reduction strategies were investigated in conjunction with a judgement task, devised by Oakhill, Garnham and Reynolds (2005). This judgement task asked participants to decide, without deliberation, whether two terms presented onscreen could refer to one person. In the absence of a stereotype-reduction training, participants consistently showed evidence of succumbing to stereotype biases on stereotype incongruent pairings (e.g. Builder/ Mother) compared to stereotype congruent pairings (e.g. Builder/ Father). However, accuracy and response time performance to these incongruent pairings were found to significantly improve from pre-training levels to post-training levels through the use of stereotype reduction strategies such as providing participants with performance-related feedback (Experiment 1, Experiment 3), social consensus feedback (Experiment 4), combined social and accuracy feedback (Experiment 6) and counter-stereotype pictures (Experiment 8). A number of individual difference measures were also administered with the behavioural tasks. These explored whether individual differences in levels of ambivalent sexism, attitudes towards sexist language, sex role perception, and, among others, sexist pronoun use could moderate performance on the judgement task. The results from these additional tasks are described in Chapter 5.

This thesis provides further evidence for the malleability of stereotype biases and delineates specific strategies through which stereotype biases can be overcome, to ultimately result in lower levels of stereotype endorsement.

Table of Contents

Statement....ii

Acknowledgements....iv

Summary....v

Table of Contents....vi

List of Figures....ix

List of Tables....x

1 Language, gender and stereotypes....1

1.1 Gender in language....1

1.2 Language and gender asymmetry....2

1.3 The development of gender stereotypes....3

1.4 The cognitive processes underlying stereotype activation....5

1.4.1 Stereotypes: Background....5

1.4.2 Automatic and controlled processes in information processing....6

1.4.3 Models of stereotype representation....8

1.5 How gender stereotypes affect language processing in text comprehension....11

1.5.1 Gender inferences....11

1.5.2 When are gender stereotypes activated?....12

1.5.3 How persistent is the stereotyping effect?....14

1.5.4 Overcoming gender biases in text....16

1.6 Overcoming stereotypes....19

1.6.1 Disadvantages of stereotypes....19

1.6.2 Challenge of overcoming stereotypes....20

1.6.3 Factors affecting stereotype use....21

1.6.4 Instigating stereotype change....23

1.6.5 Stereotype reduction....25

1.6.6 Stereotype activation vs. stereotype application....25

1.6.7 Previous strategies aimed at stereotype reduction..... 27

1.7 Individual differences in stereotyping..... 33

1.7.1 Individual differences and gender processing in language.....36

1.8 Ingredients of a successful intervention..... 37

1.8.1 Past meta-analyses..... 37

1.9 Observations from the field and suggestions for future research.....42

2 Performance-related feedback as a strategy to overcome automatic gender stereotypes..... 48

2.1 Introduction..... 48

2.2 Experiment 1: Performance Feedback..... 50

2.2.1 Introduction..... 50

2.2.2 Method..... 51

2.2.3 Results..... 55

2.2.4 Discussion..... 63

2.3 Experiment 2: Control study..... 64

2.3.1 Introduction..... 64

2.3.2 Method..... 65

2.3.3 Results..... 65

2.3.4 Discussion..... 72

2.4 Experiment 3: Long term and transfer effects..... 73

2.4.1 Introduction..... 73

2.4.2 Method..... 74

2.4.3 Results..... 75

2.4.4 Discussion..... 82

2.5 Experiments 1, 2 and 3: Combined analysis..... 84

2.6 Chapter Discussion..... 91

3 Social-consensus feedback as a strategy to overcome automatic gender stereotypes 97

3. 1 Introduction....97

3.2 Experiment 4: Social consensus feedback....98

3.2.1 Introduction....98

3.2.2 Method....99

3.2.3 Results....101

3.2.4 Discussion....106

3.3 Experiment 5: Reverse social consensus feedback....107

3.3.1 Introduction....107

3.3.2 Pilot study 1....108

3.3.3 Method....110

3.3.4 Results....110

3.3.5 Discussion....115

3.4 Experiments 4 and 5: Combined analysis....116

3.5 Experiment 6: Accuracy and social feedback....119

3.5.1 Introduction....119

3.5.2 Method....120

3.5.3 Results....120

3.5.4 Discussion....125

3.6 Experiments 4 and 6: Combined analysis....126

3.7 Performance feedback vs. Social consensus feedback....129

3.8 Chapter Discussion....134

4 Counter-stereotypic strategies aimed at overcoming immediate activation of gender biases....137

4.1 Introduction....137

4.2 Experiment 7: Counter-stereotype association learning....139

4.2.1 Introduction....139

4.2.2 Method....139

4.2.3 Results....142

4.2.4 Discussion..... 148

4.3 Experiments 8 & 9: Counter-stereotypic versus Stereotypic pictures as a strategy to overcome immediate activation of gender stereotypes..... 150

4.3.1 Pilot study 2..... 151

4.4 Experiment 8: Counter-stereotypic pictures as a strategy to overcome immediate activation of gender stereotypes..... 152

4.4.1 Introduction..... 152

4.4.2 Method..... 153

4.4.3 Results..... 155

4.4.4 Discussion..... 160

4.5 Experiment 9: Stereotypical pictures; control condition..... 161

4.5.1 Introduction..... 161

4.5.2 Method..... 161

4.5.3 Results..... 162

4.5.4 Discussion..... 167

4.6 Experiments 8 and 9: Combined analysis..... 168

4.7 Picture Booklets: analysis of participant responses..... 172

4.8 Chapter Discussion..... 183

5 Individual differences in gender stereotyping..... 185

5.1 Introduction..... 185

5.2 The Individual difference measures..... 187

5.3 Individual difference analyses..... 192

5.4 Chapter Discussion..... 217

6 General Discussion..... 222

6.1 Introduction..... 222

6.2 Main findings: A summary..... 222

6.3 Theoretical implications..... 228

6.4 Methodological limitations and future research..... 231

6.5 Final conclusion.....233

Reference List..... 235

Appendix 1: Ethical Approval Certificate.....256

Appendix 2: Role Nouns: Experiments 1-6.....257

Appendix 3: Filler Role Nouns: Experiments 1-6..... 258

Appendix 4: Information Sheet..... 259

Appendix 5: Consent Form.....260

Appendix 6: Role Nouns: Experiment 3.....261

Appendix 7: Fictitious Feedback range: Social Consensus Feedback..... 263

Appendix 8: Fictitious Feedback range: Reverse Social Consensus Feedback.....264

Appendix 9: Pilot study 1: Plausibility of the Social Consensus Feedback.....265

Appendix 10: Materials: Experiment 7.....266

Appendix 11: Filler Role Nouns: Experiment 7.....268

Appendix 12: Stereotypical & Counter-stereotypical pictures..... 270

Appendix 13: Example Booklet for Experiment 8 (Counter-Stereotype Pictures)..... 276

Appendix 14: Rater's Instructions for the booklet analysis..... 277

Appendix 15: Ethics Questionnaire.....279

List of Figures

Figure 2.1. Experiment 1: Mean percentages of correct judgements to critical word pairs across blocks……..58

Figure 2.2 Experiment 1: Mean response times (in milliseconds) of correct judgements to critical word pairs across blocks..... 60

Figure 2.3 Experiment 2: Mean percentages of correct judgements to critical word pairs across blocks……..67

Figure 2.4 Experiment 2: Mean accuracy of judgements of male and female participants to congruent and incongruent word pairs…..... 68

Figure 2.5 Experiment 2: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks..... 69

Figure 2.6 Experiment 2: Mean response times (in milliseconds) of male and female participants to congruent and incongruent word pairs…….. 70

Figure 2.7 Experiment 3: Mean percentages of correct judgements to critical word pairs across blocks..... 76

Figure 2.8 Experiment 3: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks..... 79

Figure 2.9 Mean % accuracy to critical word pairs across blocks in Experiments 1 to 3..... 85

Figure 2.10 Mean % difference in accuracy scores between Block 1 and Block 3, to critical word pairs across Experiments 1-3..... 87

Figure 2.11 Mean response times (in milliseconds) to critical word pairs across blocks in Experiments 1 to 3..... 89

Figure 3.1 Experiment 4: Mean percentages of correct judgements to critical word pairs across blocks..... 102

Figure 3.2 Experiment 4: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks..... 104

Figure 3.3 Experiment 5: Mean percentages of correct judgements to critical word pairs across blocks..... 112

Figure 3.4 Experiment 5: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks..... 113

Figure 3.5 Mean accuracy to critical word pairs across blocks in Experiments 4 and 5..... 117

Figure 3.6 Mean response times (in milliseconds) to correct critical word pairs across blocks in Experiments 4 and 5..... 118

Figure 3.7 Experiment 6: Mean percentages of correct judgements to critical word pairs across blocks..... 122

Figure 3.8 Experiment 6: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks..... 124

Figure 3.9 Mean % accuracy to critical word pairs across blocks in Experiments 4 and 6..... 127

Figure 3.10 Mean response times (in milliseconds) to correct critical word pairs, across blocks in Experiments 4 and 6..... 128

Figure 3.11 Mean percentages of correct judgements to critical word pairs across blocks in Experiments 1, 4 and 6..... 130

Figure 3.12 Mean response times (in milliseconds) to correct critical word pairs across blocks in Experiments 1, 4 and 6..... 132

Figure 4.1 Experiment 7: Mean percentages of correct judgements to critical word pairs in Block 1 and Block 2..... 144

Figure 4.2 Experiment 7: Mean response times (in milliseconds) of correct judgements to critical word pairs across Block 1 and Block 2..... 146

Figure 4.3 Experiment 7: Mean response times (in milliseconds) of correct judgements to critical word pairs across Block 1 and Block 2, for both female and male participants..... 147

Figure 4.4 Experiment 8: Mean percentages of correct judgements to critical word pairs in Block 1 and Block 2..... 156

Figure 4.5 Experiment 8: Mean response times (in milliseconds) of correct judgements to critical word pairs across Block 1 and Block 2..... 159

Figure 4.6 Experiment 9: Mean percentages of correct judgements to critical word pairs in Block 1 and Block 2..... 163

Figure 4.7 Experiment 9: Mean response times (in milliseconds) of correct judgements to critical word pairs across Block 1 and Block 2..... 165

Figure 4.8 Experiment 9: Mean RT difference (in milliseconds) between Block 1 and Block 2 performance of female and male participants in response to critical word pairs……..166

Figure 4.9 Mean percentages of correct judgements to critical word pairs across blocks in both picture experiments..... 169

Figure 4.10 Mean response times to critical word pairs across blocks in both picture experiments.....171

List of Tables

Table 4.1 Booklet analyses: Mean scores and significance levels of comparisons across groups……..175

Table 5.1 Experiments 1-9 and their associated individual difference measures..... 192

Table 5.2 Individual difference measures; Correlations (using Pearson's correlation coefficients). 194

Table 5.3 Analysis 1: Correlation coefficients of the nine predictors and Initial Performance..... 196

Table 5.4 Analysis 1: Hierarchical multiple regression analysis: Initial Performance as the dependent variable.....198

Table 5.5 Analysis 2: Correlation coefficients of the nine predictors and Initial Performance…… 199

Table 5.6 Analysis 2: Hierarchical multiple regression analysis: Initial Performance as the dependent variable.....200

Table 5.7 Analysis 3: Correlation coefficients of the nine predictors and Performance Improvement…… 201

Table 5.8 Analysis 3: Hierarchical multiple regression analysis: Performance Improvement as the dependent variable..... 203

Table 5.9 Analysis 4: Correlation coefficients of the nine predictors and Performance Improvement…… 204

Table 5.10 Analysis 5: Correlation coefficients of the BFoNE and the outcome variables of Initial Performance and Performance Improvement..... 206

Table 5.11 Analysis 5: Hierarchical multiple regression analysis of male participants' data: Performance Improvement as the dependent variable..... 207

Table 5.12 Analysis 6: Correlation coefficients of the BFoNE and the outcome variables of Initial Performance and Performance Improvement.....208

Table 5.13 Analysis 6: Hierarchical multiple regression analysis of male and female participant's data separately: RT Performance Improvement as the dependent variable..... 210

Table 5.14 Analysis 7: Correlation coefficients of the ASI and the outcome variables of Initial Performance and Performance Improvement..... 211

Table 5.15 Analysis 7: Hierarchical multiple regression analysis: Initial Performance as the dependent variable.....212

Table 5.16 Analysis 8: Correlation coefficients of the ASI and the outcome variables of Initial Performance and Performance Improvement.....213

Table 5.17 Analysis 9: Correlation coefficients of the MSS, BSRI, IAT and the outcome variables of Initial Performance and Performance Improvement..... 215

Table 5.18 Analysis 10: Correlation coefficients of the MSS, BSRI, IAT and the outcome variables of Initial Performance and Performance Improvement..... 216

Table 6.1 Accuracy (%) performance across blocks in Experiments 1-9 223

Table 6.2 Response time (ms) performance across blocks in Experiments 1-9..... 223

1. Language, gender and stereotypes

1.1 Gender in language

Gender[1] is a fundamental social category, so essential to social organisation and structure that linguistic means of referring to maleness and femaleness have developed for all speech communities (Stahlberg, Braun, Irmen, & Sczesny, 2007). While there is considerable variation in how the sexes are represented linguistically, three language types have now been identified to classify languages according to how it is denoted therein (1) grammatical gender languages where sex is encoded grammatically and gender marking is extremely commonplace (e.g. French, German, Czech) (2) genderless languages where information about a person's sex is not encoded grammatically (e.g. Turkish, Finnish) and (3) between these two extremes are natural gender languages such as English where sex is not grammatically marked on most personal nouns (such as adolescent, neighbour) but is still communicated through grammatical gender, lexical gender, referential gender and social gender (Hellinger & Bußmann, 2001; Stahlberg et al., 2007). It is the use of this final category of social gender that is the focus of this research and which I will elaborate on below.

While English has a number of personal nouns that contain either a male or female semantic property as part of their lexical definitions (e.g. father, girl, son), or are formally marked for lexical gender through the use of suffixes (e.g. waitress, landlord, landlady), the majority of human nouns in English are *not* gender specific. Instead, the sex of a human noun is typically indicated through social gender. This refers to stereotypical assumptions about appropriate male and female social roles and the typical members of these roles (Hellinger & Bußmann, 2001). Indeed this social gender is now more commonly referred to as gender (stereo)typicality in the psycholinguistic and social psychology literature and is simply defined as the likelihood of a noun referring to women or men (Irmen & Roßberg, 2004). This gender typicality plays an important role in building a cognitive representation of gender and is the reason for which people come to expect, for example, surgeons to be male and nurses to be female.

[1] Although the terms 'sex' and 'gender' are now often used to differentiate between biologically determined vs. socially constructed ideas of maleness and femaleness respectively (e.g. Pryzgoda & Chrisler, 2000; Diamond, 2000), use of the term 'gender' in this thesis does not explicitly differentiate between these two constructs. Instead, in cases where 'gender' is intended to refer to biological sex, this is typically conveyed through the use of additional person-related information (e.g. *participant* gender, a *person's* gender), use of the term 'definitional gender' (which specifies biological sex based on linguistic definitions) or other terms to aid clarification (e.g. 'specific' gender).

Such gender inferences, based on stereotypicality biases, are one example of how language contributes to the maintenance and propagation of gender stereotypes in English. While grammatical gender languages can largely avoid gender stereotypic inferences by employing gender specific personal nouns to convey maleness and femaleness (e.g. *le musician*/*la musicienne* in French versus *the musician* in English), this is rarely possible in English. Instead, linguistic structures contain asymmetries that communicate evaluations and propagate stereotypes in many ways (Stahlberg et al., 2007), a number of which are outlined below.

1.2 Language and gender asymmetry

One gender asymmetry in language is that of the 'male as norm' phenomenon which, as the name suggests, is a tendency to equate males with the norm, or as the default gender category. For example, when researching stereotypes of national groups, Eagly and Kite (1987) found judgements about 'men' and 'people' to be more similar than those about 'women' and 'people', while Lambdin and colleagues reported that people also tend to attribute maleness to seemingly genderless soft toys (Lambdin, Greer, Jibotian, Wood, & Hamilton, 2003). Indeed much evidence now points to this 'male as norm' phenomenon with the conclusion drawn that people tend to implicitly represent social categories as male much more frequently than female when specific gender information is lacking (Hegarty & Buechel, 2006). This androcentric thinking arguably has greatest implications for languages that lack grammatical gender and rely on social gender to construct a mental representation of the sex of a referent. If the default representation of an ostensibly 'neutral' group (about which no gender information has been given) is that of male, explicit gender information may be required to instigate a re-categorisation. For example, when gender norms are deviated from, explicit formal markings are often used to indicate this e.g. *female* doctor or *male* nurse (Hellinger & Bußmann, 2001). These markings imply that such pairings are not typical and, as such, indirectly contribute to the reinforcement of gender stereotypes.

Another form of language asymmetry is that of lexical gaps. This is when a vocabulary term is lacking for one sex but exists for the other. This disparity is commonly found in occupational terms and titles as males traditionally occupied the majority of labour roles, for instance, the suffix 'man' is attached to numerous job titles e.g. postman, policeman. This gender bias also occurs in the opposite direction, as a male equivalent for typically female roles such as midwife was traditionally lacking (Stahlberg et al., 2007). Happily, in some cases, measures have been introduced to address these lexical gaps. For example, female counterparts to several

'masculine' terms have recently been introduced in response to a change in the distribution of the sexes in certain roles e.g. business woman, chairwoman.

With such gender asymmetries as those outlined above, it is clear that language is central to the construction and maintenance of the gender belief system of a given society (Stahlberg et al., 2007). This thesis investigates the deeply ingrained gender-biases associated with many social role nouns in English, with the aim of devising strategies to overcome the immediate activation of stereotypical gender information when reading such terms to ultimately result in lower levels of stereotype application. This goal of stereotype reduction is approached from a cognitive perspective, beginning with a brief account of how gender stereotypes are thought to develop.

1.3 The development of gender stereotypes

Gender is a particularly salient social category with infants able to readily discriminate between male and female faces (Fagan, 1976; Fagan & Sheperd, 1982; Fagan & Singer, 1979) and voices (Miller, 1983; Miller, Younger, & Morse, 1982) from as early as 7 months old. This early recognition of gender differences may explain why gender-related stereotypes are among the first to emerge (Lenton, Bruder, & Sedikides, 2009), with evidence of their development from about age 2 (Hill & Flom, 2007). These stereotypes are thought to stem from multiple sources such as socialising with parents and peers, learning through observation and the reinforcement of sex appropriate behaviours (e.g. Poulin-Dubois, Serbin, Eichstedt, Sen, & Beissel, 2002). Indeed, experimental evidence of gender stereotyping at a young age is relatively commonplace. For instance, children have been found to show a preference for toys that are consistent with gender stereotypes from 14 to 20 months old (O'Brien & Huston, 1985; Roopnarine, 1986) and can also match the faces and voices of young girls and boys with gender-typical toys (Serbin, Poulin-Dubois, Colburne, Sen, & Eichstedt, 2001). Similarly Serbin, Poulin-Dubois, and Eichstedt (2002) have found that 2 year olds spent longer looking at pictures of male actors than female actors performing a female-typical activity (such as applying lipstick); a finding which is interpreted as reflecting the children's surprise and interest at this atypical event. From this series of findings, Serbin and colleagues argue that by 24 months, toddlers have acquired an emerging knowledge of gender stereotyped behaviour although this knowledge is *typically* slower to develop in boys, emerging around 31 months (Poulin-Dubois et al., 2002).

Gender stereotyping at a young age can have important implications for later life as exposure to gender stereotypes can influence a child's preference towards certain jobs and activities. Liben, Bigler, and Krogh (2002) found that children aged 6-11 have quite fixed opinions about whether certain roles can be applied to women and men. These children were either given occupational labels (Study 1) or shown pictures of men and women working in a number of different occupations (Study 2), and asked whether each position could belong to women, men or both. The children gave responses in line with gender stereotypes, typically stating that doctors are men and nurses are women. These results suggest that children do not comprehend that job titles can refer to both sexes, and moreover, may have important implications for later career choice. For instance, Gottfredson's (1981, 2005) theory on career development asserts that children around 6 years old begin to lose interest in occupations that are not in line with their gender self-concept.

Research also indicates that children's gender stereotypes are more restrictive for boys than girls (Wilbourn & Kee, 2010), with the likelihood being that such stereotypes will affect choices made by males regarding education and careers e.g. men may potentially ignore nursing jobs as these are typically occupied by females and are considered lower status jobs. Indeed, data from July 2011 reveals that just 10% of nurses and 0.3% of midwives in the UK were males (UK Nursing and Midwifery Council, 2011).

In a similar vein, Correll (2004) investigated how cultural beliefs about sex differences can constrain the career-related aspirations of men and women (with the implication being that a person's aspirations will influence future career choice). Participants (first year undergraduates) were informed that they were being tested for the (fictitious) ability of ''contrast sensitivity'' – a skill said to be important for both graduate school and employment success. Before beginning the task, one group of participants were told that males typically outperformed females on this task (male advantage group) while another group were told that sex differences in performance were not previously found (control). In reality the task involved quickly judging whether the colours black or white dominated in a group of rectangles and had no obvious correct choice. Importantly, on completion of the task participants were each given the same score. However, it was found that male participants from the male advantage group rated their contrast sensitivity ability significantly higher than females from this group, while no such differences emerged in the control group. More importantly, males in the male advantage group rated themselves as significantly more likely to apply for courses and jobs that required a high level of contrast sensitivity than females in the same group. These findings

have important implications for prevalent gender stereotypes in society e.g. stereotypes about superior male to female math ability and superior female to male verbal abilities. Indeed, in a related task, Correll (2004) reports that (after controlling for actual grades) male students rated their own math ability higher than female students did and that self-assessments of task competence influenced career-relevant decisions. For instance, those who rated their math ability more highly were more likely to choose a college course in the domain of science, math and engineering; a finding that helps account for the lack of women in such quantitative professions.

In sum, past research suggests that gender stereotypes lead to inequality by artificially limiting the choices on offer to both sexes. As such, it is imperative to devise interventions that challenge people's gendered perceptions and ultimately lead to a reduction in gender stereotyping.

1.4 The cognitive processes underlying stereotype activation

1.4.1 Stereotypes: Background

A central aspect of theories of language acquisition is the process of naming, labelling and categorising people and things (Anglin, 1977). This process of categorisation can occur within milliseconds (Banaji & Hardin, 1996; Devine, 1989; Dovidio, Evans, & Tyler, 1986; Zarate & Smith, 1990) and is generally based on salient cues such as sex, race or age (McCann, Ostrom, Tyner, & Mitchell, 1985; Perdue & Gurtman, 1990; Stangor, Lynch, Duan, & Glas, 1992). Allport (1954) argues that categories refer to concepts, properties or objects that share common meaning or purpose, and that it is through this process of categorisation that stereotypes are conceived. When encountering objects we first select specific features that define the object, then accentuate those and interpret the object by generalising beyond the specific characteristics. This process of categorisation enables us to take meaning from situational objects by attending to a few diagnostic cues as opposed to registering all attributes of every object. Stereotypes are the outcome of this general process of categorisation (Operario & Fiske, 2004). However, while mental categories are enormously useful in terms of simplification of an information-rich environment, stereotypes are maladaptive forms of categories as their content does not always accurately correspond to what is present or happening in the environment (Bargh, 1999).

Since introduction of the term ‘stereotype’ to the field of psychology by Lippmann (1922) thousands of studies have examined the development and processing of stereotypes, while numerous review articles have assessed the conclusions drawn from this work (Blair, 2002; Lenton et al., 2009). It is now believed that to fully understand the complexity and universality of stereotypes, they must be analysed on multiple levels. From an interpersonal perspective stereotypes echo the relationship between groups, focusing on groups’ status and interdependence. At the societal level they imitate the broader context of group life while at the cognitive level stereotypes are conceived of as functional mechanisms for considering the world (Operario & Fiske, 2004). As mentioned earlier, it is from a cognitive perspective that the processes involved in stereotype use will be examined in this thesis.

In his formative work on stereotyping, Lippmann (1922) described them as ‘pictures in our heads’ that simplify how we think about other social groups. Indeed, even at this early stage of stereotype research he recognised their functional value, reasoning that the process of stereotyping is “a fundamental human mechanism for perceiving and making sense of the world” (p. 129). While many varied definitions of stereotyping have since been proposed, that of Hamilton, Sherman, and Ruvolo (1990) is considered particularly relevant to this research. They define stereotypes as a particular type of expectancy, that may influence information processing through focusing attention on, and facilitating processing of, stereotype-consistent information or inhibiting stereotype-inconsistent information (Bruner, 1957; Olson, Roese, & Zanna, 1996). This definition fits well with the central thesis of this work, which focuses on how stereotypic expectancies influence language processing and how such expectancies may be overcome. Before reviewing the literature on stereotypes in language processing, further information about the role of automatic and controlled processes in information processing will first be discussed.

1.4.2 Automatic and controlled processes in information processing

Schneider and Shiffrin (1977) proposed a two-process theory of information processing suggesting that automatic and controlled processes are qualitatively different mechanisms that play important roles in attention. They describe automatic processing as the activation of a learned sequence of elements in long-term memory that is spontaneously carried out following activation of appropriate input(s). This automatic processing occurs without the control - or often the attention - of the perceiver and once established, is difficult to overcome. Indeed, Schneider and Shiffrin theorised that as these processes operate via a

relatively enduring set of associative connections situated in long-term memory, the development of new automatic processes would require a considerable amount of consistent training to fully develop. In contrast, these authors describe controlled processing as the activation of a temporary sequence of nodes that are triggered through the attention of the perceiver. Such processes are tightly capacity-limited (as they involve the short-term memory) but have the advantage that they are easy to set up, can be modified and applied in new contexts for which automatic sequences have not been learned. Indeed, given these advantages, controlled processes arguably hold the most potential in terms of stereotype reduction strategies.

Returning to the process of categorisation, the authors claim that category encoding must necessarily be an automatic process, otherwise a search through long-term memory would always be required in order to identify a category, thus wiping out any time processing advantages that the process of categorisation is thought to confer on a perceiver. Indeed, the claim that stereotyping (as a form of categorisation) can occur automatically, and that such automatic processes are deeply ingrained, difficult to overcome, and possibly operate without the perceiver's awareness, led to the conceptualisation of stereotypes as fixed, stable and enduring cognitive structures, a position that did not change until research published by Devine (1989) suggested that automatic stereotyping could indeed be moderated.

In a subliminal priming study, Devine (1989, Experiment 2) presented participants with either 20% or 80% stereotype-related primes about the social group of 'Black Americans'. The participants' task was to identify the position of these primes as they randomly appeared on-screen in one of four quadrants. This task was followed by an ostensibly unrelated measure, in which participants read a paragraph about a male character called 'Donald' performing ambiguously hostile actions (e.g. refusing to pay rent until his apartment is repainted) and then rated his traits along 12 dimensions. Devine discovered that those who were presented with 80% stereotype-related primes subsequently interpreted the ambiguously hostile behaviour of the character as more hostile; a finding which is of particular interest as no hostility-related traits were used as primes. This pattern of results occurred for both high and low prejudice participants (as measured on the Modern Racism Scale, McConahay, Hardee, & Batts, 1981) and is taken as evidence that when participants do not have the chance to consciously monitor their stereotype activation, they will produce stereotype congruent responses - regardless of their prejudice level.

However, in a subsequent experiment (Experiment 3, Devine, 1989), a group of high and low prejudice participants were asked to list all of their thoughts about the social group of 'Black Americans'. Participants 'low' in prejudice apparently censored and inhibited automatically activated negative stereotypes by replacing them with thoughts that conveyed their non-prejudiced values whereas 'high' prejudice participants did not achieve this[2]. Following this evidence that activated, *explicit* stereotypes could be controlled using social desirability strategies, further research endeavoured to ascertain whether *implicit* stereotyping was also amenable to change under certain conditions. This led to an upsurge of interest in the malleability of implicit stereotypes and the development of new theoretical models of stereotype representation to explain this malleability.

1.4.3 Models of stereotype representation

Further to Schneider and Shiffrin's two-process theory of information processing outlined above, Bargh (1999) claimed that over the years researchers have studied 2 main types of mental processes under a variety of names – controlled/automatic, conscious/nonconscious, explicit/implicit. Whatever the label, traditional models of stereotyping stipulate distinct roles for these two processes in stereotyping (see Bargh, 1999; Bodenhausen & Macrae, 1998; Brewer, 1988; Devine, 1989; Fiske & Neuberg, 1990). These models suggest that stereotypes begin with the activation of implicit stereotypes, followed by their application to judgement or behaviour. If motivated to do so, a perceiver may evade *explicit* stereotyping by controlling the application of stereotypes through strategies such as (1) suppressing stereotypic information, (2) compensating for the stereotype's influence by changing responding accordingly or (3) focusing on individuating information (Devine & Monteith, 1999; Dunton & Fazio, 1997; Fazio, Jackson, Dunton, & Williams, 1995; Fiske & Neuberg, 1989; Plant & Devine, 1998). However, while traditional models suggested that activation of *implicit* stereotypes could not be altered, newer models make different predictions about the flexibility of implicit stereotypes. Four different social cognitive models relating to how category information is stored in the mind and how each model predicts stereotype reduction may be possible will be briefly outlined – prototype models, exemplar models, associative networks and connectionist models.

Both prototype and exemplar based models of stereotype representation consider stereotypes to be long-term, stable cognitive structures that are not easily amenable to change across time

[2] Although there is no evidence that this particular sample of low prejudice participants had initially activated negative stereotypes, the results of Experiment 2 suggest this activation was highly likely as, participants had shown evidence of activating racial stereotypes, regardless of whether they had previously scored high or low on a prejudice scale.

or situations (Bargh, 1999; Hamilton & Sherman, 1994). Prototype models posit that people arrange category information around the most typical member, or the statistical average, of that category. They suggest that categorical information is represented in 'fuzzy sets' such that characteristics of the category have no exact boundaries, just simple association with the prototype. In short, these models suggest that modifying perceivers' thoughts about typical category members can weaken or change the stereotype (Hantzi, 1995).

On the other hand, exemplar-based models place emphasis on the role of concrete examples in mental representations (Medin & Schaffer, 1978) and result from actual experience with category members (Carlston & Smith, 1996). These models posit that mental representations entail variability, with perceivers capable of holding numerous distinct exemplars of one particular category in mind (Linville, Fisher & Salovey, 1989), and thereby accommodate the idea of creating mental representations of subgroups of larger categories. These subgroups recognize within-group differences and hint at the potential for stereotype change through the accrual of sufficient group variability (Maurer, Park, & Rothbart, 1995; Rothbart, 1996). Similarly, this model posits that stereotypes could abate when targets contrast with the exemplar that was chosen as a frame of reference by the perceiver (Stapel & Koomen, 1998).

The most widespread models of human memory are associative network models (Anderson & Bower, 1973; Collins & Loftus, 1975; Wyer & Carlston, 1979) . These focus on the patterns and strength of associative links among concepts in memory, with information theorised to be stored in distinct cognitive structures called nodes. Associative models are said to use localist representations as each node relates to a singular concept such as a name, object, personality trait, or evaluation – essentially any form of raw data (Carlston, 1994, 2010). These nodes are interconnected by links that reflect associations between the concepts of each node and ultimately constitute a person's mental representations, with strong links reflecting significant associations between concepts, while weaker links reflect less important associations. The majority of nodes are thought to lie inactive in long-term memory with a small portion active at any one time, shaping conscious or unconscious cognition (Carlston & Smith, 1996) through spreading activation (Collins & Loftus, 1975). This process of spreading activation suggests that excitation of one node travels rapidly across strong links, thus stimulating other nodes and triggering associations between connected concepts.

Associative models support findings from the stereotype literature, predominantly semantic priming studies (see Neely, 1991) that assess response times and judgements to a category target after presentation of a prime (e.g. quickly identifying the term 'manager' following

presentation of the prime ‘male’). Many well-known studies of stereotyping (Devine, 1989; Higgins, Bargh, & Lombardi, 1985; Perdue, Dovidio, Gurtman, & Tyler, 1990) investigate semantic priming; an approach proposing that stereotypes reflect the strength between 2 or more mental nodes. With this emphasis on micro-level mental structures, associative models propose that stereotyping via connected nodes takes place outside of the perceiver’s awareness (see also Banaji & Hardin, 1996 for a discussion). However, they outline the potential for stereotype change through the perceivers’ experiences (e.g. in terms of gender stereotyping, by strengthening connections between counter-stereotypical targets and the relevant concept of ‘male’ or ‘female’) and highlight the possibility of creating new nodal links following the association of previously unconnected concepts (Operario & Fiske, 2004).

While the idea of spreading activation (and spreading inhibition) incited a lot of interest and support in psychology, many theorists have now abandoned associative network models and adopted an approach that treats mental representations as weighted combinations of a fixed set of features (for a review, see Rumelhart, Smolensky, McClelland, & Hinton, 1986; Smolensky, 1988), in what is termed a connectionist approach to mental representations, sometimes referred to as parallel distributed processing (Carlston, 2010).

Connectionist models posit that memory comprises a wide network of feature units (like nodes) that share activation across weighted interconnections (like pathways). Unlike associative network models, representations are not conceived as nodes situated in the memory network. Instead, connectionist models use distributed representations as meaning is thought to stem from patterns of activation across units that create combinations of impulses (Operario & Fiske, 2004; Smith, 1996). These patterns of activation are thought to satisfy constraints from two sources; those imposed by current input that represent the immediate situation, and connection weights that develop and update over time according to the stimuli an individual encounters, thus representing long-term learning. Connectionist models propose that following recurrent stimulus exposure or experience, activation of a few input units could lead to completion of a previously learned pattern. Also, numerous patterns of activation can occur at once, mirroring people’s ability for parallel on-line cognitions (Carlston, 2010).

While few attempts have been made to apply connectionist models to stereotyping phenomena, a trend has developed focusing on the facilitatory and inhibitory mechanisms underlying stereotypic thought (see Bodenhausen & Macrae, 1998 for a review). Guided by the concept of positive and negative unit connections, research has recently focused on the variables that may weaken stereotypes (e.g. Macrae, Bodenhausen, & Milne, 1998; Macrae,

Bodenhausen, Milne, & Ford, 1997; Monteith, 1993) as opposed to those that lead to stereotyping. Another trend stemming from these models is the focus on constraint-satisfaction processes (Kunda & Thagard, 1996; Miller & Read, 1991) whereby activated units (such as individuating traits or personal goals) can oppose the powerful effect of other units (such as categorical beliefs) in impression formation or judgement (Operario & Fiske, 2004).

Overall, although the models outlined above overlap to a large extent, each also offers a unique theory as to how stereotype activation is sustained and how it may be overcome. These models helped to inform the selection of stereotype reduction strategies used throughout this thesis and provide a useful framework for interpreting how stereotype change may have occurred. However, before outlining strategies used to overcome the activation and application of gender stereotyping, the problems that gender-biased expectancies can generate in text comprehension are first considered below.

1.5 How gender stereotypes affect language processing in text comprehension

1.5.1 Gender inferences

It is now well established that readers use both implicit and explicit elements of a text to construct a mental representation of the information therein (e.g. Carreiras, Garnham, Oakhill, & Cain, 1996; Garnham & Oakhill, 1996; Johnson-Laird, 1983; Zwaan & Radvansky, 1998). Such mental representations are commonly referred to as mental models and, according to advocates of a constructionist approach to creating mental representations, may contain information about people such as their goals, emotions, physical traits, attitudes and beliefs as well as spatial and location information (Carreiras et al., 1996; Zwaan & Radvansky, 1998). The implicit elements of a text used in the construction of a mental model are called inferences and they stem from a combination of the text and the readers' general knowledge (Graesser, Singer, & Trabasso, 1994; McKoon & Ratcliff, 1992).

One particular form of inference frequently employed by English speakers is that of assuming (correctly or incorrectly) a person's gender. For example, in cases where explicit gender information about a character is lacking, perhaps following the use of a role noun (e.g. nurse, carpenter), prior knowledge in the form of a gender stereotype may be employed to supply a default gender (Carreiras et al., 1996). This category of role nouns provides an important means of investigating the role of gender information in inference-making when gender

activation is purely semantic or conceptual as opposed to morphosyntactic (Reynolds, Garnham, & Oakhill, 2006).

Indeed, despite the majority of English role nouns being gender indefinite, considerable evidence now shows that when readers encounter an occupational role noun (e.g. beautician) they *do* infer a specific gender (e.g. female) (Carreiras et al., 1996; Garnham, Oakhill, & Reynolds, 2002; Kreiner, Sturt, & Garrod, 2008; Oakhill, Garnham, & Reynolds, 2005). This only becomes a problem in text comprehension when information is subsequently encountered that is incongruent with the stereotype bias of the role name. For example, if the term 'surgeon' is followed by the pronoun 'she' in a text, does this cause the reader difficulty? A match-mismatch paradigm such as this has typically been employed by researchers to investigate gender processing and the answer is a resounding 'yes' (Carreiras et al., 1996; Garnham et al., 2002; Irmen, 2007; Kreiner et al., 2008; Oakhill et al., 2005; Reynolds et al., 2006). This paradigm presents participants with occupational role nouns followed by either gender consistent or gender inconsistent information. Unfailingly, processing difficulty is evident in the mismatch condition relative to the match condition - typically conveyed via slower judgement or reading times. Such findings are thought to reflect difficulty in integrating the unexpected gender information into the reader's mental model (Sato, Gygax & Gabriel, 2013).

1.5.2 When are gender stereotypes activated?

The question of *when* gender stereotypes are activated in online processing has received a great deal of research attention (see e.g., Duffy & Keir, 2004; Irmen, 2007; Reynolds et al., 2006). Much evidence currently supports the idea that stereotypes affect language comprehension in a backward manner, while resolving anaphors that refer back to a stereotyped role name (e.g. builder…s/he). Evidence of such backward or bridging inferences (Clark & Haviland, 1977) has been found in text processing whether investigated through self-paced reading tasks (e.g. Carreiras et al., 1996; Kennison & Trofe, 2003), eye-tracking (e.g. Duffy & Keir, 2004; Irmen, 2007; Kreiner et al., 2008) or electrophysiological studies (e.g. Osterhout, Bersick, & McLaughlin, 1997).

For instance, in the latter study by Osterhout and colleagues (1997), event-related brain potentials[3] (ERPs) of participants were recorded while they were presented with sentences

[3] Event related potentials measure voltage differences of electric impulses on the scalp that result from a specific event such as presentation of a target word in a sentence.

containing either a definitionally (e.g. mother) or stereotypically gendered role noun (e.g. beautician), followed by a reflexive pronoun of matching or mismatching gender (e.g. herself /himself). While the authors anticipated a negative deflection in the brain waves peaking at about 400ms following onset of a stereotyped role noun in the mismatch condition (an N400 effect), commonly thought to indicate a semantic anomaly (e.g. see Kutas & Van Petten, 1994), they instead found a positive peak in the brain waves 600ms following the target onset (P600 effect), commonly thought to indicate a syntactic anomaly[4] (e.g. Hagoort, Brown, & Groothusen, 1993; Osterhout et al., 1997; Osterhout & Mobley, 1995). These findings (combined with those of numerous reading studies) are frequently claimed to provide evidence that gender information is encoded in gender stereotyped role names, and subsequently used in backward inferencing. However, Pyykkönen and colleagues argue that typical match-mismatch methodologies do not allow clear investigation of the time course of stereotype processing as past results do not preclude the idea that stereotype biases are also used in a forward, elaborative manner (i.e. when they are not literally stated in the text, or needed to establish or increase text coherence) (Pyykkönen, Hyönä, & van Gompel, 2010). Indeed past studies claim that stereotype information is activated elaboratively as soon as a stereotype biased role noun is read (e.g. Garnham, 2001; Oakhill et al., 2005; Reynolds et al., 2006), yet does not cause processing difficulty until efforts are made to assimilate this information into a mental model of the discourse. Similarly, Pyykkönen and colleagues claim that plausibility judgement experiments, which suggest elaborative activation of gender stereotypes (e.g. Oakhill, et al., 2005), are not ideal as they to do not necessarily reflect normal text processing, and potentially encouraging readers to activate information they would not usually activate.

Consequently, Pyykkönen et al. (2010) sought to more definitively distinguish between elaborative activation of stereotyped information and use of this information for bridging inferences using a methodological approach that allows for more exact investigation of elaborative inferences. They conducted a visual world eye-tracking study to explore activation of information associated with gender-stereotypical role names in online language comprehension in Finnish. This paradigm was chosen as it permitted investigation of stereotype activation in cases where it does not assist text coherence i.e. before any pronoun and in a genderless language (Finnish). Participants were instructed to listen to stories and

[4] Subsequent research has argued that the ERP component found with this gender mismatch effect is actually reflective of a violation of syntactic preference (Hagoort & Brown, 1999; Van Berkum, Brown, & Hagoort, 1999). This occurs when sentences are syntactically well constructed, but where syntactic elements do not satisfy the current, preferred analysis of that sentence (Pyykkönen, 2009).

follow with their gaze the four pictures displayed on-screen, in the order in which they were mentioned in the story. It was found that when listeners heard a male biased noun such as 'chimney sweep' they were more likely to look at a male character than a female character (and vice versa for female stereotyped nouns). These results suggest that stereotype information was activated elaboratively during online discourse processing, despite not referring to any particular male or being useful for discourse coherence.

Pyykkönen et al. argue that these findings sit well with the mental models theory of text comprehension mentioned above, which posits that people make elaborative inferences in order to create a detailed representation of a particular situation. However, they sit less well with the minimalist account of text comprehension (McKoon & Ratcliff, 1986, 1992) which posits that inferences unnecessary for local coherence or not based on "easily available" information are not made online. If this easily available information is restricted to mean explicitly stated information, and propositions derived from them (as is posited by McKoon & Ratcliff, 1992), then activation of gender stereotypes is not predicted under this minimalist account unless it aids text coherence[5]. Overall, Pyykkönen and colleagues conclude that a processor may activate stereotype information upon encountering a stereotyped term in order to update and modify the mental model in an incremental manner during online processing of discourse (see also Garnham, 2001; Johnson-Laird, 1983; Zwaan & Radvansky, 1998).

This question of when stereotype information is activated is of great relevance to this thesis as a priming paradigm (described in Section 1.5.4) is used in which participants are presented with a single stereotyped role term along with a definitionally gendered word (e.g. nurse/ mother). Upon presentation of the role name it is posited that gender is elaboratively inferred by the reader.

1.5.3 How persistent is the stereotyping effect?

Another question of importance to the current research is that of how stable or fixed stereotyping effects are. Reynolds and colleagues (2006) investigated this issue by giving participants a slightly adapted version of the now well known 'surgeon riddle' of Sanford

[5] However, note that a distinction between mental models and minimalist accounts of text comprehension are not so straightforward in terms of elaborative inferences. For instance, while Garnham (1992) posits that many mental model theorists have emphasised the role of on-line elaborative inference making in text comprehension he also concedes that such on-line elaboration is not an obligatory part of a mental model theory of text comprehension. Indeed, in line with the minimalist approach, he states that some mental model theorists may hold the opinion that if information in the text is not readily available, an inference will not be made.

(1985). In this riddle, a father and son are involved in a car accident where the father dies but the son is taken to hospital for an operation. However, once there, the surgeon looks at the boy and exclaims “Oh my god, that is my son!” (Sanford, 1985: “I can’t do this operation. This boy is my son.”).The majority of readers have difficulty solving this riddle, becoming confused by an apparent inconsistency stemming from gender values they have attributed to the characters. Typically they infer that the surgeon is male and, despite knowing that the boy’s father is dead, fail to override this inference and successfully conclude that the surgeon is the boy’s mother. Indeed, the authors report that 75% of readers who had not previously seen the text, failed to resolve the inconsistency and update the gender of the surgeon in their mental representation of the text.

Across a series of studies in which Reynolds et al. varied aspects of this riddle, their findings strongly indicate that gender is elaboratively activated once a role name is encountered, as the surgeon stereotype of ‘male’ is not overcome even though assigning ‘female’ to the surgeon is the ‘simplest’ way of making sense of the passage. A different set of participants subsequently showed no difficulty in solving the riddle when the term surgeon was replaced by the term nurse. These findings with the surgeon riddle (Experiment 1A) were followed up using a self-paced reading task, again using either the original surgeon riddle or a version with the term nurse as opposed to surgeon (Experiment 2). Reading times for the final clause (which contained the gender-biased role noun) were found to be 1,000ms slower on the original riddle than the stereotype consistent version. This processing delay is much longer than past experiments that involve minor accommodations to mental representations suggest is typical i.e. approximately 200ms (e.g. Carreiras et al., 1996, Experiment 1; Haviland & Clark, 1974). Similarly, when asked after the passage whether the surgeon/nurse was the boy’s mother, response times were on average one and a half seconds longer for participants who read the original surgeon riddle compared to those who read the adapted ‘nurse’ version.

Both the offline and online evidence led Reynolds et al. to conclude that inferring gender from stereotyped role-names is at least in some part an automatic process, with the experiments highlighting how entrenched such gender inferences are and the substantial processing difficulties that they induce.

Another piece of research, conducted by Dunning and Sherman (1997), hints at the subtle pervasiveness of stereotypes and implies how difficult they may be to overcome. Across 5 experiments these authors investigated whether stereotypes lead people to make tacit inferences about individuating information i.e. inferences that modify the meaning of

information in order to affirm a stereotype. For example, reading a sentence such as “some felt that the politician’s statements were untrue” may lead people to make the tacit inference that the politician was lying, based on stereotypes about this occupational group. However this inference would be deemed less likely if the term ‘physicist’ replaced that of ‘politician’ in the text. To investigate this, the authors created sentences describing scenarios that could refer to either of two social groups, presenting one version to each participant e.g. “After a few drinks, the two (marriage counsellors/ lumberjacks) had a fight in the restaurant”. Then, for each role name a stereotype-consistent interpretation was generated (e.g. a sentence stating the marriage counsellors had a ‘quarrel’ while the lumberjacks had a ‘fist fight’). Participants were given these sentences and asked to decide whether they were new or had been previously presented. Results revealed that participants did indeed make tacit inferences, falsely recalling sentences that were consistent with stereotypes (e.g. more frequently recalling that the lumberjacks had a fist-fight as opposed to the marriage counsellors having a quarrel). Dunning and Sherman conclude that interpretations of behaviour influenced by role noun stereotypes (and interpretation of individuating information more generally) may be shaped by tacit inferences and result in impressions of others’ behaviour that do not accurately represent the reality of the situation. Moreover, in Experiment 5 it was revealed that both low and high sexist participants (as assessed by the Modern Sexism Scale; Swim, Aikin, Hall, & Hunter, 1995) made very similar levels of tacit inferences. The authors argue that if production of these inferences was under the individuals’ control, it would be more likely that the low sexist group would reveal lower levels of tacit inference production than their high sexist counterparts. Indeed the fact that perceivers seems to make these tacit inferences without conscious intent suggests that the problem of stereotyping is a particularly pervasive one. Overall, this series of experiments by Dunning and Sherman is yet further evidence of the immediate, subtle and persistent influence of gender stereotypes in language processing.

1.5.4 Overcoming stereotyped gender biases in text

Despite evidence documenting the deeply ingrained and pervasive nature of stereotypes, some researchers have attempted to reduce the spontaneous activation of gender stereotype biases.

Oakhill et al. (2005) were interested in whether gender biases are evoked for single words (as opposed to when they are in a sentence/short passage context) and the extent to which such bias information can be overcome. They devised a task in which participants were presented

with two terms: a role name (that was definitionally gendered or stereotype biased e.g. princess, beautician respectively) and a kinship term (with a specific, definitional gender e.g. uncle, aunt). Participants were then asked to quickly decide whether or not both terms can be used to refer to one person. The task required participants to link the two terms they had read. To perform well they needed to take definitional gender into account (e.g. that a mother is *always* female) but suppress stereotypical gender (e.g. that *most* beauticians are female).

Oakhill et al. reasoned that if the task involves an automatic component then participants should have difficulty suppressing the gender bias associated with the presented role noun, and consequently give a stereotype biased response. However, if there is a strategic element to the task then it should be possible to reduce or eliminate this stereotyping effect. With this basic judgement task, they conducted 6 experiments that varied stimuli presentation and instruction details. Across all experiments, results revealed that performance was mediated by the stereotype bias of the role nouns. For example, participants more frequently rejected word pairs in which the gender associated with the role noun was incongruent with the gender of the kinship term (e.g. electrician, mother) as opposed to congruent (e.g. electrician, father). This effect was still evident (although to a lesser extent) once participants were explicitly reminded that nowadays many jobs are not clearly marked for gender and that they should carefully consider whether the first term presented (i.e. the role noun) could be occupied by a man, a woman or either gender (Experiment 4). The authors conclude that their results provide strong evidence for an automatic component to responding and support the claim that stereotypical gender information associated with role nouns is typically incorporated into a mental representation of a person "immediately (and probably automatically)" (p. 982) – even when it is counter-productive to task performance. This experimental paradigm of Oakhill et al. was central to the current research as, with its use, the efficacy of different strategies aimed at overcoming the immediate activation of gender stereotypes so as to ultimately reduce levels of stereotype application was evaluated (Chapters 2-4).

But how is the processing of individual words influenced by the discourse context? Duffy and Keir (2004) set out to investigate this question in a set of two eye-tracking studies.

In Experiment 1 they examined whether or not violations of gender stereotypes led to a disruption of the reading process by measuring participants' eye movements on sentences which introduced a character using a stereotyped role name, followed immediately by a verb and a reflexive pronoun (target) that matched or mismatched the gender of the role name e.g. "The secretary treated herself/himself to a large sundae after finishing work". Duffy and Keir

anticipated that a stereotype mismatch effect would be evident on fixation times on the reflexive pronoun and the following region (of 1-4 words). Such an effect was indeed found with interference from the role noun first evident on the reflexive pronoun, followed by significantly slower go past[6] reading times for the following region as participants registered the mismatch and displayed a tendency to re-read earlier parts of the text before proceeding.

Experiment 2 then investigated whether a prior context, that explicitly specified the sex of the character introduced by a role name, would lessen this stereotyping effect. They constructed short paragraphs in which a role noun was mentioned twice, but with the reflexive pronoun placed after the second mention. For example, "The electrician was a cautious [woman/man] who carefully secured her/his ladder to the side of the house before checking the roof…the electrician taught *herself/himself* a lot while fixing the problem". Results indicated that this disambiguating context was enough to overcome the mismatch effect on the reflexive pronoun (participants who did not receive this explicit disambiguating information continued to show evidence of succumbing to the stereotype biases). Duffy and Keir conclude that their results are in line with past studies that have proposed gender stereotypes are automatically activated when certain role names are read, and that these stereotypes then have an impact on subsequent processing (e.g. Banaji & Hardin, 1996). Overall, although gender is inferred even when it has not been explicitly mentioned, it is encouraging to note that this effect can be overridden, at least temporarily, following explicit mention of a person's sex earlier in the text.

In a self-paced reading task, Lassonde and O'Brien (2013) found similar beneficial effects of explicitly mentioning the sex of a referent as a means of overcoming gender biases in text. In contrast to Duffy and Keir (2004) these authors examined whether gender neutral terms (e.g. firefighter), which are often used to replace male-biased terms (e.g. fireman), carry an implicit male bias. They constructed short paragraphs with a social role noun introduced in one of these two formats (i.e. male-biased or neutral), followed by a target sentence with a gender-specific pronoun (he/she). In Experiment 1 these target sentences were read more slowly when a female pronoun was used after both male-specific and gender neutral role names (as opposed to when a male pronoun was used), thus hinting that 'neutral' terms do indeed have an inherent male bias. However, with the addition of a sentence explicitly specifying the occupational character as a woman, the reading disruption previously found when integrating

[6] I.e. A measurement starting from the first fixation in the region until the final fixation before going on to fixate a later part of the sentence.

the gendered pronoun with the neutral role term was now eliminated (Experiment 2). This finding suggests that, when used in this way, gender neutral language can moderate against activation of gender stereotypes.

A final strategy for overcoming stereotype activation in discourse was identified by Kreiner et al. (2008) in a sentence reading task. In the first of their two eye-tracking studies, these authors again observed a processing cost with the integration of a reflexive pronoun with a stereotyped role noun. Indeed, they found a similar mismatch-cost whether using a role noun with stereotypical or definitional gender information.

However, in Experiment 2, participants were presented with sentences in which the reflexive pronoun preceded the stereotyped term e.g. After reminding *himself / herself* about the letter, the *minister* immediately went to the meeting at the office (i.e. the pronoun was cataphoric). In this experiment, a mismatch-cost was only found for nouns with definitional gender, suggesting that gender stereotype information can again be overcome by specifying a character's sex in prior discourse (as previously reported by Duffy and Keir, 2004).

Overall, a modest amount of research has been directed at overcoming the activation of gender stereotypes associated with role nouns, but with only moderate success achieved. While stereotype biases have been successfully attenuated, they have not been completely eradicated. Some of the challenges facing stereotype reduction will be outlined below, along with further strategies aimed at reducing gender stereotyping outside of the domain of text comprehension. It was hoped that information from this literature would inform the formulation of stereotype reduction strategies to be tested in this thesis.

1.6 Overcoming stereotypes

1.6.1 Disadvantages of stereotyping

Part of the reason for persistent stereotype use may be the fact that stereotypes *are* often accurate. For instance, the gender stereotype that men are better at math than women is often reflected in exam scores[7]. On the whole, it appears that perceivers use stereotypes as a

[7] For example, males have been found to outperform females in math ability at high school and college level (e.g. Hyde, Fennema, & Lamon, 1990). However, such differences were not found at earlier ages or when participant samples were chosen from the general population (thus hinting that it is indeed math ability stereotypes that were likely to be a factor in the differential performance found at school level).

cognitive shortcut when building mental representations of a situation, leading Gilbert and Hixon (1991) to aptly describe them as 'a sluggard's best friend' (p. 509).

However, despite the cognitive economies gained through stereotype use, Lippmann (1922) argued that stereotypes are also accompanied by disadvantages. He claimed that they obscure reality and misrepresent genuine experience with biased preconception, succinctly summarising the problem of every-day stereotype use by saying that "for the most part we do not first see, and then define, we define first and then see" (1922, p.81). Indeed, unregulated use of stereotypes can lead to over-generalisation, misattribution and condemnation of the personal traits and behaviours linked with particular social categories (Operario & Fiske, 2004). While the necessity to overcome stereotyping is apparent, many obstacles have been identified that stand in the way of achieving this.

1.6.2 The challenge of overcoming stereotypes

A challenge facing researchers interested in combating stereotype use is that, once a target has been categorised, various cognitive biases then work to facilitate and protect that categorisation. For example, perceivers tend to underrate differences between a category and target member (Taylor, 1981), automatically assign stereotype-consistent attributes to a target (Devine, 1989; Dovidio et al., 1986) and attend more to category-consistent information than inconsistent (Hamilton et al., 1990).

Stereotypes also benefit from informational ambiguity as, when faced with information irrelevant to the category, perceivers may still interpret it based on category expectations (Hilton & von Hippel, 1990; Nelson, Biernat, & Manis, 1990). However, perhaps more significant is that when faced with category disconfirming (i.e. counter-stereotype) information about a target, perceivers are likely to interpret this information as unrepresentative of the category in question (Krueger & Rothbart, 1990; Kunda & Oleson, 1995). For example, if a female target is perceived as a good athlete, this is very unlikely to change an individual's broader stereotype about women as being poorer at sports than men[8].

Overall, it seems that perceivers assimilate information into their established beliefs about a category, a theory which is in line with Fiske and Neuberg's (1990) idea of confirmatory

[8] However, it is worth noting that many such encounters may lead to subtyping processes and the development of a new category to account for talented female athletes. This issue is further discussed in Section 1.6.4.

categorisation. This theory posits that a perceiver's categorical representations and prior beliefs are kept intact through selective information searches. Thus stereotypes seem resistant to irrelevant and incongruent information as perceivers creatively explain away discrepant information. Indeed, meta-analyses of past studies suggest that perceivers are unlikely to recall category inconsistent information except under very specific circumstances (Stangor & McMillan, 1992) such as (1) when perceivers have weak expectancies about target members (because of little experience with their social group), (2) incongruities are strong between a target member and category and (3) when perceivers have impression-formation goals (i.e. they have a vested interest in the target and so pay closer attention to relevant individuating information). Aside from these cognitive biases that challenge the reduction of stereotypes, it is important to be cognisant of other factors that are likely to increase stereotype use.

1.6.3 Factors affecting stereotype use

One factor observed to affect stereotype activation is that of cognitive busyness (Gilbert & Hixon, 1991) i.e. the extent of our mental preoccupation with concerns other than the immediate situation. Given that stereotypes act as cognitive shortcuts allowing perceivers to efficiently categorise people based on probabilities, it is reasoned that the more distracted or cognitively busy a person is, the more they will rely on stereotypes to make judgements.

Macrae, Hewstone, and Griffiths (1993) investigated this theory, by showing participants a video of a woman (described as either a doctor or a hairdresser to participants) talking about her lifestyle and interests. Some participants were also given a distracting task to perform while watching the video (recalling an 8-digit number). It was observed that participants who completed this additional mental task recalled more items based on occupational stereotypes, and rated the woman in more stereotypical terms than those who did not partake in the distraction task.

However, Gilbert and Hixon (1991) state that cognitive busyness only results in increased stereotyping if a relevant category has actually been engaged with. In contrast to Macrae et al., these authors used cognitive distraction as a means of preventing stereotype activation in the first place. In their study, an Asian or Caucasian woman held cards upon which word fragments were written (e.g. POLI-E). Participants were required to think of as many words as possible that would fit the fragment. However, half the participants were first given an 8-digit number to recall throughout the word-fragment task (distracted condition). Results revealed that the non-distracted participants came up with more words associated with the stereotype

of Asians (e.g. polite) in the presence of the Asian woman compared to when in the presence of the Caucasian woman. However, the ethnicity of the confederate revealed no effect on the responses of distracted participants. It was therefore claimed that although category membership was known to the distracted participants (Asian vs. Caucasian), they were too preoccupied to activate associated stereotypic content.

Although these results are promising as regards the prevention of stereotype activation, situations are rare in which stereotypes and their associated content are not immediately activated when relevant social cues are present.

Another factor thought to influence stereotype use, identified by Kawakami, Dovidio, and van Kamp (2007), is that of the 'heavy handedness' of a training. They hypothesise that when an intervention is very apparent, people may attempt to alter their responses to compensate for such obvious attempts at manipulating their behaviour. In their study, Kawakami et al. asked participants to act as 'hirers' and choose a job candidate from among 4 profiles. Materials ensured that each candidate was suitable for the advertised position, the key difference being that 2 were given male names and 2 were given female names. However, before this task half of the participants first took part in a counter-stereotypic association training. This involved presenting participants with photographs of a male or a female, with two traits printed underneath (e.g. sensitive/strong). The participant's task was to identify which of the two traits is not typically associated with the sex of the person in the photograph (e.g. 'strong' if the photograph depicted a woman, 'sensitive' if it depicted a man). It was found that the choice of job candidate depended on whether participants completed the negation training and whether the hiring decision directly followed this stereotype negation training or whether participants first completed another rating task in which they had to judge the applicants on 16 job-related traits. Those who did not receive negation training consistently chose male over female candidates. However, those who did receive the negation training more frequently chose a female candidate over a male - but only when participants completed the additional rating task.

The authors argue that participants tried to influence their reactions in a systematic way in the first task after the negation training – regardless of whether it was the candidate selection task or the trait rating task. It seems that, if a stereotyping task follows a counter-stereotypic training, participants may assume that their judgements of group members are being influenced. Consequently, participants can attempt to modify their judgements in a more *biased* direction so as to compensate for this influence. Once this has been done they may not

feel the need to correct their responses in a second stereotyping task. This reasoning is in line with theories of mental correction (Wegener & Petty, 1997; Wilson & Brekke, 1994) which stipulate that if an individual is both motivated and capable, they may change their assessment to the opposite of the perceived bias (Mussweiler & Neumann, 2000). Similarly reactance theory (Brehm & Brehm, 1981), states that if people feel they are being overly influenced, they may try to oppose this effect. Reactance and correction processes are thought to differ fundamentally in that reactance theory suggests that people will react without much reflection against forces that limit their sense of behaviour freedom, while correction theory posits that people assess the potential influence in more detail and adjust their responses to compensate for its effects (Mussweiler & Neumann, 2000; Wegener & Petty, 1997; Wilson & Brekke, 1994). However, whichever theory is supported, both highlight important issues to consider before devising a stereotype reduction training.

Other factors found to influence stereotype use include (1) Emotional arousal e.g. Stroessner, Hamilton and Mackie (1992) found that being upset or anxious increases the likelihood of a person reverting to familiar stereotypes in social perceptions, and (2) the extent to which an individual and the target of their judgement are deemed to be positively interdependent (Neuberg & Fiske, 1987). Essentially, if one person depends on another for achievement of a particular goal, they are likely to look for individuating information about that person as opposed to relying on stereotypes. While it is a small consolation that individuating information is occasionally sought out, more frequently than not people will not be overly dependent on others and thus typically yield to stereotype use.

However, despite substantial evidence hinting at the likelihood and almost inevitability of stereotyping, research findings also suggest that with adequate disconfirming information and motivation, perceivers can view others as individuals as opposed to generic category members (Operario & Fiske, 2004). Following initial categorisation, perceivers can engage in more thoughtful, controlled processing to form individual impressions (Fiske & Neuberg, 1990). The processes involved in stereotype change are outlined below.

1.6.4 Instigating stereotype change

The question of how perceivers handle stereotype incongruent information has been of interest to researchers of late, with studies investigating whether this unexpected information is assimilated into the perceiver's mental representation, simply ignored or if it leads to a revision of the stereotype currently in place (Brown, 2010). For example, if you have the

stereotype that women are poor at driving, does this change once you encounter many women who disconfirm this stereotype? Or even after encountering a few striking examples of excellent women drivers?

Stereotype change is thought to be dependent on the perceiver's attention to and incorporation of new information into their established categories. Indeed the exact processes through which category change is possible have been postulated in various models. Rothbart (1981) contrasted two of these processes – the bookkeeping process and the conversion process. The bookkeeping process states that stereotypes change slowly over time, through attention to category-inconsistent targets and integrating new information into the category. Therefore, category expectations change gradually and incrementally by averaging new information with pre-existing beliefs. The conversion process asserts that stereotype change is more rapid, resulting from encounters with highly discrepant category members (Operario & Fiske, 2004). Consequently, just one distinct category member can modify the nature of the category (see Weber & Crocker, 1983 for empirical tests of this distinction). Indeed, it seems likely that both of these processes may contribute to stereotype change, depending on how entrenched the stereotype is and the nature of and frequency with which counter-stereotypical information in relation to this stereotype is encountered.

A third model of stereotype change is the subtyping model (Brewer, Dull, & Lui, 1981; Taylor, 1981). Subtyping refers to the process of mentally grouping or isolating group members who disconfirm a group stereotype (but who otherwise could be stereotypical of the group e.g. career women may be thought to share women's stereotypical concerns about physical appearance but otherwise be atypical of stereotypic females). This type of cognitive re-fencing can thus protect the stereotype from change (e.g. Johnston & Hewstone, 1992). However, with sufficient variability, this subtyping process could also be used to *encourage* change in the overall category stereotype (e.g. Pettigrew, 1981; Rothbart & John, 1985).

Alternatively, perceivers can develop more elaborated beliefs about the category through a process known as subgrouping; unlike subtyping, this process also emphasises the possibility that individuals may be grouped together because they share characteristics that *confirm* the group stereotype as opposed to solely disconfirm it (Maurer et al., 1995). For instance, business woman and homemaker are both subgroups of stereotypic females, yet both groups may display female stereotypes in different ways. Overall these models suggest that stereotype reduction through category variability is likely to be a promising route for stereotype change (Operario & Fiske, 2004).

That said, impression formation models argue that focusing on an individual's unique attributes as opposed to their category-consistent traits is the most individuated type of person perception (Brewer, 1988; Fiske & Neuberg, 1990). Unfortunately, individuation is rare - even subgrouping and subtyping processes focus more on categories rather than individuals – as it involves a great deal of effort on the perceiver's behalf. However individuation is possible when perceivers are motivated because of reasons such as outcome dependency (Erber & Fiske, 1984; Neuberg & Fiske, 1987), accountability (Tetlock, 1992), accuracy goals (Monteith, 1993) or because of other social motives (see Fiske, 1998 for a review). Additionally, if a single category fails to account for a target individual, perceivers simply treat the category as another attribute of the target. The unique traits of the target are appraised and incorporated into a more individuated impression (cf. Anderson, 1981) and re-categorisation may potentially occur in this way.

1.6.5 Stereotype reduction

Stereotype reduction has traditionally been the focus of social psychologists, with their approaches typically reliant on changing attitudes and behaviour within people's volitional control using a combination of awareness and effort (Dasgupta & Greenwald, 2001). For instance through the replacement of automatic, culturally stereotypic responses with more considered responses that reflect personal beliefs (Devine, 1989; Monteith, Devine, & Zuwerink, 1993; Monteith, 1993) or by inciting suppression of negative stereotypes (Macrae, Bodenhausen, Milne, & Jetten, 1994).

But what is meant by 'reducing' stereotypes? Lenton et al. (2009), who support the conceptualisation of stereotypes as 'states' over 'things', in line with connectionist models, argue that it is perhaps misleading to say a stereotype has been 'reduced' as this suggests they are 'stable internal structures' as opposed to more elastic concepts that allow individuals to hold multiple different representations of social category exemplars in their minds. Therefore, in line with these authors, if stereotyping is said to have been reduced in this thesis, it can be inferred that an output pattern inconsistent with gender stereotypes was produced i.e. a pattern that would not be anticipated following stereotype consistent or irrelevant input (Lenton et al., 2009).

1.6.6 Stereotype activation vs. stereotype application

One final point to note before a review of past interventions aimed at stereotype reduction, is a distinction between stereotype activation and stereotype application. These two processes

are closely tied to automatic and controlled processing; stereotype activation takes place automatically and stems from increased cognitive accessibility of traits/features connected with a specific group, while stereotype application is typically under the conscious control of a perceiver and involves the actual use of stereotypes in response to a group member (Kawakami, Dovidio, & van Kamp, 2005).

As regards stereotype reduction, Devine (1989) reasoned that since stereotypes necessarily have a longer history of activation than newly attained personal attitudes, then non-stereotypic responses would require purposeful suppression of automatically activated stereotypes and triggering of the newer belief structures. Essentially, an individual must exercise inhibition over automatic stereotyping and engage in intentional activation of non-stereotypic beliefs (Devine, 1989). Therefore, although stereotype activation is automatic, the triggering of personal beliefs and stereotype application in person perception or judgement requires conscious attention and needs to be under the motivational control of the perceiver.

Indeed both stereotype activation and application processes are of relevance to the current thesis which employs the judgement paradigm devised by Oakhill et al., (2005) throughout. As described in Section 1.5.4 this paradigm involves initial stereotype activation (upon presentation of a gender-biased role noun) which participants must try to overcome in order to correctly judge whether or not two terms may refer to one person. This judgement is a measure of stereotype application (i.e. it is a reduction in stereotype application that the strategies in this thesis are concerned with).

While the initial presentation of the gender-biased role name in the judgement task is likely to elicit implicit stereotyping through the activation of semantic cues from spreading activation, it is not certain that this is the case – it may simply be explicit stereotyping that is taking place as participants have no time restrictions within which to respond. Indeed, Carlston (2010) suggested that mental representations may not always fall neatly into the category of implicit or explicit. He argued that awareness is presumably a graduated state as opposed to 'all-or-nothing' with an intermediate condition of vague awareness. Similarly, intentionality and control may be matters of degree rather than absolutes, with the possibility that some sub-components of representations may be represented explicitly while other may not. However, regardless of whether gender stereotypes are activated implicitly or explicitly when stereotype-biased role nouns are read, their activation is known to be immediate, persistent and difficult to overcome.

The aim of this thesis is to develop interventions that will lead to reduced stereotypic responding following presentation of a gender-biased prime. Because of their relevance to the methodological paradigm used throughout this thesis, and their influence on gender stereotypic responding, interventions aimed at overcoming both stereotype activation and stereotype application will be reviewed below. Indeed, while Devine (1989) argued that only stereotype application is controllable given adequate motivation to behave in an egalitarian and non-prejudiced manner, Blair and Banaji (1996) claimed that motivational control over stereotype activation is also possible. Their research, and others', is reviewed below in an effort to ascertain the factors that may moderate levels of gender stereotyping.

1.6.7 Previous strategies aimed at stereotype reduction

Blair & Banaji (1996) were interested in the automatic activation of gender stereotypes, specifically the role of intentional strategies and cognitive resources in moderating the effect of stereotype priming. They conducted four experiments, in which participants were presented with word pairs comprised of either a personality trait or non-trait as the prime (e.g. caring/doll) and a male or female first name as the target. Their task was to quickly decide whether this target was a male or female name. Experiments 1 and 2 were devised to demonstrate baseline stereotype priming under high cognitive load conditions (Stimulus Onset Asynchrony (SOA) of 350ms or 250ms) and without any intentional stereotype strategy. Priming was indeed evident as participants demonstrated slower response times to stereotype inconsistent prime-target pairings compared with stereotype consistent pairings.

However in Experiments 3 and 4, they investigated whether stereotype priming can be overcome when perceivers have (1) sufficient cognitive resources (2,000ms SOA) vs. low cognitive resources (350ms or 250ms SOA) and (2) an intentional strategy to process counter-stereotypic information (relative to receiving an intentional strategy to process stereotypic information). The counter-stereotypic strategy advised participants to expect a male target name following a stereotypically female prime (e.g. perfume, Brian) and a female target name following a stereotypically male prime (e.g. ambitious, Betty). The opposite instructions were given with the stereotype intention strategy.

Blair and Banaji found that participants anticipating stereotype-consistent combinations succumbed to stereotype activation while those anticipating stereotype-inconsistent combinations showed evidence of overcoming this activation, for instance displaying significantly weaker priming at 350-ms SOA than those with a stereotype-consistent intention

strategy (findings that are comparable to classic findings in the word recognition literature on expectancy and processing of targets e.g. Neely, 1977). They concluded that stereotype activation may be avoidable under the right conditions as a perceiver's intentions and cognitive resources can combine to generate "a gradient of responses from stereotype priming to controlled counter-stereotypic responding" (Blair & Banaji, 1996, p. 1159).

However, Bargh (1999) on reanalysing the above results reports that in fact when participants were given the stereotype-consistent expectation in Experiment 3, the automatic stereotype priming effect was 12 times greater than when the same participants had taken part in Experiment 1 (when no expectancy strategy was operating), while the results of those given the stereotype-inconsistent expectation revealed no inhibitory effect on automatic stereotype priming. This reanalysis suggests that the stereotype-inconsistent strategy was not reducing stereotyping but that the stereotype-consistent strategy was increasing stereotyping; a finding which Bargh concluded was hardly good news for the reduction of automatic stereotyping through the manipulation of participant expectations.

Overall, despite the flawed interpretation of their findings, Blair and Banaji rightfully cautioned that stereotype research should not focus exclusively on one single variable (such as awareness, intention, or cognitive resources) or one design manipulation (such as SOA) to establish the core automatic versus controlled process. They agree with other researchers (Jacoby, 1991; Jacoby, Toth, Lindsay, & Debner, 1992; Roediger, Weldon, & Challis, 1989) in stating that no single measure can be considered as purely automatic or purely controlled and that researchers should be conscious of the complexity with which these two processes may interact to generate a response (Blair & Banaji, 1996).

Another approach to reducing automatic stereotype biases was proposed by Kawakami, Dovidio, Moll, Hermsen, and Russin (2000). They conducted three studies to investigate whether sufficient practice in the negation of stereotypic associations could enable participants to break the "stereotyping habit" (Devine, 2005; Stangor, Thompson, & Ford, 1998). This stereotype-negation strategy involved asking participants to state the word 'no' when they were presented with stereotype-consistent word pairs, and 'yes' when they received stereotype-inconsistent word pairs (while the opposite instructions were given in a stereotype maintenance condition).

Kawakami et al. reasoned that as repeatedly combining certain categories and characteristics together can lead to the formation of stereotypic representations and automatic associations

(Powell & Fazio, 1984), then repeated negation of these associations and combining other 'new' traits with specific categories should lessen automatic stereotype activation. For example, repeated exposure to newly learned counter-stereotypic stimuli responses should result in stronger and more dominant counter-stereotype associations while older automatic stimuli responses should become correspondingly weaker.

In Experiments 1 and 2, participants completed a modified version of the classic Stroop task in which participants are presented with colour-naming trials. These trials were now stereotypic traits pertaining to a particular social group and were preceded by a related social category prime (e.g. skinhead (prime)/hostile (target)), with participants asked to name the ink colour of the stereotype trait. In Experiment 3 participants completed a person categorisation task in which they were first presented with racial stereotypes/ non-stereotypes as primes, followed by photographs of black or white students. The participants' task was to simply state whether the person in the picture was black or white. In studies 1-3 above, it was anticipated that stereotypic primes would facilitate stereotypic responding (e.g. that white stereotype primes would facilitate the categorisation of white faces and that black stereotype primes would facilitate the categorisation of black faces). This was indeed found to be the case, however, such priming effects disappeared following the stereotype negation training (as opposed to a stereotype maintenance training).

Kawakami et al. report that stereotype negation is initially demanding and time consuming but that participants become increasingly efficient across trials. These findings are in line with the procedural efficiency literature (Shiffrin & Schneider, 1977; Smith, 1989), with support for the theory that, with appropriate directions and a large amount of repetition, people can become proficient at a task; in this case at negating stereotypes. Kawakami et al. also found reduced evidence of stereotype bias to newly introduced traits (which had not received negation-training, Experiment 1) and effects that were still clearly visible 24 hours after the initial training phase (Experiment 2).

These results present us with possible process-based explanations as to how perceivers may or may not create dissociations between their explicit egalitarian standards and their automatically activated stereotypes (Devine, 1989; Dovidio & Gaertner, 1998; Gaertner & Dovidio, 1986; Moskowitz et al., 1999). Indeed Kawakami et al. suggest that there are at least 3 different processes through which the negation training may have worked (1) through reinforcement and weakening of category-trait associations as the learning of new counter-stereotypic associations may have led to stereotype dilution, thus reducing stereotype

activation (2) through motivational factors with internalisation of a goal not to stereotype. i.e. with frequent and repeated activation of the goal 'to not stereotype', participants may have learned to impulsively apply a self-regulatory process – a theory closely linked to the auto-motive model of Bargh and colleagues (Bargh, 1990; Bargh & Gollwitzer, 1994; Chartrand & Bargh, 1996). This theory posits that goals and motives must be represented in the mind in the same way as other knowledge structures and should consequently be capable of becoming automatically associated with representations that they are repeatedly paired with. Therefore, if the same goal is repeatedly pursued in a given situation, it should eventually come to be pre-consciously activated in that context, independently of an individual's conscious purposes. Finally, the negation training may have worked through (3) a combination of these two effects.

While the negation training of Kawakami et al. (2000) was comprehensively vetted by the authors (who tested the durability of effects and its extension to new category-trait combinations), Gawronski and colleagues took the training a step further (Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008). In two experiments participants were either given training in negating stereotype congruent information or affirming stereotype incongruent information (as opposed to doing both together in the Kawakami et al. training). Participants then went on to complete either a measure of automatic stereotyping (Experiment 1) or automatic evaluation (Experiment 2). Gawronski et al. observed that only training in the affirmation of counter-stereotypes resulted in reduced stereotype activation and negative evaluations. Moreover, training in the negation of stereotypes actually increased stereotype activation and negative evaluations.

Gawronski and colleagues hypothesised that the negation training could only be effective if it altered the underlying associative representation of the stereotyped category through instance learning. Given that participants did not have to process the outcome of the negation training (i.e. reverse the stereotype into a counter-stereotype) it is unlikely that the meaning of the negated stereotype became independently stored and represented in memory. The authors consequently assert that the affirmation of counter-stereotypes seems to be a more effective method of reducing automatic stereotype activation as this method inherently implies that counter-stereotypical associations will be activated in memory (Gawronski, et al., 2008).

Overall, this research by Kawakami et al. (2000) and by Gawronski et al. (2008) testifies to the use of extensive practice in overcoming stereotypes while the latter research in particular

highlights the key role that the use of counter-stereotypes may play in stereotype reduction[9].

Another successful stereotype-reduction study that made use of counter-stereotypes was conducted by Blair, Ma and Lenton (2001). These authors focused on the strengthening of counter-stereotype associations through mental imagery as a means of moderating implicit gender stereotypes. In a similar vein to Kawakami et al. (2000), it was reasoned that as stereotypes and counter-stereotypes are frequently polar opposites it is unlikely that they will be represented independently of one another. Indeed it is reasonable to assume that as accessibility of one of these concepts is increased, the other will decrease because of cognitive consistency and efficiency pressures (Bodenhausen & Macrae, 1998). Therefore, Blair et al. directed participants to engage in mental imagery of counter-stereotypic information in an effort to increase its accessibility and influence relative to stereotypic information.

They devised an experiment with 4 different imagery conditions; stereotypic (participants imagined a weak woman), counter-stereotypic (participants imagined a strong woman), gender neutral (participants imagined a holiday in the Caribbean) and no imagery (participants played with a simple water game for 5 minutes) (Experiment 2; Blair et al., 2001). Implicit stereotypes were measured both before and after the five minute mental imagery task using the Implicit Association Test (IAT) (Greenwald, McGhee, & Schwartz, 1998). Participants in the counter-stereotype condition subsequently produced significantly weaker implicit gender stereotypes than those in the three other mental imagery conditions, thus providing convincing evidence for the moderating effect of counter-stereotype mental imagery on implicit stereotypes. Indeed the same pattern of results was found when this mental imagery strategy was used with two further measures of implicit stereotype bias; the Go/No-go association test (GNAT; Experiment 4) and a false memory paradigm (Experiment 5).

The GNAT assessed participants' differential ability to detect target stimuli that were stereotypically associated (e.g. female-weak) and counter-stereotypically associated (e.g. female-strong). Participants who had imagined a strong woman showed lower levels of implicit

[9] However (as briefly mentioned in Section 1.6.3) participants who played the role of 'hirer' in Kawakami et al.'s 2007 study did not necessarily show decreased stereotype application in choosing a job candidate after stereotype negation training. The authors hypothesised that if people have the time and occasion to control their responses, they may be influenced by personal principles and temporary goals. They conclude that it generally remains unclear how strategies aimed at reducing activation of stereotypes will affect other expressions of bias, specifically how stereotypes will be employed when targets of a social group are encountered (Kawakami, Dovidio & van Kamp, 2007).

bias (i.e. they were just as accurate in detecting counter-stereotypically associated stimuli as stereotypically associated stimuli) than participants in the other imagery conditions.

With the false memory paradigm, participants were presented with a list of 90 words that included 15 female-typical terms (e.g. lady), 15 gender neutral roles (e.g. author) and 15 gender neutral traits (e.g. funny)[10]. They then completed simple math tasks before a surprise recognition test. This test contained words that were both on the original list (12) and not (34), with the latter set containing 10 stereotypical female roles and traits and 10 counter-stereotypical roles and traits. Participants were asked to identify the terms that they had previously been presented with. Results showed that those in the counter-stereotype mental imagery condition made significantly fewer 'false alarm' responses to the feminine attributes than participants in the neutral imagery condition.

This comprehensive set of experiments by Blair and colleagues suggest that implicit associations can be altered by directing participants' attention to subtypes of group members or triggering counter-stereotypical links in the cognitive network (Bodenhausen & Macrae, 1998; Kunda & Thagard, 1996). Indeed, Blair et al. conclude that a practical implication of focusing on counter-stereotypes is that these group members will then become more salient. These data also suggests that constructs made temporarily accessible (through exposure to or activation of traits or behaviours) can influence an individual's subsequent attitude- and belief-based responses without their awareness (Banaji, Hardin, & Rothman, 1993).

Overall, current theories of prejudice and stereotype reduction now posit that prejudice responses may be moderated if an individual is aware of their biases, motivated to overcome them, has sufficient cognitive resources, and certain goals and situational cues are present (Allport, 1954; Blair & Banaji, 1996; Devine, Monteith, Zuwerink, & Elliot, 1991; Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000; Petty & Jarvis, 1998; Smith & Zarate, 1992; Wilson, Lindsey, & Schooler, 2000). Consequently, the belief that stereotype activation can be changed, given the right strategies and under the right conditions is becoming more widespread (Blair et al., 2001). That said, the stereotype-reduction literature is dominated by research investigating the efficacy of contextual triggers for overcoming stereotype activation and use. While focusing on this area has proved extremely beneficial for pinpointing conditions under which stereotypes are amenable to change, there remains a variable that has been largely overlooked in the stereotype reduction literature - that of the influence of an

[10] The remaining words were included to hide the purpose of the task.

individual's prejudicial attitudes and stereotypic beliefs on task performance. Accordingly, the question explored below is whether individual differences in such areas can moderate or even entirely inhibit stereotyping.

1.7 Individual differences in stereotyping

Past research has been unclear on the influence of individual differences in stereotyping. For example, Dunning and Sherman (1997, Study 5) observed that implicit gender activation (in the form of tacit inferences) is not dependent on the level of sexism exhibited by a participant, while Devine's (1989) influential article (see Section 1.4.2) suggested that high and low prejudice individuals exhibit similar levels of automatic stereotype activation because they have been equally exposed to and are comparably knowledgeable about social stereotypes. The cognitive difference between these two groups was thought to arise because the former group consciously replaces the automatically activated category information with their more egalitarian beliefs, while the latter group does not. However, evidence is now mounting to suggest that the general attitude of a person towards particular social group members may indeed play a prominent role in the moderation of stereotype knowledge activation. Indeed, given that various bottom-up (e.g. background knowledge) and top-down processes (e.g. goals and motivations) interact to create a mental representation of stereotype content, it seems likely that individual differences in attitudes or beliefs *will* dictate the nature and strength of gender inferences (Brown & Turner, 2002; Gygax & Gabriel, 2011).

In a number of studies that primed categorical[11] stimuli, participants with more egalitarian beliefs were found to show no evidence of stereotype activation whereas the prejudiced group did (Lepore & Brown, 1997; Locke, MacLeod, & Walker, 1994; Wittenbrink, Judd, & Park, 1997). For instance, Lepore and Brown (1997) report interesting evidence from a subliminal priming task in which low-prejudiced participants did not show stereotype activation following category priming of Black Americans (e.g. through the use of terms such as 'Blacks', 'Rastafarian'; Study 2) but did when the stereotypical negative content of that category was primed (through the use of terms such as 'drugs' and 'crime'; Study 3). In contrast, high prejudice participants were found to activate stereotypes in both of these studies. It appears that high and low prejudice participants may hold different cognitive representations of the social group of Black Americans in memory. While these representations do not differ in

[11] Categorical stimuli include words that are alternative labels for a particular social category or category evocative words (but does not include valenced stereotype content). For instance, other labels for the social category of black people include 'Blacks' or 'Afro-Carribean' (Lepore & Brown, 1997).

content (Study 1), they do appear to differ in the strength with which the positive and negative traits are associated with the category label. Lepore and Brown conclude that it is an individual's stereotype endorsement (as opposed to stereotype knowledge, as suggested by Devine, 1989) that is likely to influence group representation in memory, through reinforcement between a particular category label and certain stereotypic attributes over other attributes.

As regards stereotype application, Bargh (1999) remarks that individual differences in stereotype use such as those found by Lepore and Brown (1997) may have arisen because automatic stereotypic associations may not have developed in the first place. Monteith (1993) addressed this issue more directly by investigating how stereotypes may be overcome once they are known to be in place.

Monteith's research focused on prejudice-related discrepancies, and revealed that many people show evidence of a conflict between how they *think* they should react to members of various groups and how they *would* react to them – frequently acting in ways that are more negative than their personal standards deem appropriate. She hypothesised that once such discrepancies are noted by people who hold egalitarian attitudes, feelings of guilt or self-criticism may increase. Such feelings may in turn lead to the activation of self-regulatory processes in the hope of reducing future discrepancies between an individual's explicit and implicit stereotypic responses (Devine & Monteith, 1993; Monteith, Zuwerink, & Devine, 1994; Monteith, 1993). In Monteith's study (1993), participants were asked to rate the suitability of a homosexual applicant for a specific job. Once done, they were then informed that their ratings were lower than those of other participants who evaluated an applicant who was described identically except stated to be heterosexual. Having had their prejudice highlighted, participants who had previously expressed non-prejudiced values displayed an increased effort to be non-prejudiced in a subsequent part of the experiment in which they rated jokes about gay men (among others) on dimensions such as how funny they were. However, those who had not expressed such non-prejudiced values remained unaffected by their prejudiced ratings. Monteith concluded that once individuals are made aware of their prejudice, motivation to avoid such behaviour may ensue.

While Monteith examined the role of individual differences on stereotype application, Kawakami, Dion, and Dovidio (1998) investigated the role of individual differences in stereotype activation. Using a priming-related pronunciation task they report that, following the presentation of category primes (black or white), high prejudice participants showed a

facilitation of articulating target black stereotypic traits in both automatic (300ms SOA) and controlled (2000ms SOA) conditions. No such facilitation was found for low stereotype participants in either of the two SOA conditions. In a second phase of the study participants were asked to complete 3 prejudice scales and one measure investigating their personal endorsement of stereotypic traits. As in the work of Devine (1989) and Lepore and Brown (1997), both high and low prejudice participants were found to be equally knowledgeable about the stereotypes used in the study, and differed only in their endorsement of these stereotypes. While both groups showed significant stereotypic associations for blacks, the high prejudiced group revealed a much stronger effect.

With evidence mounting for the moderating effect of individual differences on automatic stereotype activation (e.g. Kawakami et al., 1998; Lepore & Brown, 1997), Moskowitz et al. (1999) further questioned whether an individual's commitment to a goal could result in control over *pre-conscious* activation of stereotypes. They hypothesised that those who were committed to gender-related egalitarian goals (termed 'chronic egalitarians') would avoid stereotype activation through preconscious stereotype control relative to those who were not committed to such goals ('nonchronics'). Moskowitz and colleagues reasoned that a situational cue may first be required to initiate egalitarian goals along with conscious effort and intent on the perceiver's behalf, but that over time this goal may be internalised and lead to stereotype control when specific environmental features are encountered.

In a word pronunciation task, participants were presented with photographs of men or women followed by an attribute that was either consistent or inconsistent with the stereotype of women which they had to articulate as quickly as possible. The results revealed that when using short SOAs of 200ms (so as to avoid conscious processing), chronic egalitarians did not benefit from the classic response time advantage found with congruent prime-target combinations while the nonchronics did (Experiment 3). This difference in performance between chronics and nonchronics was not observed when an SOA of 1500ms was used, thus suggesting that stereotypes had not been activated for chronics at 200ms SOA but that they attempted to suppress the stereotypes at 1500ms SOA. The authors conclude that stereotype activation is dependent on the preconsciously operating goals of participants, as commitment to egalitarian goals can inhibit stereotype processes.

In a similar vein Devine, Plant, Amodio, Harmon-Jones, and Vance (2002) conducted 3 studies investigating the internal motivations (e.g. personal beliefs) and external motivations (e.g. concerns about approval from others) that drive participants to respond without (racial)

prejudice on a range of tasks. They report that explicit race bias (as measured on the Attitude Towards Blacks scale, Brigham, 1993) was indeed moderated by internal standards to answer without prejudice, while implicit bias (as measured on a sequential priming task, Study 1, and an Implicit Association Test, Studies 2 and 3) was moderated by both internal and externally motivating factors. Specifically, participants with high internal but low external motivation to avoid prejudiced responding showed lower levels of implicit race bias than the other participants[12].

1.7.1 Individual differences and gender processing in language

A topic of particular pertinence to this thesis is how individual differences in sexism may moderate performance on processing of gender-biased role nouns. However, little research has focused on this issue. That said, Matheson and Kristiansen (1987) report that people holding more negative attitudes towards non-traditional women (as measured on the Attitude Towards Women scale of Spence & Helmreich, 1972) correlated with greater levels of gender-biased pronoun use on a sentence completion task in which a gender-biased role noun was first introduced e.g. " As a librarian sorts books…".

In a similar vein, Gabriel, Garnham, Sarrasin, Gygax, and Oakhill (2010, unpublished[13]) investigated whether sexist beliefs could influence the way readers process gender-biased role names, in English, French and German. Their objective was to establish the influence of the different grammatical systems in these languages on the processing of gender-biased role nouns (which although written in the masculine form in French and German, should be interpreted generically according to grammatical rules), and the influence that the use of the plural pronoun 'they' would have on the role noun interpretation. While 'they' is gender neutral in English, the equivalent French terms 'ils' has a masculine gender marking while the German term 'sie', though interpretable as referring to either men or women, has the same form as the feminine singular. An example of the sentences used is "*(a) The neighbours came out of the cafeteria. They went away. (b) Because of the cloudy weather one of the women [men] had an umbrella*". They hypothesised that if sexism leads to increased stereotyping,

[12] While it may seem surprising that those scoring high in both implicit and explicit motivations to avoid stereotyping did not show the lowest levels of implicit bias, the authors posit that this finding is in line with findings of previous studies (Plant & Devine, 1998; Devine et al., 2002), where it is thought that the combination of internal and external motivations is associated with less effective regulatory processes - particularly on implicit responses which are theoretically more difficult to control.

[13] Note that this paper was subsequently published, but with the sexism component of the paper removed. The order of authors also changed with the article now retrievable under Garnham, Gabriel, Sarrasin, Gygax and Oakhill (2012).

then those scoring higher in sexism (on the Modern Sexism Scale, Swim et al., 1995) should be more influenced by stereotyped gender information in a text. Specifically, in the grammatical gender languages of French and German, those scoring higher in sexism should be less influenced by the grammatical form of the role noun form than those scoring lower in sexism.

Gabriel et al. observed that in English, the mental representations of readers were indeed moderated by sexist beliefs, suggesting top-down influences are at work when reading role nouns. No such moderation was found in French and German with the authors suggesting this may be because (1) the male bias of the role noun (prompted by use of the grammatical masculine form) may have been too strong for moderating factors to have any influence or (2) more likely, participants' sexist beliefs may have affected how this masculine form was processed (as opposed to simply influencing the stereotype bias associated with the role noun). For example, those with lower sexist beliefs may be more sensitive to the use of gender biasing forms and avoid specific interpretations of the plural masculine terms. However, these two possibilities have yet to be disentangled and further investigated (Gygax & Gabriel, 2011).

Overall, this evidence that individual differences in sexism can affect processing of stereotypical gender information in English is of theoretical relevance to this thesis. With a dearth of literature on this issue, the current research aims to more comprehensively examine several individual difference measures and investigate how they may moderate performance on a judgement task involving activation of stereotypical gender information. More generally, with evidence mounting to suggest that personal beliefs, attitudes and goals do indeed play a role in stereotype processing (e.g. Devine et al., 2002; Kawakami et al., 1998; Lepore & Brown, 1997; Monteith, 1993), the importance of examining individual differences so as to develop a truly integrated theoretical account of human behaviour is patently clear. Collectively, the findings outlined in this section highlight the value of including both implicit and explicit measures of individual differences alongside behavioural tasks in research. In so doing, researchers may indentify person-related factors that moderate stereotyping, and ultimately ascertain the conditions under which stereotyping can be overcome.

1.8 Ingredients of a successful intervention

1.8.1 Past meta-analyses

Lenton et al. (2009) conducted a meta-analysis of studies concerning the reduction of automatic gender stereotypes, in an effort to investigate the relative success of strategies used

(through effect size comparisons) or conversely the stability of automatic stereotypes. They posited that in order to assess how effective an intervention strategy is, it is important to know how much automatic stereotypes have been reduced in the past. However, given the over-reliance on single session studies in the literature, they warn that their meta-analysis essentially investigates the effectiveness of interventions at changing current output patterns but not necessarily the underlying stereotypes. Specifically, the potential moderators that Lenton et al. examined in their analysis included (1) the intervention method used, (2) the specificity of the intervention i.e. whether it aimed to overcome automatic gender stereotypes in general or focused on stereotypes related solely to women (3) the type of indirect measure used, (4) nationality of the sample, (5) gender composition of the sample, (6) publication status of studies and (7) sex of the first author.

They discovered that interventions targeting reduction of automatic gender stereotypes have generally been successful, although the average effect size was small. However, significant heterogeneity was evident in the sample of effect sizes indicating the presence of moderators. Indeed three moderators were found (1) published studies were associated with larger effect sizes than unpublished studies (the latter did not differ significantly from zero) (2) US studies resulted in larger effect sizes than those with EU respondents (the latter did not differ significantly from zero); they suggest this might be because frequently used implicit measures (in particular those relying on semantic priming) were developed according to the attitude and belief structures of North Americans and may not be valid beyond this region, and (3) the intervention method used was important, as those based on distraction and heterogeneity (i.e. stereotype inconsistent information) were both found to be more successful at overcoming automatic gender stereotypes than those based on suppression. No evidence was found to suggest that the remaining variables had any notable effects on gender-stereotype reduction strategies[14].

As I could not influence 2 of the 3 potential moderators before conducting my research (publication status of the strategy or nationality of the sample), it is the third moderator which is of most relevance to this research; intervention method. Lenton et al. grouped interventions into 3 different categories (A) Interventions aimed at distracting or redirecting participant's attention *before* category activation (B) interventions that facilitate the holding of multiple, different representations within the activated stereotype i.e. stereotype consistent and

[14] However, as regards the type of indirect measure used, the Go/No-Go Association test was found to be unreceptive to or unable to detect change in automatic stereotypes. That said, this finding should be treated with caution as it was based on a very small sample.

inconsistent information and (C) interventions that focus on stereotype prevention or inhibiting expression of stereotypes. As mentioned above, the meta-analysis revealed that interventions based on category A or B interventions were more successful than those based on suppression of automatic stereotypes. While this is interesting, the focus of this thesis was on overcoming stereotype application once stereotype activation had occurred through the presentation of a stereotype-biased role noun. Therefore, methods of reducing stereotype activation through distraction or re-directing of participants' attention before category engagement were not further investigated (Category A). Instead, Category B and C interventions were focused on i.e. once stereotypes had been activated, strategies aimed at creating awareness of category heterogeneity (Category B), and aimed at reducing levels of stereotype application (Category C) were introduced.

Overall, the conclusions of this meta-analysis are somewhat mixed concerning the success of interventions aimed at automatic stereotype reduction. Although the authors found that automatic attitudes are malleable and susceptible to certain single-session interventions (Blair, 2002), the small effect sizes indicate interventions do not guarantee success. Indeed, the studies typically fail to reduce automatic stereotyping to zero, or to generate reliable, counter-stereotypic responding (Gregg, Seibt, & Banaji, 2006). More dishearteningly, Lenton et al. claim that only considering the findings of published studies leads to an overestimation of the success rate of interventions, with the reality being more modest than the published studies suggest. Despite these issues, the authors have provided a platform from which to formulate future research by highlighting some issues that need to be addressed within this field. These will be outlined in further detail in Section 1.9 below, following details of another meta-analysis by Paluck and Green (2009).

As opposed to Lenton et al.'s (2009) meta-analysis, which focused on the issue of overcoming automatic activation of gender stereotypes, Paluck and Green (2009) conducted a broader meta-analysis of prejudice reduction in research. Although prejudice and stereotyping are different processes, they decided to include stereotyping and many related constructs in their definition of prejudice (discrimination, intolerance and negative emotions towards one group), rendering the findings of their meta-analysis of interest to this thesis[15].

[15] Definitions of prejudice and stereotyping have tended to simplify over the years, with prejudice now defined as negative attitudes towards a group or group members, while stereotype refers to the traits that quickly come to mind when we think about different groups (Stangor, 2009).

The aim of their comprehensive review was essentially to establish 'what works' in the field of prejudice reduction. They examined laboratory interventions that take an intergroup approach to prejudice reduction with the aim of modifying group boundaries and interactions, as well as interventions that take an individual approach to prejudice reduction, with the aim of influencing a perceiver's feelings, cognitions or behaviour. It is the latter, individual approach that will be the focus of this research given that all studies in this thesis were conducted on participants individually. Paluck and Green (2009) state that past interventions used in this domain have focused on strategies such as "instruction, expert opinion and norm information, manipulating accountability, consciousness-raising, and targeting personal identity, self-worth, or emotion"(p. 347). The most relevant of these for the current thesis will be outlined in further detail below.

Instruction: Stephan and Stephan (1984) assert that a major source of prejudice is ignorance. A number of laboratory experiments have used various instructional techniques as a means of prejudice reduction, their aim generally being to change the perceiver's way of thinking, for example through training in complex thinking or statistical logic. Researchers reasoned that such training could help people avoid faulty group generalizations. Indeed such techniques have achieved modest success. For example, following training, students have written more positive stories about images depicting interracial encounters, shown more friendliness towards ethnic out groups (Gardiner, 1972), and avoided stereotyping characters displayed in a vignette (Schaller, Asp, Roseil, & Heim, 1996).

Consciousness-raising: The use of suppression as a stereotype reduction strategy has been tried in the past but with limited success due to rebound effects (e.g. Macrae et al., 1994). While some findings suggest rebound effects are avoidable (e.g. Monteith, Sherman, & Devine, 1998), particularly when suppression of stereotypes is combined with cognitive re-training exercises (Kawakami et al., 2000), the consensus is that suppression is not a very effective strategy for the reduction of prejudice. Therefore, taking the opposite approach, researchers have also tried to raise awareness of attitudes and beliefs relating to stereotyping and prejudice. For instance, Son Hing and colleagues asked participants to think of a time when they treated an Asian person in a prejudiced way. Students who had previously shown high prejudice on an implicit prejudice test were more likely to experience guilt from this memory and, in a subsequent survey, were more likely to indicate support for the funding of an Asian student association at their university than those who scored low in prejudice (Son Hing, Li, & Zanna, 2002).

Oakhill et al. (2005) also aimed to reduce stereotype application using a consciousness-raising strategy (briefly mentioned in Section 1.5.4). They took the approach of explicitly reminding participants in a gender stereotype judgement task that some jobs are not clearly marked for gender, then advising that participants should carefully consider whether the first term presented (the stereotype biased role noun) could be filled by a man or woman only, or whether it could be done by either sex (Experiment 4). While this approach did not completely eradicate the effects of stereotype priming, there was evidence of improved task performance relative to similar studies of theirs in which other elements of the design were varied (such as presentation duration of the stimuli and presenting word-pair terms together or separately). Thus, the research of Oakhill et al. again hints at the value of using a consciousness raising strategy to reduce stereotyping.

Reconditioning: Another approach used in laboratory interventions is that of reconditioning implicit attitudes and beliefs. Such interventions can involve the use of classical conditioning techniques such as pairing stigmatised groups with positive images or words. For example it was found that presenting positive pictures of well-known black people (e.g. Martin Luther King) and negative pictures of well-known white people (e.g. Charles Manson) led to a reduction in implicit racial prejudice on an IAT while conscious attitudes remained unchanged (Dasgupta & Greenwald, 2001; Wittenbrink, Judd, & Park, 2001).

Expert opinion and norm information: Much evidence now suggests that prejudice attitudes and behaviours are strongly influenced by social-norm information (Crandall & Stangor, 1995) and expert opinion (under certain conditions) (Kuklinski & Hurley, 1996). Social norms are defined as “perceptions that are descriptive of what people are doing or prescriptive of what people should do (as a member of a group, an organization, or a society)” (Paluck & Green, 2009, p. 347).

Indeed, Paluck and Green (2009) suggest that knowledge of the power of authority and conformity, stemming from landmark research by Milgram (1963), Asch (1956) and Zimbardo (1972), has not been fully utilised in laboratory based prejudice-reduction research. That said, exceptions to this are starting to emerge as evidence has now been found that perceivers frequently adjust their intergroup attitudes and behaviours to align with those modelled by their peers. For instance, participants have been found to adapt their beliefs in line with the perceived peer consensus on stereotyping when asked to state their personal racial stereotypes (Sechrist & Stangor, 2001; Stangor, Sechrist, & Jost, 2001) and to show more tolerance of prejudice against minorities and women following racist or sexist jokes (Ford &

Ferguson, 2004; LaFrance & Woodzicka, 1998). Indeed, even when a peer's position on prejudice is subtly signalled though an anti-racism t-shirt, the perceiver's unconscious prejudice can be influenced (Lun, Sinclair, Whitchurch, & Glenn, 2007; Sinclair, Lowery, Hardin, & Colangelo, 2005). Peer influence on prejudice is a promising area of research with positive results following peer-related interventions already revealing the communicative and normative character of prejudice change.

Various theories have been offered to explain this peer influence on intergroup prejudice, which is now thought to result from the basic human objectives of understanding and social connection (Asch, 1956; Cialdini & Goldstein, 2004). For instance, Social Reality Theory posits that striving for understanding and connection pushes people to validate their experiences with others and to display actions and beliefs that others value (Hardin & Conley, 2000). Similarly, Group Norms Theory suggests that people assume the perceived attitudes and behaviours of others who exemplify admirable in-group identities in an effort to socially connect with the group (Crandall, Eshleman, & O'Brien, 2002; Kelman, 1958; Sherif & Sherif, 1953). While both of these theories predict that individuals would adjust their behaviour and attitudes in line with those of the valued reference group, other theories predict a contrasting pattern of results. Specifically, Deviance Regulation Theory (Blanton & Christie, 2003) claims that people may reject the perceived attitudes and behaviours of their peers as a means of self-definition, while the Focus Theory of normative conduct predicts that peer values will only influence individuals when they are made salient, which is not always the case across contexts and time periods (Kallgren, Reno, & Cialdini, 2000). Overall, given that peer influence may be driven by basic human objectives, and has been shown to alter both pro-social and anti-social beliefs and behaviour, it would be useful to further explore the extent of its influence in terms of stereotype and prejudice reduction.

1.9 Observations from the field & suggestions for future research

Despite the many strategies devised over the years to moderate gender stereotyping, Bargh (1999) argued that the field of social cognition has become overly optimistic about tackling the 'cognitive monster' that is automatic stereotype activation. He claims that conclusions being drawn from research are overestimating the degree to which automatic stereotypes can be moderated with the use of good intentions and effortful thought. Indeed following their respective meta-analyses both Paluck and Green (2009) and Lenton et al., (2009) conclude with their observations of what is lacking in current research. The work in the current thesis

has drawn on this advice and sought to devise and implement interventions that address some of the issues highlighted by these authors, outlined further below.

a) Durability of effects:

Most notably, both sets of authors commented on the durability of stereotype reduction effects induced by past interventions. Paluck and Green noted that the vast majority of laboratory studies in the field of prejudice-reduction test quick fixes in which prejudice is first created or measured, then modified and assessed during the course of an hour. While short manipulations have been found to show powerful effects (e.g. Bargh, Chen, & Burrows, 1996) studies very infrequently assess the durability of effects that have been found. Indeed in their review, Lenton et al. examined just one study that went beyond a single-session experiment (Dasgupta & Asgari, 2004, Study 2)). As learning is typically a slow, incremental process, there is likely to be a limit on the extent to which automatic responding can be reduced with a single session intervention, with Lenton and colleagues positing that such stereotype reduction effects will generally be 'moderate at best' (p. 184). This is cause for concern given that the vast majority of empirical studies examine stereotype reduction in single research sessions.

While not ideal, short intervention studies allow the authors to at least assess how current input can overcome the default pattern of activation that has become embedded over time with the slow-learning system (Smith & Conrey, 2007; Smith & DeCoster, 1999). However, with this obvious need for more comprehensive testing of stereotype reduction strategies, Experiment 3 of this thesis investigated the use of a stereotype reduction strategy based on performance feedback over a one week period. With this experiment it was hoped to establish the efficacy of the feedback strategy as a more long term means of reducing gender-biased responding.

b) Subtlety:

Another issue that Paluck and Green (2009) observed in relation to laboratory based interventions is that of their subtlety. For instance, past strategies have aimed to influence the perception of others using techniques based on seating assignments, minor changes in instructions or even t-shirt colour. In contrast, real world institutions are much more forceful and, for example, can influence intergroup perception and behaviour with their strict citizenship requirements and immigration quotas. Laboratory-based interventions are frequently detached and abstracted from their real life application, not accounting for "larger

institutions and social processes in which interventions are embedded" (Paluck & Green, 2009, p. 349). This detachment may affect the influence of an intervention that has been devised in a lab but then used outside of this context. Although subtle manipulations are of important scientific value, an exclusive focus on such methods means that a fuller range of situational interventions will not be uncovered. While this observation and its implications for future research are arguably more suited to a social psychological approach to reducing gender stereotyping than that of the cognitive approach adopted here, the advice of Paluck and Green was adapted for the current research; for the most part, quite explicit training strategies were devised relative to methods that have previously been used in the field of stereotype reduction e.g. subliminal priming, evaluative condition paradigms. Examples of the explicit trainings that were conducted include use of overt performance-related feedback on participant performance (Experiments 1 and 3), and use of striking imagery to highlight occupational roles that can be occupied by both men and women in society (Experiments 8 and 9).

c) Men vs. Women:

Thirdly, Lenton et al. found no difference in the effectiveness of interventions aimed at changing stereotypes specifically about women, versus more generally about both sexes. However, it is not possible to ascertain whether male and female stereotypes would be equally influenced by past trainings given the dearth of studies attempting to change only the male stereotype. Lenton and colleagues (2009) urge researchers to address this research imbalance by investigating the extent to which male stereotypes are susceptible to stereotype reduction strategies relative to female stereotypes, or gender stereotypes more generally. They hypothesise that there would be numerous advantages to such a research focus. For example, addressing this bias may help to shed some light on why the male role is thought to have changed less in the past 50 years than the female role (Diekman & Eagly, 2000) and may indirectly provide support for Lenton et al.'s argument that the male stereotype is less heterogeneous than that of the female stereotype (Lenton et al., 2009). Moreover, given evidence that men are generally liked less than women (Eagly, Mladinic, & Otto, 1991; Rudman & Goodwin, 2004), it appears there is certainly scope for modifying people's attitudes about and expectations of men. In this thesis the judgement task of Oakhill et al. (2005) was adopted as a measure of gender stereotyping. As described in Section 1.5.4, this task involves presentation of both male-biased and female-biased role names in gender incongruent word pairs (among others), responses to which are indicative of gender stereotyping or not. This

research sought to overcome gender biases in reaction to *both* of these sets of role nouns, with the judgement paradigm allowing measurement of reaction times and response accuracy to each set independently of the other. In this way, it can be investigated whether different patterns of responding emerge in respect to male and female role nouns.

One further point regarding men and women, was that of the gender balance of the experiment sample. As the topic of this research was gender stereotyping, it was imperative that participant sex was balanced as far as possible in each experiment. In this way it could be inferred that the stereotype reduction trainings successfully extended to both sexes as opposed to just one (unless of course the data proved otherwise after analysis). However, no specific hypotheses were made as regards this issue.

d) Social norms:

Lenton et al. (2009) assert that more research is needed on how motives (self or social; Blair, 2002; Sedikides & Strube, 1997) influence automatic gender stereotypes, while Paluck and Green (2009) claim that the concepts of authority and conformity are under-used in stereotype-reduction. These two suggestions were combined in this thesis and the influence of social motives on stereotype reduction was examined through the presentation of fictitious social-norm information (which aimed to induce conformity of the participant towards the perceived consensus of their peers). Past research in this area has already been described in Section 1.8.1, yet more specifically, the exact form that the current social consensus feedback should take was influenced by literature suggesting that the source of this norm information is an important variable in exerting influence over a person's attitudes or beliefs. This research suggests that information coming from a valued in-group will be highly effective in modifying an individual's beliefs (e.g. Sechrist & Stangor, 2001; Stangor et al., 2001). For this reason feedback was based on the participants' peer group of 'previous students who had completed the same task' in the hope that they would regard displaying similar behaviour to this group as both important and desirable (see Experiments 4, 5 and 6).

e) Verbal stimuli:

Another limitation of the research to date identified by Macrae and Bodenhausen (2000) is the over reliance on verbal stimuli (category labels) to examine the category activation process. They argue that in reality perceivers can classify people along multiple dimensions i.e. people are a much more complex stimulus than verbal labels suggest. It therefore cannot be assumed

that the processing of verbal labels equates to the processes involved in person perception. While future research would undoubtedly benefit from an investigation of the processes involved in stereotyping when perceivers encounter real people, the current research has taken a step in this direction and incorporated images of real people as a central part of the experimental design of Experiments 8 and 9.

f) Extension of results to other stimuli: transfer effects

A further observation made in relation to past research is that of a distinct lack of studies testing whether interventions still succeed in lowering stereotyping or prejudice beyond the stimuli on which participants initially received training (an exception to this being Kawakami et al. (2000) who tested their stereotype negation training on stimuli relating to two social groups – skinheads and racial categories). Too often training is conducted on a narrow set of target stimuli and simply assumed to work beyond this situation. An effort to address this issue was made in Experiment 3 of this thesis, as the efficacy of a previously successful stereotype reduction strategy was examined in relation to a novel set of role nouns (which participants had not received training on). This practice of evaluating a stereotype reduction strategy beyond the immediate training context is imperative in determining the true worth of a training. While strategies that positively impact on a very specific set of stimuli can certainly have their uses, it is stereotype reduction strategies that obtain results beyond a very specific context that are ultimately likely to prove most valuable.

g) Individual differences in stereotyping

As mentioned in Section 1.7, current research on the moderating effects of individual differences in stereotyping is ambiguous. Therefore, using a variety of self-report measures and the Implicit Association Test, this topic was explored in considerable detail in this thesis. While some measures were employed in just one experiment (and compared to performance on the stereotyping task), others were used across numerous experiments, with data combined to increase statistical power and the chances of accurately identifying potential stereotype moderators. These results are comprehensively discussed in Chapter 5.

Overall, across the following Chapters 2-4, a number of stereotype-reduction strategies that incorporate the above suggestions are investigated. These strategies were inspired by or adapted from analyses of past literature with similar aims, whether from a social or cognitive perspective. With a focus on the gender stereotype biases associated with many role names in

English, this research aims to overcome initial activation of this bias and increase levels of non-stereotyped responses in a judgement task. Ultimately it is hoped to devise a laboratory intervention that will contribute to more durable changes in stereotype reduction and identify whether or not particular individual differences may moderate stereotype use.

2. Performance-related feedback as a strategy to overcome automatic gender stereotypes

2.1 Introduction

In order to engage in successful social interaction, a perceiver must be able to manage unexpected information by resolving the conflict between activated expectations and the current stimuli (Hastie & Kumar, 1979). Evidence suggests that such resolutions can be difficult to achieve in language processing, in particular when gender-related expectancies clash with explicitly stated gender information (e.g. Carreiras et al., 1996; Garnham et al., 2002; Irmen, 2007; Kreiner et al., 2008). Therefore, the focus of this chapter is on overcoming the gender-related expectancies that are immediately evoked upon encountering certain role nouns in English.

Given that stereotype activation can take place automatically and often outside of a perceiver's awareness or control, the current research was aimed at devising a strategy that would both (a) alert participants to their stereotype biases and (b) facilitate their overcoming of those biases. A number of explicit training strategies have previously reported success in overcoming stereotype activation and application. For instance, research addressing explicit racial prejudice and stereotyping often employs a type of diversity training in which students or workers are typically challenged to reflect on ways in which they may unknowingly oppress others (e.g. see Paluck, 2006 for a discussion). Researchers hypothesise that learning about one's own biases in this way may result in increased motivation and effort to become more egalitarian. Furthermore, Rudman, Ashmore, and Gary (2001) report such training can also lead to a reduction in *implicit* prejudice (i.e. the automatic negative associations that may occur without a perceiver's awareness or intention).

In the field of language processing, explicit strategies aimed at reducing stereotype bias have also been identified. For example, Kreiner et al. (2008; Experiment 2) investigated the use of cataphora (as opposed to anaphora) in sentence reading as a means of definitively establishing the sex of a character in a text *before* encountering a role name (at which point a reader would typically rely on stereotype bias information to infer a character's sex) e.g. "After reminding *himself/herself* about the letter, the *minister* immediately went to the meeting at the office". This approach eliminated the processing difficulty they previously found with the integration of the reflexive and the stereotyped role noun in sentences containing anaphors e.g.

"Yesterday the *minister* left London after reminding *himself/herself* about the letter" (Experiment 1).

Another approach was taken by Oakhill et al. (2005) who devised a judgement task in which participants had to quickly decide whether two terms presented on-screen could refer to one person. In the absence of a stereotype-reduction training, participants consistently showed evidence of succumbing to stereotype biases on stereotype incongruent pairings (e.g. *builder/mother*) compared to stereotype congruent pairings (e.g. *builder/father*). However, when Oakhill and colleagues provided participants with explicit instructional reminders that nowadays many jobs are not restricted to one sex, they reported significantly improved performance on stereotype incongruent pairings, although the stereotype bias was not completely overcome (Experiment 4).

Simply put, Strack and Hannover (1996) state that in order for someone to bypass or adjust for the potential influence of unwanted stereotypic thoughts, that person must first recognise that these influences are a possibility. Without such recognition, a perceiver will take no measures to avert or alleviate the biases that may follow (Bodenhausen & Macrae, 1998; Greenwald & Banaji, 1995; Stapel et al., 1998). Conversely, once a perceiver is made aware of the potential for bias in encounters with others, they can employ various regulatory mechanisms to control for these effects (Macrae & Bodenhausen, 2000). For example, social judgements may be adjusted to the opposite direction of the bias (e.g. Wegener & Petty, 1997; Wilson & Brekke, 1994) or attempts may be made to completely prevent stereotypic thoughts through suppression (e.g. Bodenhausen & Macrae, 1998; Monteith et al., 1998). In a similar vein, Bargh (1992, 1999) posits that strategic efforts to overcome automatic stereotypes and prejudice may require some combination of awareness, motivation, skill and resources in order to succeed – albeit to a lesser extent with practice (e.g. Kawakami et al., 2000; Monteith, 1993), while others state that stereotype change is possible when a perceiver has been given sufficient individuating information (e.g. Hilton & Fein, 1989) and when they are motivated to attend to it (see Fiske, 2000 for a review).

With these factors in mind, the stereotype reduction strategy of Chapter 2 was designed primarily to create awareness of stereotype biases. It involved providing participants with performance-related feedback as they completed the judgement task of Oakhill et al. (2005; briefly mentioned above and described further in Section 2.2.2). It was anticipated that this feedback would alert participants to their personal stereotyping tendencies, thus triggering control processes and helping to reduce subsequent levels of stereotype application. The

feedback should essentially remind participants that certain roles can be fulfilled by either sex (irrespective of stereotype norms), assist them in learning from their mistakes and ultimately lead to improved performance on subsequent trials. Together with the accuracy feedback, participants were also provided with their own cumulative percentage scores after each response. It was anticipated that this additional information may help to boost participants' engagement with the task, as they could effectively track their ongoing performance and aim to increase their overall accuracy.

A secondary aim of Chapter 2 was to investigate whether an individual's pre-existing, personal beliefs can moderate or even entirely inhibit stereotyping on the judgement task. Participants were therefore administered a battery of individual difference measures with each experiment, which explored personal differences in levels of sexism, attitudes towards the use of sexist and non-sexist language, sex role perception, and amount of sexist pronoun use. The inclusion of these measures was a conscious attempt to address the frequently overlooked role played by individual differences in the field of stereotype reduction. However, the individual difference measures will not be examined further until Chapter 5 at which point they will be combined with data from later experiments so as to facilitate a more in-depth investigation into the role these elements play in stereotyping.

2.2 Experiment 1: Performance Feedback

2.2.1 Introduction

The purpose of Experiment 1 was threefold; (a) to further investigate whether gender-stereotyped information is automatically elicited from single words (as has been found in previous experiments (Banaji & Hardin, 1996; Oakhill et al., 2005), (b) to examine whether this stereotyping effect can be attenuated with the introduction of performance-related feedback (evidenced by improved performance to stereotype incongruent trials), and (c) to investigate the extent to which individual differences modulate performance on the behavioural task.

As will be further described in Section 2.2.2, the participants' task was to judge, as quickly as possible, whether or not two terms could be used to refer to one person. The word pairs presented to participants fell into five distinct categories: stereotype congruent, stereotype incongruent, neutral, definitionally matching and definitionally mismatching. However, it was performance on the stereotype incongruent pairings that was of most theoretical interest.

These incongruent pairs were comprised of one term with a stereotypical gender bias (e.g. the term *Bricklayer* which is strongly male-biased) and one term of opposing definitional gender (e.g. the term *Sister* which refers exclusively to females). In order to successfully respond, participants were required to resolve the incongruity between prior expectations and the presented stimuli by suppressing stereotypical gender information and relying on definitional gender information. The final judgement of the participant was used as a measure of stereotype application.

With the above information in mind, this experiment had three blocks of word pair trials, with feedback on responses offered only in the second block. It was hypothesised, in keeping with previous findings (Garnham, et al., 2002; Irmen, 2007; Kreiner et al., 2008; Oakhill et al., 2005; Reynolds et al., 2006), that participants would initially respond more slowly and less accurately to trials made up of stereotype incongruent word pairs (e.g. *nurse/father*) than to stereotype congruent word pairs (*nurse/mother*) in Block 1. However, with the introduction of feedback in Block 2 it was anticipated that alerting participants to their stereotype tendencies would lead to better control of these biases. Therefore in Block 2 it was hypothesised that the processing cost associated with the stereotype incongruent condition in Block 1 would be attenuated, and that this improved performance would extend into Block 3, as evidenced by higher accuracy and faster reaction times to the critical trials i.e. reduced stereotype application.

2.2.2 Method[16]

Participants

Fifty-one monolingual, native English speakers (25 male, 26 female) from the student population of the University of Sussex took part in the study. Participants' ages ranged from 18 to 28 years (*M*: 20.21; *SD*: 2.40). They received either £5 or 4 course credits for taking part in the session which lasted approximately 45 minutes.

Materials

Gender-biased role nouns

[16] The chosen measure of stereotype application used throughout this thesis was a judgement task, originally devised by Oakhill et al., (2005). Details of this judgement task are outlined at length in this section. Therefore, for purposes of clarity and brevity, the methodology sections of subsequent experiments will refer back to this one. However, any further details that are experiment or chapter specific will be outlined as appropriate in the relevant sections.

Across all experiments, gender-biased role nouns were chosen from norms compiled by Gabriel, Gygax, Sarrasin, Garnham, and Oakhill (2008). The selected items were those rated as being most highly male-biased (e.g. *bricklayer, president*), most highly female-biased (e.g. *beautician, fortune teller*) or neutral (e.g. *pedestrian, proof reader*), with 12 exemplars chosen in each of these three conditions. A full list of the stereotyped terms used, and their associated bias ratings is provided in Appendix 2. These ratings reveal that the bias scores of the 12 male-biased items extend across a narrower range (11.10%) than the bias ratings of the 12 female-biased items (17.55%)[17]. This suggests that, on the whole, the female-biased items were not judged as being as strongly stereotype-biased as the male items, with ratings of the former approaching closer to neutral. For this reason, participants may show less difficulty in overcoming stereotype biases to female-biased role nouns relative to male-biased nouns throughout this thesis[18]. Finally, ratings of the neutral terms extended across a very narrow range of 5.29%. Combined with the fact that participants should have ample experience of males and females fulfilling neutral roles across their lifetimes, this narrow range of bias ratings is another reason that these terms should prove unproblematic for participants.

Kinship terms

As in Oakhill et al. (2005), 6 kinship terms (3 male, 3 female) were also selected to be used as one of the terms in the word pairs. These terms were *father, mother, brother, sister, uncle, aunt*. Importantly, these words incorporate a specific gender into their definitions e.g. the term 'father' can only refer to a person of male sex.

Critical word pairs

Word pairs were formed by combining the 12 male-biased, 12 female-biased and 12 neutral role nouns with the 6 kinship terms to produce a set of stereotype congruent, stereotype incongruent and neutral pairings. In the congruent condition, male and female stereotyped role names were paired with a kinship term of congruent definitional gender – for example, *pilot/brother* or *nurse/sister*. In the incongruent condition the stereotyped terms were paired with a kinship term of incongruent definitional gender – for example, *nurse/brother* or *pilot/sister*. Finally, in the neutral condition, neutrally rated role nouns were paired with each

[17] An independent samples *t*-test found this to be a significant difference, $t(22) = 3.53$, $p = .002$.

[18] While this difference in bias ratings was not ideal, it was deemed more pertinent to choose the most strongly biased role nouns for each sex (as evidence of overcoming stereotyping to these role nouns should logically extend to role nouns with a weaker bias rating) than to choose role nouns with matching degrees of typicality.

of the male and female kinship terms to create neutral word pairs – for example, *artist/father* and *artist/mother*. Overall, each of the 12 male-biased, female-biased and neutral role terms was teamed once with each of the 6 kinship terms resulting in 72 word pairs in each of the three congruence conditions, totalling 216 critical trials.

Filler trials

Two-hundred and forty filler trials were also created, made by pairing the 6 kinship terms with role nouns that are also gender-specific by definition. In this way, filler trials were gender unambiguous pairings to which participants could respond *yes* or *no* to with relative ease and certainty. The selected role nouns were either explicitly marked for gender (e.g. *waitress*, or *policeman*), official titles (e.g. *count* or *countess*) or other terms that carry gender as part of their meaning (e.g. *lady* or *husband*). These role nouns were sourced from norming studies conducted by Hamilton (2008) and Kennison and Trofe (2003). A full list of the filler terms used is provided in Appendix 3.

Item overview

In total, participants were presented with 456 word pairs, divided into three equal blocks of 152 trials. Each of the stereotyped terms appeared twice in each block, once with a male kinship term and once with a female kinship term i.e. in both a congruent and an incongruent condition. The 6 kinship terms were counterbalanced so as to appear with the critical items an equal number of times in each block. Altogether 276 items, including all critical items, were intended to elicit a *yes* response while 180 required a *no* response.

Performance-related feedback

In this experiment, performance-related feedback was presented to participants as a strategy aimed at reducing levels of gender-stereotype application. This feedback was provided after each response in Block 2 of the judgement task only. It consisted of a statement of accuracy in which participants were simply informed whether their response was 'Correct' or 'Incorrect' along with a report of their cumulative percentage score. Therefore, feedback consisted of a statement such as "Correct! 75% average correct". This feedback remained on screen for 1,000ms before ceding to the next trial.

Design

There were two independent variables, the first of which was the congruency condition of the word pairs (stereotype congruent, stereotype incongruent or neutral), treated as within-

subjects but between-items (as each pair was treated as a separate item). The second independent variable was the stereotype-reduction training (in this case the provision of performance-related feedback), operationalised as performance on different blocks of trials. Two dependent variables were analysed: the proportion of correct answers and response time of judgements to correct trials.

The experiment was programmed using Eprime, version 2.0. In the judgement task, terms were presented one at a time in the centre of a computer screen. A role term was first displayed for 1000ms, followed immediately by a kinship term (inter-stimulus interval of 0), which remained on-screen until a response was made. After responding, feedback immediately appeared on-screen (0 delay) and remained for 1,000ms. This was followed by a 500ms delay before onset of the next trial. The word pairs were divided into three fixed sets to form the blocks of the experiment, while the sequence in which these blocks were presented to participants was counterbalanced. Within each block, trial order was (pseudo-)randomised separately for each participant, using the standard E-Prime procedure. A PST (E-prime manufactured) button box was used for responding, with one button clearly marked 'Y' for *yes* and another 'N' for *no*. Participants made a judgement about every word pair.

Procedure

Participants were tested individually in a quiet laboratory. They first read a short information sheet with details of the study, and if happy to proceed, signed a consent form (see Appendix 4 and 5 respectively)[19]. Onscreen instructions then informed participants to read each pair of words and decide (without excessive deliberation) whether the two terms could apply to the same individual. These instructions provided the participants with two examples of such (definitional) word pairs – one that required a *yes* response and one that required a *no* response. Participants were further informed that they would receive feedback in the second block of judgement trials only, and explained what this feedback entailed. The instructions and examples were then repeated verbally. Next, a short practice session using a representative sample of fillers and critical word pairs was given to familiarise the participants with the experimental task. This consisted of eight trials and involved role terms that were not subsequently used in the experimental blocks. Once the practice session was complete, participants were prompted to ask the experimenter any questions they had before beginning the experiment proper. Once satisfied with the procedure, participants were then left alone to

[19] Examples of these documents are not included for further studies due to reasons of similarity.

complete the judgement task. At the end of the study, participants were fully debriefed as to the aims of the experiment before being thanked and reimbursed for their time[20].

2.2.3 Results

Data screening

The analyses reported below excluded data for word pairs that contained the neutral term *adolescent*. Accuracy of responses to pairs involving this term was low, resulting in only 66% correct responses in Block 1 compared to > 90% accuracy for all other neutral role nouns. On reflection, this may be due to age considerations as opposed to gender stereotyping - the term *adolescent* typically refers to an individual in their teens and was paired with kinship terms that generally imply an older generation e.g. *uncle, aunt, mother, father*. This resulted in word pairs such as '*adolescent/father'*, which proved more difficult for participants to accept as correct than '*adolescent/brother'*, despite both being possible combinations. This resulted in the removal of 1.32% of the data.

Pre-analysis

Next, response times for all errors of judgement were identified and excluded (representing 7.50% of the total data) as were extreme response times, below 150ms and above 4,000ms (representing a further 2.11%), totalling 9.61% of the data. Next, the Subject by Block mean was calculated for each participant. Data points 2.5 standard deviations above or below the Subject by Block mean were replaced with the relevant upper or lower cut off point[21].

Analysis

Across all experiments, both accuracy of judgements and response times (RTs) were analysed using two mixed-design analyses of variance (ANOVAs) on the correct responses: firstly with participants treated as the random variable and secondly with items treated as the random variable. In the by-participants analysis (F_1), the mixed ANOVA had three repeated factors –

[20] Note that, before completing this behavioural task, participants first completed a "questionnaire section" which involved answering a battery of four individual difference measures. However, as mentioned earlier, details of the individual difference data for this, and subsequent experiments, will be outlined collectively in Chapter 5.

[21] The data trimming measures described above and analyses described below were carried out for each experiment (unless otherwise specified). Therefore, only exact figures pertaining to the amount of data replaced in each study will be stated from this point forward, as opposed to repeating the process by which these extreme data points were replaced.

stereotype bias of the role name (Stereotype: Male/Female/Neutral), gender of the kinship term (Kinship term gender: Male/Female) and block of trials (Block: Block1/Block2/Block3). Participant Gender was included as a between-subjects factor. In the by-items analyses (F_2), Stereotype was included as a between-items factor while Kinship term gender, Block and Participant Gender were included as within-item variables. In both sets of analyses, where sphericity was not satisfied, Greenhouse-Geisser (when $\varepsilon < 0.75$) or Huynh-Feldt ($\varepsilon > 0.75$) corrected degrees of freedom and *p* values are presented (as recommended by Girden, 1992). With all paired *t*-tests, within-subject or within-item effect sizes were estimated using Cohen's *dz* while with the independent-samples *t*-tests estimates of between-subject or between-item effect sizes were estimated using Cohen's *d*[22]. Finally, average scores stated throughout this thesis pertain to the by-participants data as opposed to the by-items data (unless otherwise specified), as this protocol appears to be more frequently practiced in the literature.

A note on Congruency

It is important to note that an interaction of Stereotype by Kinship term gender is in essence an effect of Congruency as it is the combination of the levels of these two factors that give rise to the three critical conditions – congruent, incongruent and neutral. As such, all Stereotype by Kinship term gender interactions in this thesis are referred to as Congruency effects (though primarily in relation to the male and female stereotyped terms). This approach was taken (as opposed to including a 'Congruency' variable in the analysis) because Congruency is not independent of the Stereotype and Kinship term gender variables and therefore cannot be included separately. Furthermore, word pairs involving neutral terms differ from pairs involving the male and female stereotype biased terms; when the latter are combined with male and female kinship terms, two different conditions are formed e.g. *pilot/father* is a congruent pairing, while *pilot/mother* is incongruent. However, when the neutral role names are combined with male and female kinship terms, two different conditions are *not* formed e.g. *swimmer/father* and *swimmer/mother* are both neutrally rated pairings. In other words, congruency is not defined for the neutral terms. Consequently, it makes sense to use the design variables (Stereotype by Kinship term gender) as opposed to including a 'Congruency' variable, despite this adding some complication to the interpretation of results.

[22] Cohen's *dz* is calculated as the mean of the differences of two variables divided by the standard deviation of these differences while Cohen's *d* is calculated as the mean of one variable minus the mean of another variable, divided by the square root of the pooled standard deviation of the two variables. Standard interpretation of Cohen's *d* and Cohen's *dz* is as follows: Small effect size = 0.2, Medium effect size = 0.5, Large effect size = 0.8.

Accuracy

The analysis revealed a main effect of Stereotype, F_1 (1.60, 78.31) = 17.12, p < .001; F_2 (2, 32) = 6.91, p = .003, with significantly higher accuracy to word pairs that contained a neutral role term (M = 98.1%), than those that contained male-biased (M = 94.5%) or female-biased terms (M = 94.1%). This result was not surprising as participants should have substantial experience of seeing both men and women occupying neutral roles but less experience of women occupying typically male roles and vice versa. An interaction between Stereotype and Block was also found, F_1 (3.39, 166.27) = 6.80, p < .001; F_2 (4, 64) = 5.67, p = .001 with accuracy improving across Blocks 1 to 3, most notably to word pairs involving female-biased (7.9%) and male-biased (7.4%) role names compared to neutral (2.7%). Again, this is likely due to the latter being judged most accurately from the outset, thus leaving less room for improvement.

A significant Stereotype by Kinship term gender interaction was also found (i.e. an effect of Congruency), F_1 (1.06, 52.02) = 22.77, p < .001; F_2 (2, 32) = 47.61, p < .001. As expected there was significantly lower accuracy to incongruent word pairs (M = 89.7%), compared to congruent (M = 98.9%) or neutral (M = 98.2%) pairs; evidence that participants were influenced by gender stereotype biases[23].

There was also a significant main effect of Block, F_1 (1.19, 58.20) = 19.68, p < .001; F_2 (2, 64) = 72.80, p < .001. Contrasts revealed a linear trend, with a continuous improvement in accuracy from Block 1 to Block 3, F_1 (1, 49) = 21.93, p < .001; F_2 (1, 32) = 138.94, p < .001.

Importantly, an interaction of Block by Congruency was also found, F_1 (1.63, 80.08) = 15.96, p < .001; F_2 (4, 64) = 36.80, p < .001. This interaction indicates that participants' accuracy varied according to the congruency of the word pair presented and the block of trials they were on, a pattern that can be seen in Figure 2.1 below. It is evident that accuracy increased across all three conditions as the blocks progressed, with the most dramatic improvement seen in response to stereotype incongruent pairings (+14.5% across blocks). This pattern suggests

[23] Note that here, and throughout this thesis, *congruent* means are calculated from responses to word pairs comprising female stereotypes with female kinship terms + male stereotypes with male kinship terms, *incongruent* means are calculated from responses to word pairs that presented female stereotypes with male kinship terms + male stereotypes with female kinship terms, and *neutral* means are calculated from responses to word pairs that presented neutrally rated role nouns with female kinship terms + neutrally rated role nouns with male kinship terms. Also, it is worth noting that there are no further differential effects of male versus female role nouns/kinship terms than those outlined in the thesis i.e. on occasion male roles are responded to more accurately/faster than female roles, while on other occasions the opposite pattern of responding emerged, reasons for which remain unknown. However, typically, responses across male and female roles were very similar.

there was a reduction in stereotypic responding following the presentation of performance-related feedback in Block 2.

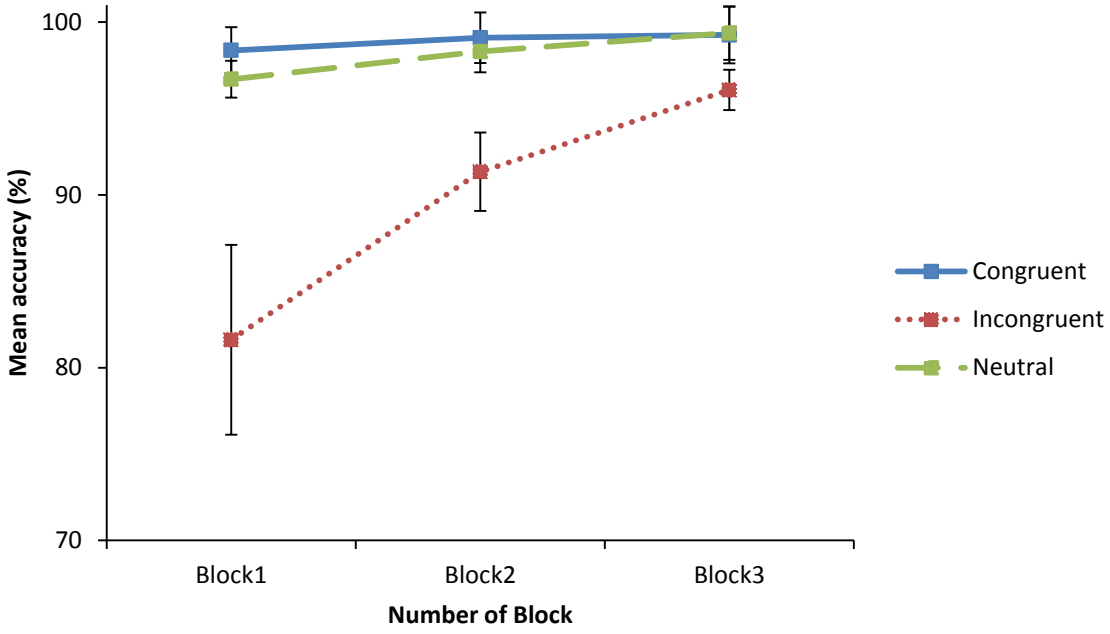


Figure 2.1[24]. Experiment 1: Mean percentages of correct judgements to critical word pairs across blocks. Error bars indicate the 95% confidence intervals.

A series of paired *t*-tests (with Bonferroni corrections) was performed to investigate the Block by Congruency interaction further. It was anticipated that incongruent items would prove more difficult to respond to than congruent items in Block 1, and also that performance on these incongruent pairings would improve after the feedback trainings (i.e. from Block 1 to Block 3). Therefore, one-tailed *t*-tests were used for these comparisons while all remaining differences were examined using two-tailed tests.

As expected, congruent items were responded to significantly more accurately than incongruent items in Block 1, t_1 (50) = 4.91, $p < .001$, $dz = .69$; t_2 (23) = 9.50, $p < .001$, $dz = 1.94$. However, with the feedback training, a significant improvement in accuracy of incongruent pairings from Block 1 to Block 3 was found, t_1 (50) = 4.82, $p < .001$, $dz = .68$; t_2 (23) = 10.97, $p < .001$, $dz = 2.24$, suggesting that the provision of feedback greatly aided performance on these word pairs. Similarly, variance of responses to incongruent pairings was seen to lessen across blocks. However, regardless of this improvement in performance, a significant difference

[24] Note that, unless otherwise stated, all figures in this thesis displaying accuracy data will begin at 70% for purposes of clarity.

between accuracy of congruent and incongruent pairings still remained in Block 3, t_1 (50) = 3.16, $p < .001$, $dz = .44$; t_2 (23) = 4.08, $p < .001$, $dz = .83$. These results suggest that, despite the feedback training, accuracy of incongruent pairings had not yet reached the high level of accuracy found in response to congruent pairings.

Similarly, a significant difference between accuracy of incongruent and neutral pairings[25] was revealed in Block 1, t_1 (50) = 5.01, $p < .001$, $dz = .70$; t_2 (26.58) = 8.05, $p < .001$, $d = 3.12$, but was still evident in Block 3 after the feedback training, t_1 (50) = 3.27, $p = .002$, $dz = .46$; t_2 (26.54) = 3.73, $p = .001$, $d = 1.45$.

Finally, a significant difference between accuracy of congruent and neutral pairings was found in Block 1, t_1 (50) = 2.55, $p = .014$, $dz = .36$; t_2 (44) = 2.25, $p = .030$, $d = .68$, but this disappeared in Block 3 after the training, t_1 (50) = 0.35, $p > .7$; t_2 (44) = 0.32 $p > .7$, as accuracy to neutral pairings increased to fall in line with that of the congruent pairings.

Finally, the analysis revealed a main effect of Participant Gender in the by-items analysis only, F_2 (1, 32) = 11.05, $p = .002$, despite a very small difference in the mean accuracy scores of male and female participants (95.0% vs. 96.2% respectively). Given that there were many fewer participants than items in this experiment (51 participants vs. 456 item pairs per participant) it is highly likely that this effect was only significant by-items because the standard errors of the condition means are likely to be much lower in the by-items analysis than in the by-participants analysis, if the variances are roughly equal. For instance, the average standard deviation of responses to critical word pairs was 8.54% in the by-participants data while just 3.43% in the by-items analysis. It is worth noting that a similar imbalance between subject numbers and item numbers runs through all the studies in this thesis and consequently this kind of pattern (a significant effect by-items but not by-participants) will frequently recur.

[25]Note that in the t_1 analyses, paired *t*-tests were conducted when examining performance on neutral pairings. However, analyses of the neutral terms in the t_2 analyses were less straightforward. Due to the omission of the term *adolescent*, paired *t*-tests were replaced with independent-samples *t*-tests when comparing performance on neutral pairings with that of the other two congruency conditions (as there were now less items in the neutral condition). This procedure was also followed for subsequent experiments in which the term *adolescent* was omitted from analyses.

Response times (RTs)

The analysis revealed a main effect of Congruency[26], F_1 (2, 98) = 35.03, $p < .001$; F_2 (2, 32) = 32.47, $p < .001$, with fastest response times to congruent word pairs (M = 774ms), followed by neutral (M = 808ms) and incongruent pairings (M = 891ms) respectively.

There was also a main effect of Block, F_1 (1.68, 82.1) = 70.54, $p < .001$; F_2 (2, 64) = 225.05, $p < .001$. Contrasts again revealed a linear trend with response times decreasing from Block 1 to 3, F_1 (1, 49) = 123.49, $p < .001$; F_2 (1, 32) = 371.78, $p < .001$.

A significant Block by Congruency interaction was also revealed, F_1 (4, 196) = 8.45, $p < .001$; F_2 (3.75, 59.95) = 5.46, $p = .001$. As can be seen in Figure 2.2 below, RTs decreased across all conditions as the blocks progressed, with the greatest reduction seen in response to incongruent pairings (407ms faster in Block 3 than Block 1 vs. 306ms and 245ms reductions for neutral and congruent pairings respectively).

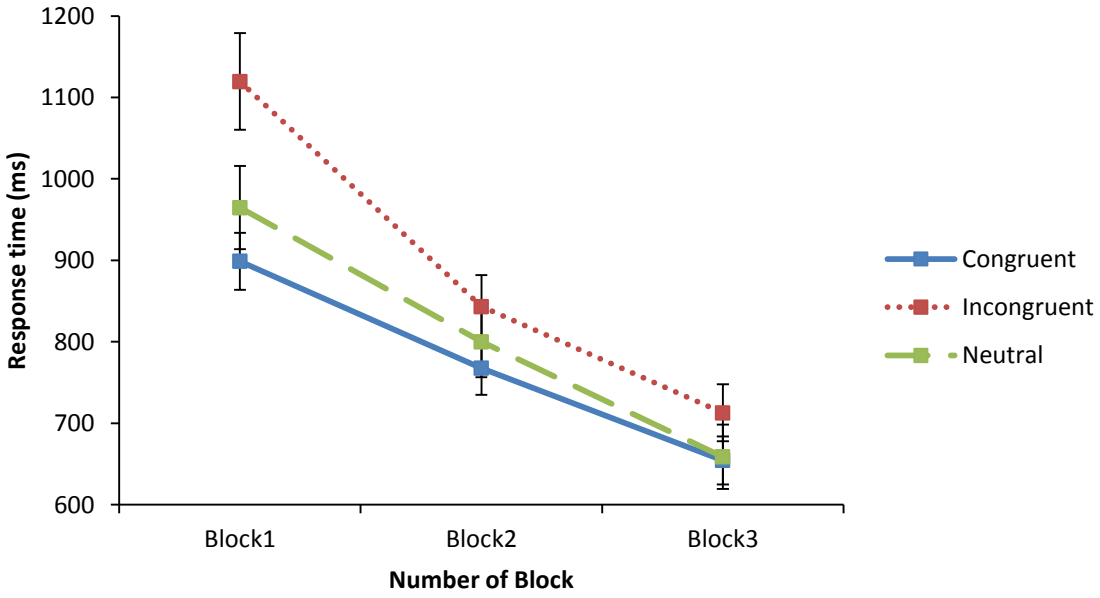


Figure 2.2[27]. Experiment 1: Mean response times (in milliseconds) of correct judgements to critical word pairs across blocks. Error bars indicate the 95% confidence intervals.

Paired *t*-tests (with Bonferroni corrections) were performed to investigate this interaction further, as described with the accuracy data.

[26] i.e. an interaction of Stereotype bias by Kinship term gender.

[27] Note that, unless otherwise stated, all figures in this thesis displaying response time data will begin at 600ms for purposes of clarity.

Congruent items were responded to significantly faster than incongruent items in Block 1: t_1 (50) = 6.83, $p < .001$, $dz = .96$; t_2 (23) = 6.94, $p < .001$, $dz = 1.42$, with this same pattern found in Block 3: t_1 (50) = 2.85, $p = .006$, $dz = .40$; t_2 (23) = 2.60, $p = .016$, $dz = .53$. Therefore, despite significantly faster reaction times to stereotype incongruent pairings from Block 1 to Block 3, t_1 (50) = 10.53, $p < .001$, $dz = 1.47$; t_2 (23) = 10.68, $p < .001$, $dz = 2.18$, these pairings were still responded to more slowly than stereotype congruent pairings. Similarly, a significant difference between RTs of incongruent and neutral pairings was also found in Block 1, t_1 (50) = 5.25, $p < .001$, $dz = .76$; t_2 (34.61) = 3.70, $p < .001$, $d = 1.26$, although effect sizes reveal the magnitude of this difference was reduced by Block 3, t_1 (50) = 2.26, $p = .028$, $dz = .32$; t_2 (44) = 1.88, $p = .067$, $d = .57$.

Finally, a significant difference between RTs of congruent and neutral pairings was found in Block 1, t_1 (50) = 2.61, $p = .012$, $dz = .37$; t_2 (44) = 3.10, $p = .003$, $d = .94$, but this disappeared in Block 3 after the training t_1 (50) = 0.20, $p > .8$; t_1 (44) = 0.26, $p > .7$, as responses to neutral word pairs speeded up to parallel those of stereotype congruent pairings, thus mirroring the results of the accuracy data.

The analysis also revealed a marginal effect of Kinship term gender with faster RTs to word pairs that contained male kinship terms (M = 813ms) than female terms (M = 836ms), F_1 (1, 49) = 3.89, $p = .054$; F_2 (1, 32) = 55.55, $p < .001$. A two-way interaction of Kinship term gender by Participant Gender was also found, F_1 (1, 49) = 18.33, $p < .001$; F_2 (1, 32) = 27.27, $p < .001$. Means revealed that, overall, male participants responded faster to word pairs that involved a male kinship term than a female one (784ms vs. 857ms respectively, mean difference = 73ms), whereas female participants responded faster to word pairs that involved a female kinship term than a male one (814ms vs. 842ms respectively, mean difference = 28ms). This pattern of results suggests a response time advantage for judgements involving kinship terms congruent with a participant's own gender, particularly for male participants**.**

Fillers - Accuracy

Performance on filler trials was somewhat variable with an average of 96.30% accuracy on definitionally matching word pairs versus an average of 88.31% on mismatching word pairs across the experiment[28]. Results of the matching condition are in line with those of Oakhill et

[28] Note that as responding to the filler trials was not the main focus of this thesis, tests of significance were not carried out on this data throughout; only a descriptive analysis of results is presented.

al. (2005) who found accuracy of responses to fillers to be uniformly high at around 95% across conditions. However, in the current study a deterioration in accuracy of the definitionally mismatching word pairs was found. This unexpected finding is driven by poorer accuracy to word pairs involving definitionally male (82.61%) as opposed to definitionally female role names (94.01%). It appears that participants are interpreting certain male terms (e.g. *host, hero*) as generically applicable to both sexes until they are alerted to the fact that they should be stricter in their linguistic definitions – this information is either signalled through performance-related feedback or by encountering the definitionally female counterpart to a male term that may previously have been presented e.g. once the term *hostess* has appeared, the term *host* is less likely to be interpreted generically. Therefore, as opposed to the gender stereotype bias evident in response to stereotype incongruent trials, poor performance on male definitionally mismatching trials suggests participants were responding in a more open manner, accepting the male role terms as suitable referents to both sexes. Indeed, when performance of the mismatching fillers is analysed solely in Block 3 (following the feedback training), accuracy is much more in-line with past findings at 92.88%[29].

Fillers - Response times

With the response time data, average RTs to both male and female definitionally matching word pairs were quite similar (951ms vs. 930ms respectively). However, more variation was again found with the definitionally mismatching word pairs, as average RTs to the female mismatching pairings were 58ms faster than to the male mismatching pairings (945ms vs. 1003ms respectively). This finding supports the pattern seen in the accuracy data, with longer processing times likely to reflect participants' deliberation over certain role nouns, which have female-specific counterparts and which should, therefore, be taken as male specific e.g. *host* or *hero*.

[29] Despite this difficulty shown with certain definitionally masculine role nouns, it was decided to retain these terms in further experiments as performance improved from Block 1 to Block 3 following the introduction of a training strategy, and, as mentioned above, when participants realized that they should be stricter in their linguistic definitions (based on exposure to female-specific versions of these male terms) i.e. participants did not consistently reject these terms but learnt to interpret them based on definitional gender.

2.2.4 Discussion

The results of Experiment 1 provide support for all experimental hypotheses. Firstly, the findings suggest that gender stereotype information is indeed automatically evoked from single words, as has been claimed in past research (Banaji & Hardin, 1996; Oakhill et al., 2005). This was evident with significantly lower accuracy and slower response times to stereotype incongruent word pairs in Block 1 compared to stereotype congruent and neutral word pairs.

Secondly, the results reveal that the inferior performance associated with the stereotype incongruent condition in Block 1 was attenuated in Block 2 and, moreover, that this improvement in performance was extended into Block 3. More specifically, accuracy of judgements to stereotype incongruent word pairs increased significantly across blocks while response times to stereotype incongruent word pairs decreased significantly across blocks. Both of these findings suggest that the introduction of performance-related feedback proved an effective means of overcoming automatic inferences so as to produce lower levels of gender-stereotyped responding.

However, despite this improvement in performance to incongruent word pairs, performance on congruent and neutral pairings remained significantly more accurate and faster than on incongruent word pairs by the end of the experiment. Therefore, while the stereotyping effect was not completely overridden, the above findings do provide further support for the malleability of stereotype biases and provide evidence that such biases can be attenuated with the provision of performance-related feedback.

Overall, creating awareness of an individual's personal stereotype-tendencies through providing feedback on their behaviour appears to be a straightforward means of overcoming the immediate activation of gender stereotypes and allows for reduced stereotype application, in the short term at least. While explicit training strategies have been used to tackle stereotype application in the past, Experiment 1 is unique in that a participant's own performance accuracy was used as a means of both reminding and re-educating themselves about the social roles occupied by women and men. That said, given that the experiment involved a large number of trials, and that the participants could probably easily discern the nature of the pairs they were judging (Oakhill et al., 2005), it was possible that participants simply improved at the task (in particular the stereotype incongruent pairings) as they progressed, benefitting from practice effects.

Therefore, in order to determine whether the improved performance to stereotype incongruent word pairs across blocks was indeed a direct result of the feedback offered, or alternatively due to practice effects, a second experiment was run with the feedback component removed.

2.3 Experiment 2: Control study

2.3.1 Introduction

The purpose of Experiment 2 was to provide a baseline condition against which to compare the results of Experiment 1. In this way it was hoped to establish whether or not the reduction of stereotype bias in Experiment 2 (specifically to stereotype incongruent pairings) was indeed due to the feedback manipulation, or alternatively a result of practice effects, with participants naturally improving as the task went on. To achieve this, the experimental design was kept identical to that of Experiment 1, but with the feedback component in Block 2 of the judgement trials removed[30].

However, it is important to note a distinction between speeding up with practice and becoming less stereotyped with practice. While participants may naturally speed up with practice due to the repetitive nature of a task, becoming less stereotyped requires more deliberate and controlled processes i.e. a person must expend effort into consciously dismissing stereotyped associations, and may then over time show less evidence of succumbing to such biases. Therefore, in Experiment 2, if participants are found to exhibit faster response times across blocks to incongruent pairings, this is not necessarily evidence of reduced stereotyping - an equivalent increase in accuracy would also be required for this.

[30] It should be noted at this point that similar conditions (such as the performance feedback condition of Experiment 1 and the control condition of Experiment 2) were consistently tested in separate experiments (as opposed to with a between-subjects design within one experiment) across this thesis. This approach was taken in an effort to minimise anticipated difficulty with participant recruitment - as all experiments in this thesis involved the same judgment paradigm, different participants were required across each study. Therefore, if a training was not found to effectively reduce stereotyping, results of a control condition were not required i.e. participants were not unnecessarily used by running similar conditions as separate experiments. The different experiments (i.e. conditions) were subsequently examined together in a combined analysis with 'Experiment' included as a between-subjects/within items factor. With this approach, the issues of error variance and different base rates between experiments were examined using t-tests to investigate whether initial Block 1 performance varied significantly across blocks.

As the feedback in Experiment 1 informed participants of their (in)accuracy (thereby alerting them to their personal stereotype biases) it was anticipated that its removal would lower participants' awareness of these biases, and consequently, the likelihood of overcoming them. Therefore, as before, it was hypothesised that participants would respond more slowly and less accurately to stereotype incongruent word pairs (e.g. *nurse/father*) than to stereotype congruent word pairs (*nurse/mother*) in Block 1. However, without the provision of feedback in Block 2, it was further hypothesised that this effect would remain consistent across all three blocks.

2.3.2 Method

Participants

Thirty monolingual, native English speakers (13 male, 17 female) from the student population of the University of Sussex took part. Participants' age ranged from 18 to 31 years (*M*: 21.13; *SD*: 3.24). They received either £5 or 4 course credits for taking part in the session, which lasted approximately 45 minutes.

Materials & Design

The stimuli and design used were identical to those of Experiment 1 (see Section 2.2.2 for details), the only modification being the removal of the feedback component after responses in Block 2. Therefore participants completed three blocks of the judgement trials without receiving any performance-related information.

Procedure

The procedure was identical to that of Experiment 1, but with instructions updated as regards the removal of all feedback-related information.

2.3.3 Results

Data screening

As before, the analyses reported below were carried out following exclusion of the neutral term *adolescent* over concerns that low accuracy (68% in Block 1, 76% overall) to word pairs containing this term stemmed from considerations of age appropriateness over gender. This resulted in removing 1.32% of the data.

Response times below 150ms, and above 4,000ms were identified and excluded prior to analysis (2.36% of the total data) along with times for all errors of judgement (a further 11.89%)[31] totalling a loss of 14.25% of the data. Further data trimming procedures were conducted as outlined in Section 2.2.3.

As in Experiment 1, accuracy and response times of judgements were examined using two mixed-design ANOVAs with further details again explained in Section 2.2.3.

Accuracy

A main effect of Stereotype was found, F_1 (1.64, 45.94) = 13.24, $p < .001$; F_2 (2, 32) = 8.91, p = .001, with higher accuracy to word pairs that contained a neutral role term (M = 94.7%), than those that contained male-biased (M = 87.6%) or female-biased terms (M = 87.9%).

A main effect of Congruency[32] was also found, F_1 (1.04, 29.14) = 19.30, $p < .001$; F_2 (2, 32) = 87.44, $p < .001$. As expected, accuracy was lower to stereotype incongruent word pairs (M = 78.7%), than congruent (M = 98.2%) and neutral (M = 94.7%) pairs.

Importantly, neither a main effect of Block nor interaction of Congruency by Block was revealed, ps > .1. Upon examination of the data in Figure 2.3 below it is clear there was no significant improvement in accuracy of responses to stereotype incongruent word pairs across blocks, t_1 (29) = 0.51, p = .610, dz = .09; t_2 (23) = 0.70, p = .491, dz =.14, while variance of responses also remained consistently high throughout. These results suggest that the steady increase in accuracy towards stereotype incongruent word pairs in Experiment 1 was indeed due to the provision of performance-related feedback, and not just a result of practice.

[31] This is a 4.33% greater loss in data than in Experiment 1. However, this is not surprising; as no improvement was made across blocks, more errors occurred and their accompanying times were replaced.

[32] i.e. an interaction of Stereotype bias by Kinship term gender.

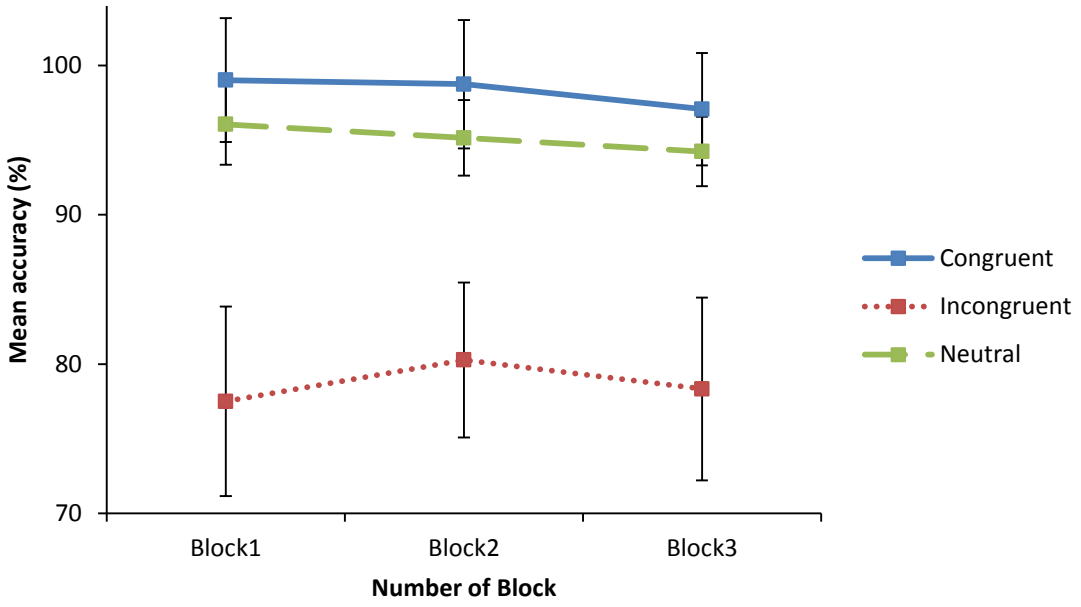


Figure 2.3. Experiment 2: Mean percentages of correct judgements to critical word pairs across blocks. Error bars indicate the 95% confidence intervals.

Unexpectedly, a main effect of Participant Gender was also found, F_1 (1, 28) = 4.71, p = .039; F_2 (1, 32) = 97.34, p < .001, with females responding more accurately than males (95.0% vs. 85.2% respectively) on critical trials[33].

A two-way interaction of Kinship term gender by Participant Gender was also found, F_1 (1, 28) = 4.84, p = .036; F_2 (1, 32) = 9.78, p = .004, with male participants responding more accurately to word pairs involving a male kinship term than a female one (87.0% vs. 83.4% respectively), while female participants responded slightly more accurately to word pairs that involved a female kinship term rather than male term (95.5% vs. 94.4% respectively). This pattern of results was also found in the RT data of Experiment 1, and again reveals a performance advantage for kinship terms congruent with a participant's own gender, particularly for the male participants. It was also evident above that female participants responded more accurately than male participants across both male and female kinship terms.

Finally, a marginally significant interaction of Congruency by Participant Gender was also revealed, F_1 (1.04, 29.14) = 4.07, p = .052; F_2 (2, 32) = 57.05, p < .001, with the pattern of responding displayed in Figure 2.4 below. Firstly, a significant difference emerged between accuracy of responding to female congruent pairings by female (M = 99.35%) and male (M =

[33] This is a surprising finding which was not anticipated based on findings from Experiment 1 or previous studies involving this paradigm by Oakhill et al. (2005).

95.73%) participants, t_1 (15.76) = 2.63, p = .019, d = 1.32. Female participants also outperformed males in response to male incongruent pairings (M = 88.73% vs. 63.89% respectively), t_1 (17.09) = 2.08, p = .053, d = 1.0, suggesting that male participants succumbed to stereotype biases to a greater extent than females in the absence of performance feedback. The error bars reveal that both sexes displayed greater variability in responses to incongruent pairings as opposed to the congruent pairings, but also that the accuracy of male participants typically varied to a greater extent than that of the female participants.

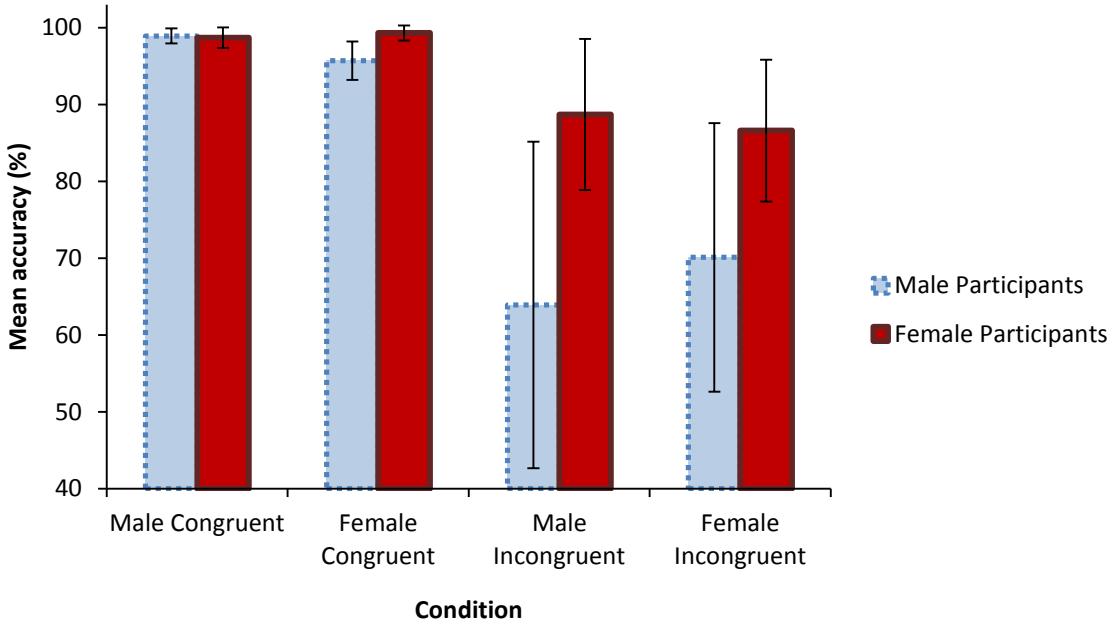


Figure 2.4. Experiment 2: Mean accuracy of judgements of male and female participants to congruent and incongruent word pairs. Error bars indicate the 95% confidence intervals. The vertical axis begins at 40% for purposes of clarity.

Response times

A significant main effect of Congruency[34] was found, F_1 (2, 56) = 23.10, p < .001; F_2 (2, 32) = 19.88, p < .001, with fastest response times to congruent word pairs (M = 820ms), followed by neutral (M = 855ms) and incongruent pairings (M = 949ms) respectively.

There was also a main effect of Block, F_1 (2, 56) = 16.12, p < .001; F_2 (2, 64) = 30.11, p < .001. Contrasts revealed a linear trend with response times decreasing from Block 1 to 3 respectively, F_1 (1, 28) = 23.36, p < .001; F_2 (1, 32) = 53.88, p < .001.

[34] i.e. an interaction of Stereotype bias by Kinship term gender.

A significant interaction of Block by Congruency was also found, F_1 (2.98, 83.42) = 6.75, p <.001; F_2 (4, 64) = 3.0, p = .025. As can be seen in Figure 2.5 below, RTs decreased across all three conditions as the blocks progressed. The greatest reduction was found in response to stereotype incongruent pairings with a significant decrease in response times of 247ms from Block 1 to Block 3, t_1 (29) = 4.42, p < .001, dz = .81; t_2 (23) = 5.12, p < .001, dz = 1.05. Response times to neutral pairings also decreased significantly across blocks (189ms), t_1 (29) = 4.93, p < .001, dz = .90; t_2 (21) = 5.70, p < .001, dz = 1.22, as did RTs to stereotype congruent pairings (81ms), t_1 (29) = 2.12, p = .042, dz = .39; t_2 (23) = 3.47, p = .002, dz = .71. Although these RT improvements are not as large as those reported in Experiment 1, it is evident that RTs *do* naturally decrease with practice across blocks.

By the end of the experiment in Block 3, it was also found that RTs to stereotype incongruent word pairs remained significantly slower than to neutral word pairs, t_1 (29) = 2.27, p =.031, dz = .41; t_2 (44) = 1.81, p = .081, d = .55, and stereotype congruent word pairs, t_1 (29) = 2.73, p = .011, dz = .50; t_2 (23) = 1.72, p = .099, d = .72, although these effects were only marginally significant in the by-items analysis.

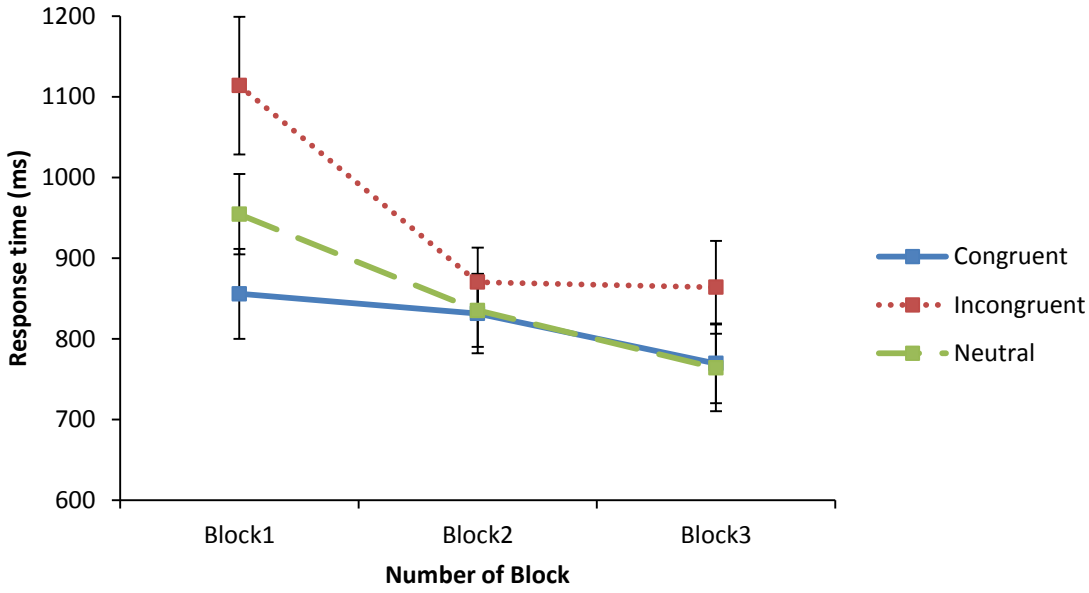


Figure 2.5. Experiment 2: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks. Error bars indicate the 95% confidence intervals.

There was a main effect of Participant Gender in the by-items analysis only, F_2 (1, 32) = 21.04, p < .001, with females faster to respond than males (829ms vs. 908ms respectively). Similarly, as with the accuracy data, an interaction of Congruency by Participant Gender was found but

in the by-participants data only (marginally significant by-items), F_1 (2, 56) = 4.79, p = .012; F_2 (2, 32) = 2.61, p = .089. As seen in Figure 2.6 below, both sexes responded faster in the congruent condition than the incongruent condition, with the male participants typically much slower at responding than the female participants (e.g. M = 1046ms vs. 875ms respectively to stereotype incongruent pairings). However, it should be noted that variance of responding was relatively high for both sexes across all conditions.

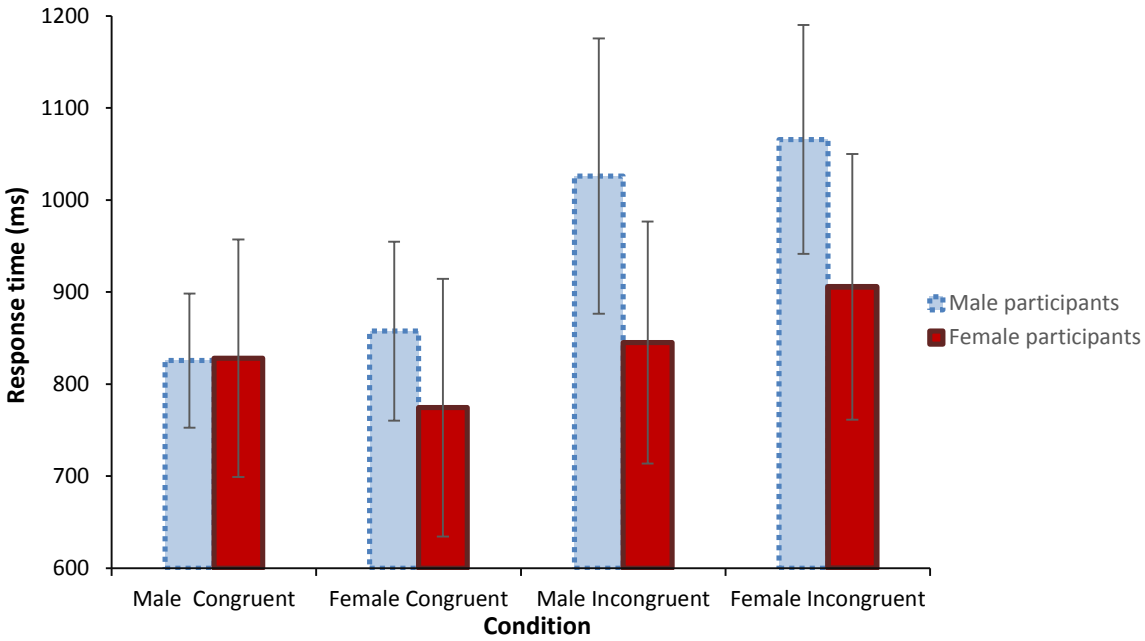


Figure 2.6. Experiment 2: Mean response times (in milliseconds) of male and female participants to congruent and incongruent word pairs. Error bars indicate the 95% confidence intervals.

Combined with the accuracy data, these Congruency by Participant Gender interactions suggest that in the absence of feedback male participants struggle with the stereotype incongruent word pairs to a greater extent than females; despite taking longer to make a response, male participants frequently made more incorrect judgements than females. However, an inspection of Block 1 performance to these incongruent pairings revealed that male participants' accuracy was much lower from the beginning in Experiment 2 than Experiment 1 (65.7% vs. 80% respectively), while such variation was not found with the female participants' scores (83.2% vs. 86.6% respectively). This same pattern was found to a lesser extent with the RT data as male participants' RTs were slower from the beginning in Experiment 2 than Experiment 1 (1206ms vs. 1078ms respectively, mean difference of 128ms). However, the opposite trend was found with the female participants as they achieved faster

RTs from the beginning of Experiment 2 than Experiment 1 (1043ms vs. 1158ms respectively, mean difference of 115ms). Overall, the reason(s) for which male participants displayed particularly variable performance remain unclear, as both experiments were very similar at this pre-training point, differing only in information provided about the feedback training in Block 2 of Experiment 1.

Fillers - Accuracy

As in Experiment 1, participants showed higher levels of accuracy to definitionally matching word pairs (*M* = 94.7%) than to definitionally mismatching word pairs (*M* = 83.9%). The latter result is again driven by poorer accuracy to word pairs involving definitionally male role names (*M* = 74.5%) as opposed to definitionally female role names (*M* = 93.2%), likely due to participants interpreting male terms generically throughout the experiment.

However, unlike Experiment 1 in which participants received feedback alerting them of their errors, no such guidance was provided in this experiment. Consequently, participants in Experiment 1 reached higher accuracy on definitionally mismatching pairings in Block 3 (*M* = 92.9%) than in this control experiment (*M* = 76.0%). Although this finding was somewhat unexpected, it is further support for the use of performance-related feedback as a means of alerting participants to their use of gender-related information.

Fillers - Response times

The response time data tell a similar story. Average reaction times to male and female definitionally matching word pairs were very similar (*M* = 935ms vs. *M* = 921ms respectively) across the experiment. However, more variable performance was found with the definitionally mismatching word pairs. Reaction times to the female mismatch pairings were faster (*M* = 960ms) than to the male mismatch pairs (*M* = 1055ms) by 95ms. This finding is again thought to be indicative of participants' deliberation over certain, definitionally male, terms that are nowadays often used in reference to both sexes.

Moreover, participants in Experiment 1 were faster to respond to definitionally mismatching pairings in Block 3 (*M* = 849ms) than in this control experiment (*M* = 901ms). This finding is in line with the accuracy data and provides further support for the use of performance-related feedback as an aid to performance on this judgement task.

Overall, the data of Experiment 2 suggest that performance related feedback resulted in better task performance in Experiment 1 relative to this control study. However, whether or not these experimental differences are statistically significant will be later explored in Section 2.5.

2.3.4 Discussion

The current control experiment revealed that, in the absence of feedback, participants do not overcome automatic activation of stereotype biases in their accuracy judgements. It was found that accuracy of responses to stereotype incongruent word pairs was significantly lower than to stereotype congruent and neutral pairs from the outset. More importantly, this pattern remained consistent across blocks, with no hint of the Congruency by Block interaction that was found in Experiment 1.

With the response time data it was found that speed of response decreased across blocks in each of the three congruency conditions, despite the fact that no feedback was provided. This suggests that participants benefitted from a practice effect, naturally speeding up at the task as the experiment progressed. Nevertheless, RTs to stereotype incongruent pairings remained significantly slower than to stereotype congruent and neutral word pairs by the end of the experiment, and were ultimately still slower than those reported in Experiment 1.

Male participants in particular exhibited difficulty with the incongruent word pairs in this study, with lower accuracy and slower response times to these pairings than the female participants. This pattern of responding was not found in Experiment 1 as both male and female participants improved performance across blocks upon the provision of performance feedback (with females achieving higher accuracy than males in the by-items analysis). As mentioned earlier, this effect was primarily driven by much lower Block 1 accuracy for male participants in Experiment 2 than Experiment 1. However, the reasons for this variability remain unclear as both experimental designs were almost identical up until Block 2 where a training strategy was introduced (Experiment 1) or not (Experiment 2).

Overall, these results suggest that the improvement in response accuracy to stereotype incongruent word pairs in Experiment 1 was indeed a result of the feedback provided. Furthermore, although faster RTs in Experiment 1 were, at least to some extent, likely to result from practice effects, the data also show that the performance-feedback facilitated RTs to critical incongruent pairings in particular. Therefore, on the whole, it is concluded that performance-related feedback proved an effective means of lowering stereotype application.

By alerting participants to their personal stereotype biases, and helping them ignore or avert these biases through feedback, performance was found to significantly improve across blocks, relative to this control experiment. However, having established the value of this stereotype-reduction strategy in the short-term, two important issues remained unaddressed (a) whether the success of this training generalises beyond the stimuli on which feedback was received i.e. to other terms with an equally strong gender-stereotype bias and (b) the durability of the training results – would they persist**,** at least over a period of days? Experiment 3 was designed to investigate these issues.

2.4 Experiment 3: Long term and transfer effects

2.4.1 Introduction

Many studies that claim to reduce stereotype activation and application frequently examine only the short-term effects of their manipulations. Indeed, in their meta-analytic reviews of the stereotyping and prejudice literature, Lenton et al. (2009) and Paluck and Green (2009) have both highlighted an over-reliance on single session studies in past research, with the latter authors terming them 'quick fixes' (p. 349). Considering this dearth of literature on the durability of stereotype-reduction effects, the following experiment set out to address this issue and examine the value of performance-feedback as a longer-term stereotype-reduction training.

A second aim of the study was to investigate the generalisability of the feedback training. Again, few past studies have provided evidence that the effects of their training extend beyond the immediate training context and influence responses to novel stimuli (exceptions include Kawakami et al., 2000, and Blair et al., 2001). Therefore, this study aimed to address this issue and examine whether participants would continue to exhibit improved performance to stereotype incongruent trials in Block 3 (after the feedback training), when presented with word pairs containing novel stereotype-biased role nouns as opposed to the role nouns already presented in Blocks 1 and 2 (as was the case in Experiment 1).

To investigate these issues, the experimental design of this study was kept identical to that of Experiment 1, with two exceptions: (1) an entirely new set of male-biased, female-biased and neutrally rated role terms was introduced in Block 3 so as to investigate whether the feedback training extended to a new set of stimuli and (2) participants were asked to return to the

laboratory one week after Session 1 to complete one final block of the judgement trials, in order to assess the durability of the training effects.

Specifically, it was again hypothesised that participants would respond more slowly and less accurately to trials made up of stereotype incongruent word pairs (e.g. *nurse/father*) than to stereotype congruent word pairs (*nurse/mother*) in Block 1. However, on receipt of feedback in Block 2, it was hypothesised that this stereotyping effect would be reduced, and improved performance found in Block 3, despite the introduction of new stimuli. Finally, it was hypothesised that this improved performance would also be evident one week later with participants having learnt to exert control over automatic gender stereotype biases.

2.4.2 Method

Participants

Thirty-six monolingual, native English speakers (18 male, 18 female) from the student population of the University of Sussex participated. Participants' age ranged from 18 to 40 years (*M*: 19.89; *SD*: 4.53). They received either £8 or 6 course credits for taking part in this two-part study. Session 1 lasted approximately 45minutes while Session 2 lasted approximately 15 minutes.

Materials & Design

Session 1

Session 1 consisted of three blocks of judgement trials, with performance-related feedback provided in Block 2 only. While the materials and design were largely as described in Section 2.2.2, the experimental stimuli were adapted to incorporate a new set of critical role nouns in Block 3. To achieve this, an additional group of 12 male-biased, 12 female-biased and 12 neutrally rated role terms was combined with those used in Experiments 1 and 2. The majority of these terms were again sourced from Gabriel et al. (2008) while some were also selected from Kennison and Trofe (2003) (A full list of the original and new terms is provided in Appendix 6 along with the mean bias ratings for each term). As role nouns with the strongest stereotype bias were originally selected for use in Experiments 1 and 2, the stereotype ratings of this new set were necessarily weaker (although were again chosen based on their relatively high stereotype ratings). Therefore, to ensure 2 groups of equal stereotype bias were formed, the two sets of role nouns were combined and then individually matched based on their

stereotype ratings e.g. the two strongest male stereotype role names were paired together, followed by the next two strongest and so on. One set of role nouns was presented in Blocks 1 and 2 while the other was presented in Blocks 3 and 4 (counterbalanced across participants).

Finally, in this study the neutral term *swimmer* replaced *adolescent* as there was evidence that the latter term was being responded to based on age considerations over gender considerations in Experiments 1 and 2.

Session 2

This session took place one week after each participant's first session and involved one block of the judgement trials[35]. This block was the same block of trials that participants had completed in Block 3 of Session 1 i.e. it contained the terms they had not previously received feedback on.

Procedure

The experimental procedure was exactly as described in Section 2.2.2 but with instructions updated so as to remind participants to return one week later and take part in Session 2. Also, participants were reminded that they would not be fully debriefed as to what the experiment was investigating until the end of Session 2 (but told they were free to withdraw from the study at any stage).

2.4.3 Results

Data screening

Response times below 150ms, and above 4,000ms were identified and removed before analysis (representing 2.61% of the total data in Session 1, 2.76% in Session 2) along with times for all errors of judgement (representing a further 7.52% in Session 1, 4.42% in Session 2) totalling a loss of 10.12% of the data in Session 1 and 7.18% in Session 2.

Further data trimming measures and two mixed-design ANOVAs were again conducted as outlined in Section 2.2.3, but with the Block factor updated so as to incorporate 4 blocks as opposed to 3. Although two different sets of critical role nouns were used for Blocks 1 and 2

[35] One participant returned 8 days after Session 1 while all others returned 7 days after Session 1.

vs. Blocks 3 and 4, Block remained a within-items factor in the following analyses as all role nouns were individually matched for strength of stereotype bias i.e. a related design was used.

Accuracy

As in Experiments 1 and 2, a main effect of Stereotype was found, F_1 (1.2, 41.39) = 12.96, p < .001; F_2 (2, 33) = 9.52, p = .001 with higher accuracy to word pairs involving a neutral role name (97.3%) as opposed to a male-biased (93.4%) or a female-biased role name (92.4%).

There was also a main effect of Block, F_1 (2.09, 71.16) = 4.87, p = .009; F_2 (1.66, 54.73) = 18.38, p < .001, with contrasts revealing a significant linear trend, F_1 (1, 34) = 8.29, p = .007; F_2 (1, 33) = 27.79, p < .001.

Again, there was a main effect of Congruency[36], F_1 (1.02, 34.52) = 12.17, p = .001; F_2 (2, 33) = 82.21, p < .001, with higher accuracy to stereotype congruent (M = 98.45%) and neutral word pairs (M = 97.30%) than stereotype incongruent word pairs (M = 87.45%). Importantly, there was also a significant interaction of Congruency by Block, F_1 (3.33, 113.28) = 4.70, p = .003; F_2 (4.02, 66.39) = 5.95, p < .001. Accuracy performance across blocks in each of the three congruency conditions is displayed in Figure 2.7 below.

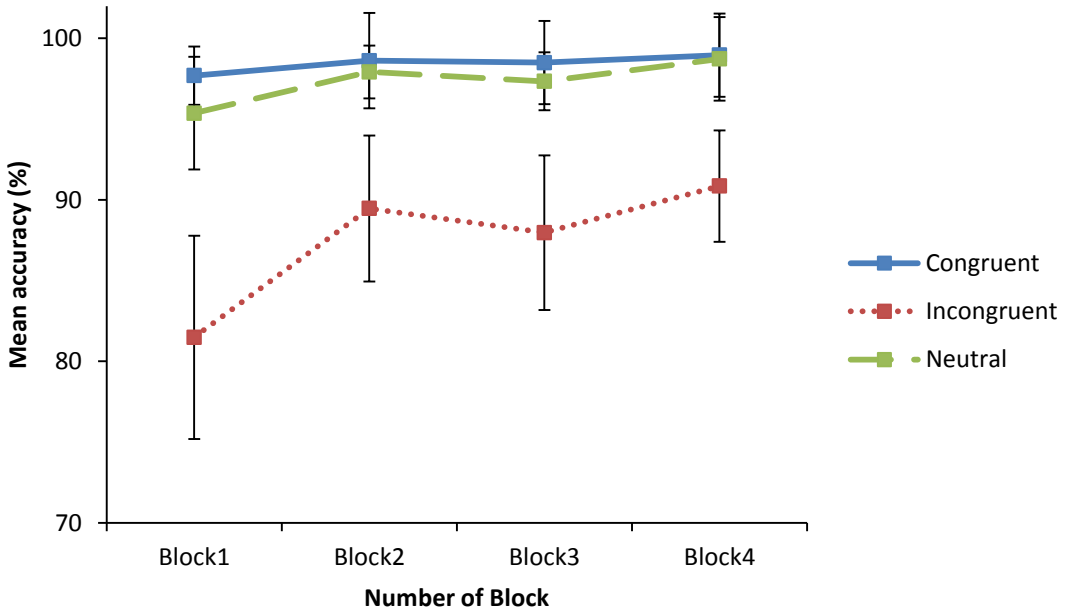


Figure 2.7. Experiment 3: Mean percentages of correct judgements to critical word pairs across blocks. Error bars indicate the 95% confidence intervals.

[36] i.e. an interaction of Stereotype bias by Kinship term gender.

As expected, significantly higher accuracy to stereotype congruent (*M* = 97.7%) and neutral word pairs (*M* = 95.4%) versus stereotype incongruent word pairs (*M* = 81.5%) was found in Block 1 (*p*s <.001[37]).

Generalisability

To investigate whether the effects of this feedback training generalised to a new set of items, accuracy of stereotype incongruent pairings in Block 1 vs. Block 3 was first examined (i.e. pre-training accuracy vs. post training accuracy). A significant 6.5% increase in accuracy was found across these blocks, t_1 (35) = 2.37, p = .024, dz = .39; t_2 (23) = 4.41, p < .001, dz =.90, suggesting that performance-related feedback *does* successfully help reduce stereotype application to a new set of items.

Durability

The durability of this stereotype-reduction effect was next examined. To investigate this, performance accuracy between Block 3 and Block 4 (which took place one week later) of the judgement trials was analysed. No significant difference between stereotype incongruent performance on these two blocks was found in the by-participants analysis, yet was found in the by-items analysis, t_1 (35) = 1.57, p = .125; t_2 (23) = 4.02, p = .001, dz = .82, with accuracy rising a further 2.9% from Block 3 to Block 4 in both sets of analyses.

However, despite these encouraging findings, accuracy to stereotype incongruent word pairs remained significantly lower than to congruent, t_1 (35) = 2.92, p = .006, dz = .49; t_2 (23) = 5.97, p < .001, dz = 1.22, and neutral pairings, t_1 (35) = 2.87, p = .007, dz = .48; t_2 (23) = 4.87, p < .001, dz = .99, respectively by the end of the experiment (Block 4). It should also be noted that accuracy to congruent and neutral word pairs was excellent from the outset (with a mean accuracy of 96.9% and 94.4% in Block 1 respectively) and had little scope for improvement across the experiment.

Participant Gender

Next, a series of findings involving Participant Gender also emerged. Firstly, there was a three-way interaction of Stereotype by Block by Participant Gender, F_1 (3.51, 119.23) = 2.84, p =.033, F_2 (5.36, 88.48) = 4.34, p = .001. Female participants were found to outperform males across blocks in each of the stereotyped conditions, with males scoring particularly poorly in response

[37] Congruent vs. Incongruent: t_1 (35) = 4.13, p < .001, dz = .69; t_2 (23) = 8.33, p < .001, dz = 1.70 and Neutral vs. Incongruent: t_1 (35) = 3.62, p < .001, dz = .60; t_2 (23) = 7.91, p < .001, dz = 1.62.

to female stereotyped terms. Performance was also found to consistently improve across blocks (except that performance of male participants slightly deteriorated from Block 2 to Block 3 upon the introduction of novel stimuli).

However, other effects involving Participant Gender were significant in the by-items analysis only. There was an interaction of Participant Gender by Congruency, F_2 (1, 33) = 18.88, p < .001. This interaction was primarily driven by poorer accuracy of male participants to male stereotype incongruent word pairs (84.5%) than female participants on these pairings (92.0%). Similarly there was an interaction of Participant Gender by Kinship term gender, F_2 (1, 33) = 8.32, p =.007 and an interaction of Participant Gender by Stereotype bias, F_2 (2, 33) = 7.80, p =.002, with male participants achieving lower accuracy scores to both sets of kinship terms and to all three sets of critical role nouns than female participants. Again, it is posited that these effects were revealed in the by-items analysis only because of lower levels of variance than in the by-participants analysis.

Response times

The response time data were next analysed in the same way as the accuracy data.

A main effect of Block was first found, F_1 (2.77, 94.17) = 30.03, p < .001; F_2 (2.58, 84.97) = 91.95, p < .001, with response times steadily decreasing across blocks. Contrasts revealed this was a significant linear trend, F_1 (1, 34) = 58.17, p < .001; F_2 (1, 33) = 162.69, p < .001.

Again, there was a main effect of Congruency[38], F_1 (1.42, 48.25) = 20.12, p < .001; F_2 (2, 33) = 20.82, p < .001, with faster responses on average to stereotype congruent (M = 807ms) and neutral word pairs (M = 846ms) than to stereotype incongruent word pairs (M = 920ms). While a marginally significant interaction of Congruency by Block was found in the by-participants analysis, F_1 (3.47, 117.91) = 2.31, p = .07, this interaction was highly significant in the by-items analysis; F_2 (6, 99) = 3.20, p = .007. The pattern of RTs across blocks in each of the three congruency conditions (for the by-participants analysis) is shown in Figure 2.8 below.

[38] i.e. an interaction of Stereotype bias by Kinship term gender.

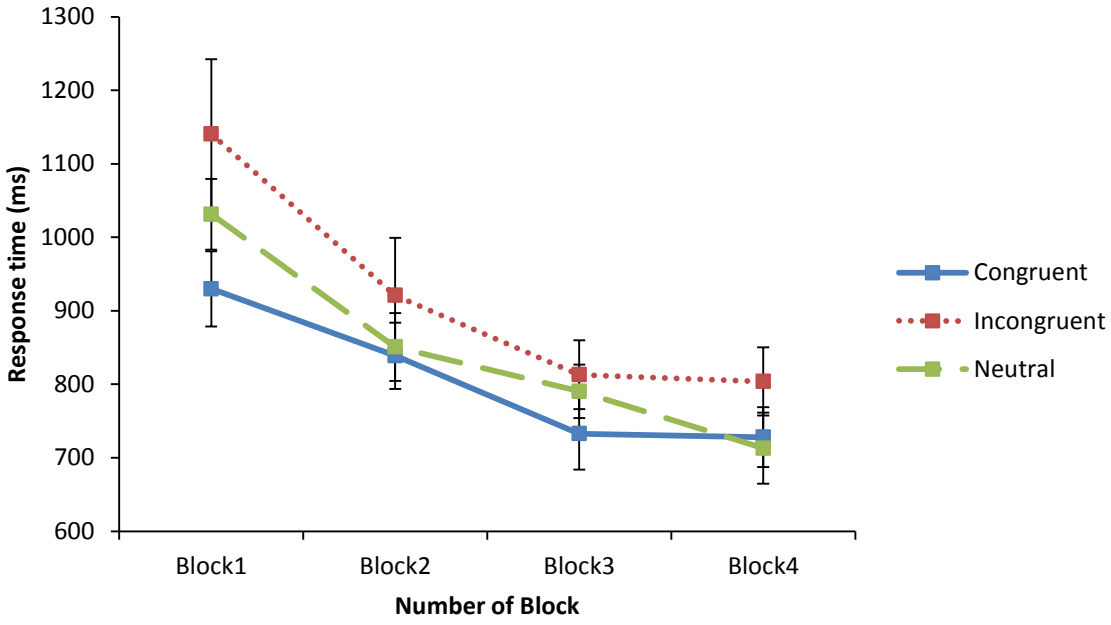


Figure 2.8. Experiment 3: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks. Error bars indicate the 95% confidence intervals.

As expected, response times to stereotype congruent word pairs were significantly faster than to stereotype incongruent word pairs in Block 1, t_1 (35) = 3.76, p = .001, dz = .63; t_2 (23) = 5.14, p < .001, dz = 1.05, while there was a marginally significant difference of RTs between stereotype incongruent and neutral word pairs in the by-participants analysis only, t_1 (35) = 1.90, p = .066, dz = .32; t_2 (23) = 1.55, p = .135, dz = .32, although effect sizes were identical for both sets of analysis.

Generalisability

In Experiment 1 it was found that, following the feedback training in Block 2, response times to stereotype incongruent pairings decreased in Block 3. This same pattern was found in the control experiment, although to a lesser extent. However, to investigate whether this feedback training also facilitated responding to a novel set of stereotype incongruent role names, pre-training RTs (Block 1) versus post-training RTs (Block 3) were first examined.

Response times to stereotype incongruent role names did significantly decrease 327ms across the study from Block 1 to Block 3, t_1 (35) = 4.78, p < .001, dz = .94; t_2 (23) = 7.33, p < .001, dz =

.93, suggesting that performance-related feedback does facilitate responding to a novel set of stereotyped items in this judgement task[39].

Durability

Next, to investigate the durability of the training effect, RT performance between stereotype incongruent word pairs in Block 3 and Block 4 was compared. No significant difference between these two blocks was found, t_1 (35) = 0.31, p = .758; t_2 (23) = 0.64, p = .526, with response times at a very similar level across both (813ms vs. 804ms respectively).

However, while RTs in all three congruency conditions decreased across blocks in a largely linear fashion, RTs to stereotype incongruent word pairs remained significantly slower than to congruent, t_1 (35) = 3.90, p < .001, dz = .65; t_2 (23) = 3.20, p = .004, dz = .65, and neutral word pairs in Block 4, t_1 (35) = 3.85, p < .001, dz = .64, t_2 (23) = 3.49, p = .002, dz = .71.

Overall, it is evident that while performance-related feedback proved a useful aid for decreasing RTs to stereotype incongruent pairings, this training did not wholly succeed in overcoming the processing latencies induced by stereotype bias.

Participant Gender

There was a marginally significant and highly significant effect of Participant Gender in the by-participants and by-items analyses respectively, F_1 (1, 34) = 3.19, p =.083; F_2 (1, 33) = 105.70, p < .001, with female participants responding faster than males overall (780ms vs. 935ms respectively).

An interaction of Stereotype by Participant Gender was also revealed, F_1 (2, 68) = 3.78, p = .028; F_2 (2, 33) = 3.53, p = .041, with female participants slower to respond to word pairs involving a male stereotype term (799ms) than female (770ms) while male participants showed the opposite pattern, responding slower to word pairs involving a female stereotype term (965ms) than male (919ms). Response times to neutral terms fell between these two stereotype conditions for both sets of participants. Similarly, an interaction of Kinship term gender by Participant Gender was found, F_1 (1, 34) = 8.91, p =.005; F_2 (1, 33) = 4.33, p = .045. Female participants again responded faster to word pairs that involved a female kinship term (760ms) as opposed to a male kinship term (800ms) while male participants showed the

[39] Response times to stereotype congruent and neutral role names also significantly decreased across the study from Block 1 to Block 3, t_1 (35) = 5.60, p <.001, dz = .93; t_2 (23) = 7.12, p <.001, dz = 1.45 vs. t_1 (35) = 7.43, p <.001, dz = 1.24; t_2 (23) = 7.97, p <.001, dz = 1.63 respectively.

opposite pattern, responding faster to word pairs that involved a male kinship term (921ms) than a female kinship term (949ms).

Finally, there was also a significant three-way interaction of Congruency by Block by Participant Gender in the by-items analysis only, F_2 (6, 99) = 2.60, p = .022, driven by performance on neutral pairings, as male participants made a much greater improvement across blocks (404ms) than females (213ms).

Overall, the above effects involving Participant Gender suggest a response time advantage when responding to terms that are congruent with a participant's own sex, but with females typically faster to respond than men.

Fillers - Accuracy

As expected, participants showed higher levels of accuracy to definitionally matching word pairs (*M* = 96.00%) than definitionally mismatching word pairs (*M* = 90.09%) overall. The lower accuracy to mismatching pairs was again more apparent with male role terms (*M* = 86.39%) than female role terms (*M* = 95.33%). As before, it is posited that this relatively poor performance to male mismatch fillers is due to participants interpreting male terms generically in Block 1, before learning to become stricter in their linguistic definitions of these terms. Indeed, with the male definitionally mismatching word pairs, accuracy improves 23.78% from Block 1 to Block 4 compared to just 4.89% for female definitionally mismatching word pairs (ultimately resulting in final mean scores that were very similar; 95.8% and 97.6% respectively). These means are somewhat higher than the average accuracy to definitionally mismatching pairings attained in the final block of Experiment 1, *M* = 92.88% (and unsurprisingly, the control, *M* = 76%). This improved performance is likely due to participants completing 4 blocks of trials in the current experiment as opposed to 3 in previous experiments (as accuracy of the current fillers was 92.33% in Block 3).

Fillers - Response times

Overall, participants were faster to respond to definitionally matching word pairs (*M* = 956ms) than definitionally mismatching word pairs (*M* = 991ms). However, unlike Experiment 1 and 2, in which average reaction times to male and female definitionally matching word pairs were very similar[40], in this experiment participants were noticeably faster at responding to female

[40] Experiment 1: *M* = 931ms for female, vs. *M* = 951ms for male definitionally matching word pairs; Experiment 2: *M* = 921ms for female vs. *M* = 935ms for male definitionally matching word pairs.

definitionally matching word pairs (*M* = 868ms) than male pairs (*M* = 1043ms; mean difference = 175ms).

Response times to the definitionally mismatching word pairs showed a similar pattern. Average reaction times to the female mismatch pairings were faster (*M* = 937ms) than to the male mismatch pairs (*M* = 1045ms; mean difference = 108ms). As in the previous experiments, it is posited that slower RTs to male terms over female is likely to reflect participants' deliberation over certain role nouns that are male-specific by definition but which are frequently used to refer to both sexes.

Overall it should be noted that mean RTs to male mismatch pairings are a mere 2ms slower than to male match pairings. The reason for these unusually slow RTs to male definitionally matching word pairs remain unclear, as theoretically, they should not prove difficult for participants. Nevertheless, it seems plausible that, as with the mismatch pairings, participants may simply have taken longer to consider these pairs in an effort to successfully respond. This idea is supported by the accuracy data, as average accuracy to the male definitionally matching fillers was very high from the outset (*M* = 95.3% in Block 1, *M* = 96% overall).

Overall there was a large reduction in response times across the blocks to definitionally mismatching word pairs, with participants responding an average of 418ms faster from Block 1 to Block 4.

2.4.4 Discussion

This two-part experiment again confirmed that the provision of performance-related feedback is an effective means of overcoming immediate activation of gender stereotype biases and results in reduced levels of stereotype application.

It was found that accuracy scores improved significantly from Block 1 to Block 3 indicating that this training not only assisted performance on stereotyped items that participants had previously received feedback on (as was found in Experiment 1), but also helped performance on an entirely new set of stereotyped terms. This finding suggests that by highlighting participants' stereotyping tendencies, they learn not only to exert control over stereotype biases to a particular set of role nouns, but to extend this strategy so as to exert control over further gender-biased role nouns they encounter.

Strategies for reducing activation of gender stereotypes are rarely examined for the durability of their effects. This study sought to address this issue by examining performance on stereotype incongruent pairings one week after the initial training. It was found that the high level of accuracy found in Block 3 in response to stereotype incongruent word pairs was still evident one week later, providing strong support for the use of performance-related feedback as a useful strategy for combating the initial activation of stereotype bias in the longer term. That said, further scope for improvement was also evident, as accuracy on stereotype congruent and neutral pairs remained greater than that of incongruent word pairs in the final block of trials.

With the response time data, it was found that participants responded increasingly quickly across blocks, with a similar pattern found in all three congruency conditions. Importantly, response times were faster in Block 3 (following the feedback training) compared to in Block 1, despite the fact that participants were responding to novel stereotyped items.

Furthermore, although response times did not improve significantly one week later, they did stay at the same post-training level as in Block 3 thus providing further support for the durability of the effects of this feedback training on stereotype application. Again, RTs to the incongruent word pairs were still slower than to the other congruency conditions at the end of the experiment. Therefore, while results are promising, scope for further improvement remains.

It is also worth noting that female participants once again outperformed males in both their accuracy and response time data; a pattern that is consistent with the findings of Experiments 1 and 2. This issue will be briefly elaborated on in Section 2.6.

Overall, these results provide strong support for the use of performance-related feedback as a strategy for reducing the influence of immediate gender activation upon reading gender-biased role nouns in English. This training strategy proved effective at reducing levels of stereotype application, to both previously seen and novel stimuli, and in both the short term and long term. However, while the individual analyses of Experiments 1, 2 and 3 suggest there were mean performance differences across experiments, the following section set out to investigate whether these differences were indeed statistically significant.

2.5 Experiments 1, 2 and 3: Combined analysis.

Introduction

In this section, the data of Experiments 1, 2 and 3 were combined so as to more comprehensively compare performance across all three studies. This was an important step towards ascertaining whether the performance-related feedback training led to (a) significantly reduced levels of stereotype application compared to a control condition (b) whether the post-training performance was still significantly better than control levels when new role nouns had been introduced and (c) whether Experiment 1 achieved better performance than Experiment 3 in which novel role nouns were introduced. Data from Blocks 1 to 3[41] of each experiment was thus combined to investigate these issues.

Results

Analysis

Combining the trimmed data from Experiments 1, 2 and 3, both accuracy of judgements and response times were again analysed using two mixed-design analyses of variance (ANOVAs) on the correct responses, as outlined in Section 2.2.3. However, in the by-participants analyses, Experiment (i.e. Experiment 1/2/3) was further added as a between-subjects factor, while it was added as a within-items factor in the by-items analyses.

Note that the findings reported below do not include effects that were covered in the individual experiment analyses (e.g. main effects of Block, Congruency etc.), but instead focus on effects that emerged with the Experiment variable, and more specifically on the main interest of this combined analysis; performance on critical incongruent trials across experiments.

Accuracy

A main effect of Experiment was first observed, F_1 (2, 111) = 3.25, p = .043; F_2 (2, 97) = 19.66, p < .001, with accuracy of Experiment 1 found to be higher (M = 95.1%) than accuracy in Experiment 3 (M = 93.8%) and in the control experiment (M = 89.5%). Several interactions with the factor of Experiment were also obtained and are outlined below.

[41] Although Experiment 3 had 4 blocks of trials, the 4th block was unique to this experiment and could not be examined across experiments, thus it was not included in this analysis.

Of most interest was a significant three-way interaction of Congruency by Block by Experiment, F_1 (4.25, 235.88) = 3.23, p = .012; F_2 (8, 194) = 4.60, p < .001. A graph of this interaction is shown in Figure 2.9 below; note that Experiment 2 (the control) is displayed last so as to best display the pattern of responding to incongruent word pairs across experiments.

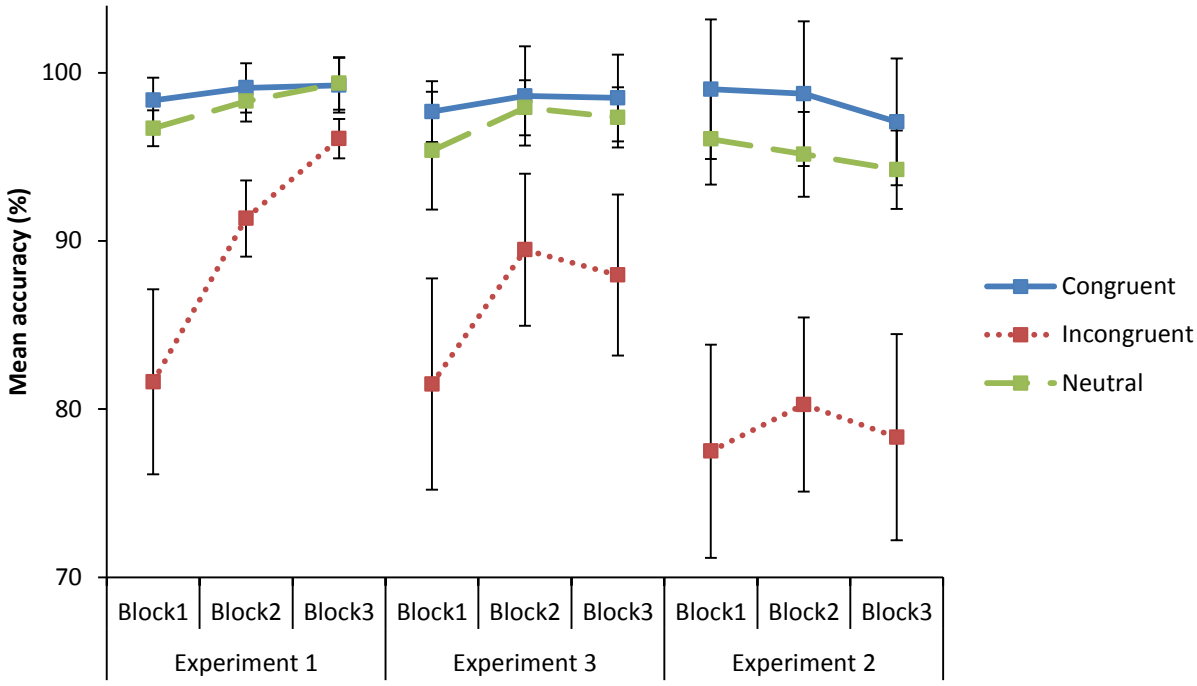


Figure 2.9. Mean % accuracy to critical word pairs across blocks in Experiments 1 to 3. Error bars indicate the 95% confidence intervals.

Figure 2.9 reveals that the Congruency by Block by Experiment interaction was primarily driven by variable performance on stereotype incongruent trials, as ceiling effects were evident in response to stereotype congruent and neutral pairings across experiments. In Experiment 1, accuracy to incongruent pairings was found to rise steadily across blocks, stemming from the provision of feedback in Block 2. In Experiment 3, accuracy of incongruent trials again rose from Block 1 to Block 2 due to the feedback manipulation, yet performance dipped slightly in Block 3 with the introduction of a new set of stereotyped role nouns. However, accuracy to the incongruent word pairs in Experiment 2 was found to be consistently poor across blocks. Also, while variance of performance was revealed to decrease across blocks in Experiment 1, little change occurred in Experiments 2 and 3.

Next, so as to examine performance on these stereotype incongruent pairings in more detail, a second set of ANOVAs was conducted on the incongruent data alone to examine (a)

Experiment (1 vs. 2) by Block (1 vs. 3) performance, (b) Experiment (2 vs. 3) by Block (1 vs. 3) performance and finally (c) Experiment (1 vs. 3) by Block (1 vs. 3) performance.

The comparison of Experiment 1 (performance feedback) and Experiment 2 (the control) revealed a significant Experiment by Block interaction, F_1 (1, 79) = 10.98, $p < .001$; F_2 (1, 23) = 65.03, $p < .001$. Although Block 1 accuracy appears somewhat lower in Experiment 2 than Experiment 1 (4.12%), this difference was not significant in the by-participants analysis, yet was significant in the by-items analysis, t_1 (79) = 0.65, $p = .518$; t_2 (23) = 4.08, $p < .001$, $dz = .83$[42]. However, Block 3 accuracy was highly significant in both sets of analyses, reflecting the substantial improvement across blocks in Experiment 1 relative to the control, t_1 (31.27) = 3.20, $p = .001$, $d = 1.14$; t_2 (23) = 11.10, $p < .001$, $dz = 2.27$.

Next, the comparison of Experiment 2 (the control) and Experiment 3 (long-term feedback) revealed a significant Experiment by Block interaction in the by-items data only, F_1 (1, 64) = 2.84, $p = .097$; F_2 (1, 23) = 5.54, $p = .027$. While Block 1 accuracy was again found to be somewhat lower in Experiment 2 than Experiment 3, this was not significant in the by-participants analysis, and only marginally significant by-items, t_1 (64) = 0.57, $p = .57$; t_2 (23) = 2.03, $p = .055$, $dz = .41$. However, Block 3 accuracy was significantly different across both set of analyses (marginally so by-participants), t_1 (54.47) = 1.44, $p = .078$, $d = .39$; t_2 (23) = 4.59, $p < .001$, $dz = .94$. This pattern of results again illustrates the superior accuracy performance after feedback training (this time to a novel set of role nouns) relative to the control condition in which no training was received.

Finally, comparison of Experiment 1 (performance feedback) and Experiment 3 (long-term feedback) revealed a significant Experiment by Block interaction (marginally so by-participants), F_1 (1, 85) = 3.53, $p = .064$; F_2 (1, 23) = 11.79, $p = .002$. From Figure 2.9 it can be seen that Block 1 accuracy is almost identical across both experiments (mean difference of .14%), however, two-tailed independent-samples *t*-tests revealed that Block 3 accuracy was significantly different across experiments, t_1 (40.35) = 2.00, $p = .052$, $d = .63$; t_2 (23) = 5.00, $p < .001$, $dz = 1.02$. It therefore appears that the feedback training was significantly more successful when the same role nouns were used post-training compared to a novel set.

[42] Note that *t*-tests involving the control experiment and either of the training experiments were two-tailed for Block 1 comparisons (as no difference in performance was hypothesised at this stage), yet one-tailed for Block 3 comparisons (as it was hypothesised that performance would be significantly poorer in the control relative to the others). *T*-tests involving comparisons of Experiments 1 and 3 together were two-tailed as no specific predictions were made about performance before or after training. This same procedure was followed with the RT data.

Next, to further examine performance on critical word pairs across experiments, mean differences between Block 1 and Block 3 scores were calculated and are presented in Figure 2.10 below[43]. Note that positive numbers indicate improved performance in Block 3 compared to Block 1.

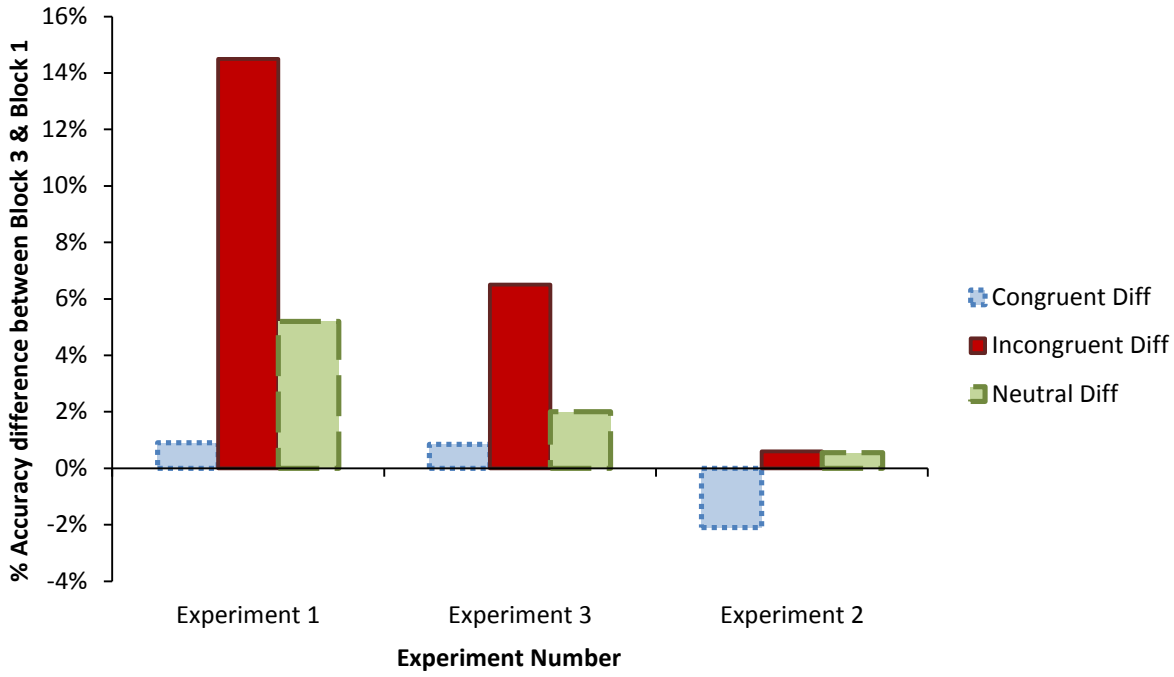


Figure 2.10. Mean % difference in accuracy scores between Block 1 and Block 3, to critical word pairs across Experiments 1-3.

Figure 2.10 reveals that accuracy of these incongruent word pairs increased 14.5% across blocks in Experiment 1, compared to 6.5% in Experiment 3 while only improving 0.6% in the control experiment. A similar, though less extreme, pattern of results was found in response to neutral word pairs, while accuracy of stereotype congruent word pairs showed little change across blocks.

Briefly, subsumed within the three-way interaction of Congruency by Block by Experiment described earlier were a number of two-way interactions involving the same factors. Firstly, an interaction of Block by Experiment was found, F_1 (2.81, 156.22) = 5.24, p = .002; F_2 (2.57, 82.12) = 12.18, p < .001 with an average of 6.87% improvement in accuracy across blocks in Experiment 1, a 3.12% improvement in Experiment 3 but conversely a 0.32% reduction in accuracy in Experiment 2.

[43] While this approach ignores data from Block 2, this was not deemed problematic as this three-way interaction appears driven by pre- vs. post-training accuracy.

Similarly, there was a significant interaction of Congruency by Experiment, F_1 (2.10, 116.59) = 3.52, p = .031; F_2 (4, 64) = 57.42, p < .001, again driven by poorer performance to stereotype incongruent pairings in Experiment 2 (the control) (M = 77.30%), in comparison to Experiment 3 (M = 86.35%) and Experiment 1 (M = 89.65%) respectively.

Response times

A significant main effect of Experiment was *not* found in the by-participants analysis, F_1 (2, 111) = .77, p = .465, yet was revealed in the by-items analysis, F_2 (1.82, 58.2) = 14.64, p <.001. Mean RTs in the by-items data showed that fastest responses emerged in Experiment 1 (M = 816ms), followed by Experiment 2 (M = 868ms) and then Experiment 3 (M = 875ms) i.e. a predictable pattern of results based on the training (or lack thereof) provided. This RT difference between Experiments 1 and 2 was significant (t_2 (68) = 3.54, p = .001, d =.86) as was the difference between Experiments 1 and 3 (t_2 (68) = 4.65, p < .001, d = 1.13). However, in the by-participants analysis, no significant differences between any of the experiments were found (ps > .3), despite the same pattern of results emerging as in the by-items data i.e. RTs were again fastest in Experiment 1 (M = 836ms), followed by Experiment 2 (M = 880ms) and 3 (M = 894ms) respectively. As explained in Section 2.2.3 it is highly likely that this effect was only significant by-items because the standard errors of the condition means are likely to be lower in the by-items analysis than in the by-participants analysis, if the variances are roughly equal (due to the much greater number of items than participants in the respective analyses).

Contrary to expectations, no evidence of an Experiment by Congruency by Block interaction was found in the by-participants, F_1 (8, 444) = .86, p = .55, or by-items analysis, F_2 (8, 128) = .566, p = .804. However, to examine the F_1 pattern of responding across blocks, RTs to critical pairings in each of the 3 experiments can be seen in Figure 2.11 below. Note that Experiment 2 (the control) is again displayed last so as to best display the pattern of responding to incongruent word pairs across experiments.

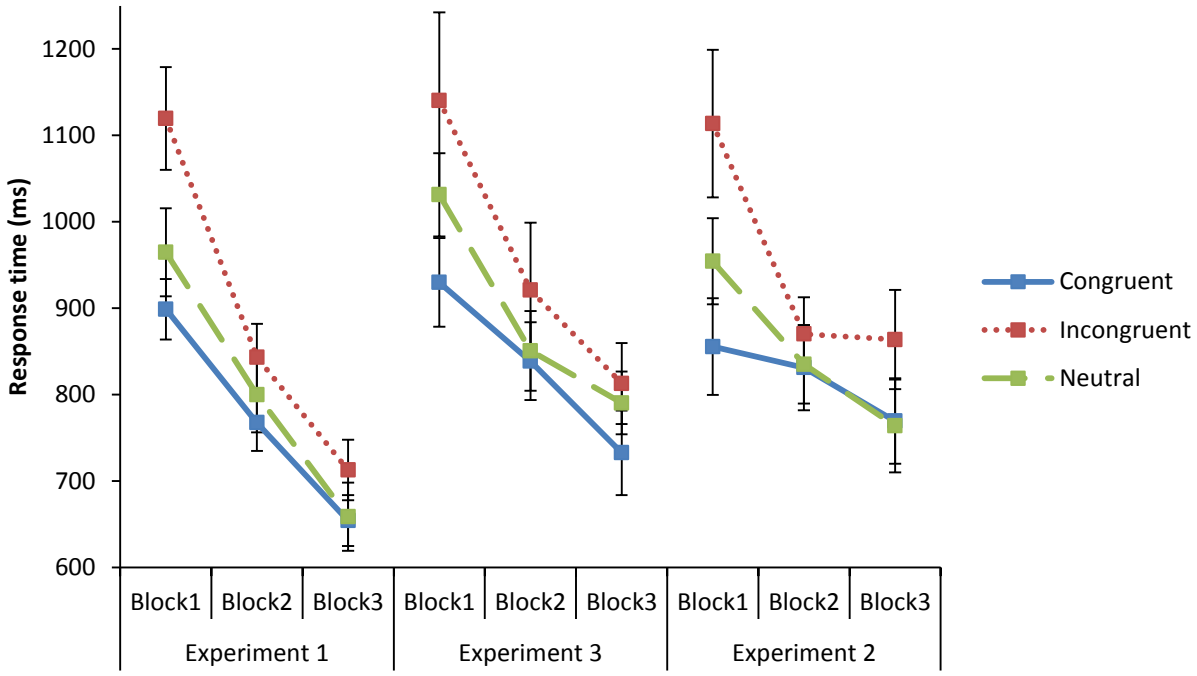


Figure 2.11. Mean response times (in milliseconds) to critical word pairs across blocks in Experiments 1 to 3. Error bars indicate the 95% confidence intervals.

Figure 2.11 reveals a relatively similar pattern of responding across experiments. However, so as to more closely examine the stereotype incongruent data, a second set of ANOVAs was conducted on the incongruent data, investigating (a) Experiment (1 vs. 2) by Block (1 vs. 3) performance, (b) Experiment (2 vs. 3) by Block (1 vs. 3) performance and finally (c) Experiment (1 vs. 3) by Block (1 vs. 3) performance.

To begin, the comparison of Experiment 1 (performance feedback) and Experiment 2 (the control) revealed a significant Experiment by Block interaction, F_1 (1, 79) = 5.77, p = .019; F_2 (1, 23) = 10.82 p = .003. While no significant difference in Block 1 RTs emerged, t_1 (79) = 0.30, p = .768; t_2 (23) = 0.98, p = .337, this difference was significant in Block 3, t_1 (79) = 2.22, p = .015, d = .50; t_2 (23) = 3.72, p < .001, dz = .76. Therefore, despite RTs improving across blocks in the control condition due to practice effects (as described in Section 2.3.3), the above pattern of results illustrates that RTs improved to a greater extent following the feedback training relative to the control.

Next, the comparison of Experiment 2 (the control) and Experiment 3 (long-term feedback) failed to reveal a significant Experiment by Block interaction, F_1 (1, 64) = .72, p = .398; F_2 (1, 23) = 1.70, p = .206. Similarly, t-tests revealed that there was no difference between starting RTs in Block 1 across both experiments, t_1 (64) = .25, p = .801; t_2 (23) = .10, p =.328, or final RTs in

Block 3, t_1 (64) = .64, p = .522; t_2 (23) = 1.06, p =.300. This pattern of results suggests that when novel items are introduced after the feedback training, final RTs are not significantly faster than in the control condition.

Finally, comparison of Experiment 1 (performance feedback) and Experiment 3 (long-term feedback) again revealed that there was no significant Experiment by Block interaction, F_1 (1, 85) = 1.28, p = .260; F_2 (1, 23) = 1.67, p = .210. *T*-tests again found there was no difference between starting RTs in Block 1 (p > .7) across experiments, however, final RTs in Block 3 were significant in the by-items data only, t_1 (60.18) = 1.46, p = .149; t_2 (23) = 2.50, p =.020, dz = .51. This last finding provides some evidence the feedback training was more successful when the same role nouns were used post-training compared to a novel set, as was found with the accuracy data.

Discussion: Combined analysis

A combined analyses of Experiments 1-3 was conducted to establish whether or not performance-related feedback is a valuable strategy for reducing stereotype application, and whether the effects of this training can successfully extend to a new set of stimuli.

Firstly, with accuracy of responses, the results were largely as predicted. Accuracy of stereotype incongruent word pairs was significantly higher in Experiments 1 and 3 (where feedback was provided) relative to Experiment 2 (the control). However, despite earlier findings of Experiment 3 revealing that the feedback training resulted in significantly higher accuracy to novel stimuli (relative to pre-feedback levels), this combined analysis showed that accuracy remained significantly higher in Experiment 1 than Experiment 3 (i.e. higher when the same set of role nouns was presented after the feedback training as opposed to a new set). It can be concluded that performance-related feedback is an effective means of reducing stereotype application when a participant is presented with specific items on which feedback was received. However, this training demonstrates limited transfer when novel items are introduced.

As regards the response time data, the results of Experiment 2 had suggested that RTs to incongruent pairings decreased significantly across blocks independently of the feedback training. However, this combined analysis revealed that RTs to stereotype incongruent trials in Block 3 were significantly faster in Experiment 1 than the control, thereby again endorsing the use of performance-related feedback as a stereotype reduction strategy. On the other hand,

when novel stimuli were introduced in Block 3 of Experiment 3, RTs to incongruent pairings did not prove significantly faster than those in the control experiment. Similarly, final RTs in Experiment 1 were faster than those of Experiment 3. Together, these findings are largely in line with those of the accuracy data and suggest that performance-related feedback is very effective as a specific, stereotype-reduction strategy, with its effects generalising to novel stimuli with moderate success.

2.6 Chapter Discussion

The aim of this chapter was to investigate the efficacy of performance-related feedback as a strategy for reducing gender stereotype application in a judgement task. Across three experiments, participants showed more difficulty (i.e. lower accuracy and higher response times) when responding to stereotype incongruent word pairs as opposed to stereotype congruent or neutrally rated pairings. These findings support those of past researchers who have suggested that gender stereotyped information is automatically elicited from single words (Banaji & Hardin, 1996; Oakhill et al., 2005) and that gender resolution can be difficult to achieve in language processing when gender-related expectancies clash with explicitly stated gender information (e.g. Carreiras et al., 1996; Irmen, 2007; Kreiner et al., 2008).

Experiment 1 was designed as a preliminary investigation into whether this stereotyping effect could be attenuated with the provision of performance-related feedback. Results showed that both accuracy and response times improved significantly following the provision of feedback. However, performance to stereotype incongruent pairings remained worse than to stereotype congruent and neutral pairings by the end of the experiment. Thus, this experiment provided support for the use of performance-related feedback as a stereotype reduction strategy, but also revealed further scope for improvement.

Experiment 2 was conducted as a control experiment against which the performance of Experiment 1 could be compared. This was deemed necessary so as to rule out the possibility that improved performance to stereotype incongruent word pairs in Experiment 1 was due to practice effects. Results of the control experiment revealed that accuracy of the word pairs did not significantly increase across blocks when feedback was not provided, however response times were found to naturally decrease as the experiment progressed. Again, performance to incongruent pairings was found to be significantly poorer than to congruent and neutral pairings at the end of the experiment. These results thus provide partial support in favour of

the feedback training, in particular its value in raising awareness of gender biases (and consequently leading to an increase in accuracy of responses in Experiment 1) as opposed to improving response times to the word pairs.

Experiment 3 investigated the more applied value of performance feedback as a means of stereotype reduction. Both the generalisability of this training to novel stimuli, and the durability of training effects were explored. Beginning with generalisability, it was found that post-feedback accuracy and RTs to incongruent pairings were significantly better than pre-feedback levels, thus providing support for the generalisability of the training. Similarly, performance one week after the initial training session was significantly higher than pre-feedback levels, thereby providing evidence for the durability of the feedback training effects. In this experiment, performance to stereotype incongruent pairings was once more poorer than to congruent and neutral pairings by the end of the experiment, again demonstrating scope for further improvement.

Finally, when all three experiments were combined for analysis, some interesting findings were observed. With the accuracy data, results were in line with predictions. Judgements of stereotype incongruent pairings were significantly more accurate in Experiment 1 than in Experiment 3 (where novel stimuli were introduced) while this experiment in turn achieved higher accuracy than Experiment 2 (the control). Next, with the RT data, it was found that post-feedback response times to incongruent pairings were significantly faster in Experiment 1 than Experiment 2 (the control). However no significant difference was found between Experiment 3 and the control, suggesting that the introduction of new role nouns in Experiment 3 led to reduced levels of RT improvement from Block 2 to Block 3, as compared with Experiment 1. Finally, there was no significant difference in Block 3 RT performance of Experiment 1 and Experiment 3 in the by-participants analysis, yet this was significant in the by-items analysis. Overall, means revealed that participants were faster in responding to word pairs that they had received feedback on (Experiment 1) as opposed to novel items (Experiment 3).

An unexpected series of findings in relation to Participant Gender also emerged across Experiments 1-3. Female participants were consistently found to outperform males on both the accuracy and response time data. This finding was not anticipated (based on the results of Oakhill et al., 2005 where no such effect was reported) and it is not entirely clear why this pattern of results emerged. However, as it is *typically* females as opposed to males that are the objects of sexism, one potential explanation for this finding may simply be that females have

more experience of recognising sexism than males, and thus may ultimately hold more flexible representations of sex roles. Also, evidence from the early nineties suggests that females have been entering male-dominated occupations to a greater extent than males have entered female-dominated occupations (Reskin & Roos, 1990). Overall, females may be more sensitive to these changes in the gender distribution of certain social roles and thus generally be more open to accepting stereotype incongruent pairings than their male counterparts. However, this theory cannot be verified here.

The processes underlying stereotype reduction

A pertinent issue involves the mechanism(s) through which unwanted, yet automatically activated, biases are overcome. Wegner (1994) designed a model of thought suppression (relevant to the bypassing of stereotype bias in this chapter) which suggests that two cognitive processes are jointly involved in this goal. Firstly, monitoring processes are hypothesised to scan the mental environment in a search for evidence of the unwanted thoughts (e.g. stereotype biases). If found, then another operating process will work to direct attention away from the unwanted thought and onto a distracter (e.g. counter-stereotype information, definitional gender information). This monitoring stage is thought to occur in quite an automatic manner while the operating process is proposed to require effort and sufficient cognitive resources to function successfully. Wegner (1994) consequently asserts that detecting stereotyped information may be achieved with relative ease but that replacing such information is more demanding and necessitates adequate attentional resources. However, with adequate cognitive resources and high motivation, the operating processes may succeed in maintaining attention on distracters as opposed to the stereotypic material (Wegner, 1994; Wenzlaff & Wegner, 2000). This appears to have been the case with the studies outlined in this chapter as, with practice, participants became more adept at overcoming stereotypic information and taking definitional gender information into account in their judgements.

In a similar vein, the results of this chapter provide support for the claims of past researchers who posit that some combination of awareness, motivation, skill and resources is required to overcome immediate stereotype biases (e.g. Bargh, 1992, 1999; Hilton & Von Hippel, 1996; Macrae & Bodenhausen, 2000). More specifically, the performance-related feedback training sought to first make participants aware of their biases and then induce a motivated attention to the stereotype incongruent information so as to overcome subsequent biases. Furthermore, the lack of time constraint in responding facilitated performance as cognitive resources were not overly stretched.

Parallels can also be drawn between the design of the studies in this chapter and the work of Kawakami et al. (2000). As described in Chapter 1, Kawakami et al. (2000) used a stereotype negation/counter-stereotype affirmation training task across 480 trials to successfully reduce levels of stereotype activation. While at first glance this strategy differs somewhat from the feedback training used in the current chapter, it could be argued that the judgement task used thus far encourages a type of mental negation of the gender stereotype and conscious affirmation of the counter-stereotype in order for participants to successfully respond. It is therefore likely that the underlying processes involved in reducing levels of stereotyping across the two studies are quite similar. Indeed, Kawakami and colleagues proposed three mechanisms through which they believe their training may have operated that are now relevant to the current studies.

They firstly posit that cognitive changes may have resulted from the differential reinforcement and weakening of certain category-trait associations. The learning of new associations may have led to stereotype dilution, and in turn, reduced stereotype activation. This proposal seems equally plausible in terms of the performance-feedback training, with participants creating stronger associations between previously weak category-member associations, and vice versa.

Secondly, Kawakami et al. (2000) posit that motivational factors may have played a role in the success of their training. Through repeated activation of the goal 'to not stereotype', participants may have learned to spontaneously apply a self-regulatory process (Bargh, 1990; Bargh & Gollwitzer, 1994; Moskowitz, Gollwitzer, Wasel, & Schaal, 1999). As described in Chapter 1, this is a theory closely linked to the auto-motive model of Bargh and colleagues (Bargh, 1990; Bargh & Gollwitzer, 1994; Chartrand & Bargh, 1996). In this model, goals and motives must be represented in the mind in a way akin to that of other knowledge structures, and thus be capable of becoming automatically associated with representations that they are repeatedly paired with. Therefore, given the large number of trials in the three experiments outlined in this chapter (Experiments 1 and 2 had 456 trials while Experiment 3 had 608), it is highly likely that regulatory processes and the goal of stereotype-free responding became automated to a certain extent, thereby resulting in reduced levels of stereotype application.

Thirdly, it may be that a combination of the two processes described above led to successful stereotype reduction (Kawakami et al., 2000).

Returning briefly to the number of trials used in Experiments 1-3, the role of repetition merits further consideration. Again, the findings provide support for the assertion of Kawakami and

colleagues (2000) who argue that, in addition to the immediate efforts of the participants to overcome a stereotype, the successful reduction of bias was also owing to substantial, extended practice. Despite responding being initially demanding, performance became increasingly efficient across trials. Indeed, literature from the field of skill acquisition posits that extensive practice can lead to automatic responses (Shiffrin & Schneider, 1977; Smith, Branscombe, & Bormann, 1988; Wyer & Hamilton, 1998). It is posited that, with repetition, a newer response to a particular stimulus can come to dominate an old (automatic) response (Kawakami et al., 2000), or in this case, that new non-stereotypic responding can come to dominate the previous gender stereotypic responding.

Further considerations

Despite the documented success of performance-related feedback as a stereotype reduction training, there is an important design issue which may have influenced the results thus far. As part of the performance-feedback information, participants were provided with their cumulative percentage scores. In hindsight, the combined inclusion of cumulative percentage scores and explicit accuracy information ('Correct' vs. 'Incorrect'), has added complication to the interpretation of results. It cannot be ascertained which of these factors contributed most to the reduction in levels of stereotype application. While it is likely that the latter, more direct, approach had the most impact, the independent contributions of each component remain impossible to disentangle with the current design. Future research could tease apart these effects with a between-subjects design, providing participants with only one aspect of this performance feedback.

Another avenue for future research points to whether this feedback training could also aid performance on stereotype *activation* as opposed to stereotype *application*. Participants in the current studies were first presented with a prime for 1,000ms, with no time limit imposed within which to respond to the target. However, based on a study by Neely (1977), the field has largely adopted a 500ms cut-off as a boundary up to which automatic priming effects can be assessed (Blair & Banaji, 1996). Therefore, by adapting the design of the experiments used in this chapter so as to adhere to tighter time restrictions on role noun presentation and response time limits, the effects of the feedback training on stereotype activation could be assessed.

Overall, the studies outlined in this chapter provide support for (a) the use of performance feedback as a stereotype reduction strategy, (b) the value of extensive practice in overcoming stereotyping and (c) the importance of investigating beyond the immediate effects of a

training strategy to explore the generalisabilty and durability of reported findings. More specifically, the use of performance-related feedback as a stereotype reduction strategy has proved beneficial, both in the short and long-term. However, given that the stereotype bias associated with incongruent word pairs was not completely eradicated in Experiments 1-3 (relative to performance on stereotype congruent and neutral pairings), the following chapter investigates the use of another form of feedback, documented in the stereotype literature, as a means of stereotype reduction. This feedback is based on social norm information.

3. Social-consensus feedback as a strategy to overcome automatic gender stereotypes

3. 1 Introduction

A sizeable body of research has now been devoted to the influence of other people's beliefs on an individual's own beliefs, with evidence emerging that perceivers frequently modify their intergroup attitudes and behaviours in order to align with those modelled by members of groups that they value. For example, informing participants that stereotyping is not typical of their in-group has previously been found to reduce stereotyping against groups such as racial minorities (e.g. Stangor et al., 2001; Wittenbrink & Henly, 1996) and people suffering from obesity (Puhl, Schwartz, & Brownell, 2005), while, conversely, research has documented that discrimination against racial minorities and women is more tolerated when a racist or sexist joke has been heard (Ford & Ferguson, 2004; LaFrance & Woodzicka, 1998). Across three experiments, this chapter explores the use of (fictitious) social consensus feedback as a gender stereotype reduction strategy. This feedback involves presenting participants with social norm information in an attempt to highlight the attitudes or behaviours of a social group (their peers) towards the topic of role-based gender stereotypes.

Peer influence on intergroup prejudice is thought to be driven by the basic human goals of understanding, social connection, and self-definition (Paluck, 2011). Indeed, Prentice and Miller (1993) posit that individuals will experience discomfort if they perceive their attitudes to be different from the normative attitude of their peer group. However, this discrepancy can be resolved in three ways (1) by moving an individual's personal attitudes towards that of the perceived norm, (2) bringing the norm closer to the individual's attitude, or (3) complete rejection of the group. Prentice and Miller maintain that the most straightforward way for an individual to reduce a perceived discrepancy in attitude, is to bring their private attitudes in line with those of the group norm. It is via this route that behaviour change is hypothesised to occur in this chapter. Furthermore, Stangor et al. (2001) posit three reasons why individuals should be particularly likely to be swayed by the opinions of others on the issue of stereotyping (1) the accuracy of stereotypes is difficult to assess objectively (2) stereotyping is a socially sensitive topic and (3) people are likely to be highly motivated to learn about the traits of individuals from different social groups.

While many studies have claimed that people reform their attitudes and beliefs so as to mirror those of their peer group, less research has explored how well people can initially identify

these social norms. In fact, it is now clear that people can make significant errors in their estimation of opinions held by others (e.g. Prentice & Miller, 1993). Such errors frequently occur through the phenomenon of pluralistic ignorance; the belief that one's private attitudes and judgements differ from those of others, despite both parties displaying identical public behaviour (Miller & McFarland, 1991). The use of fictitious norm information as a stereotype reduction strategy in this chapter benefits from the fact that people are often poor at estimating social norms, yet are strongly influenced by what they perceive these norms to be. For example, as will be revealed, the social feedback presented in Experiments 4 and 6 varies greatly from the feedback presented in Experiment 5, yet participants show some evidence of compliance with the presented norms across all three experiments (i.e. participants are not particularly adept at gauging responses of their peers and so readily accept the norms they are presented with as being true).

Finally, the exact form that the social consensus feedback in this chapter should take was influenced by research stating that the source of norm information is an important variable when aiming to exert an influence over a person's attitudes or beliefs. Research indicates that information coming from a valued in-group will be most highly effective in this regard (Levine, Resnick, & Higgins, 1993; Sechrist & Stangor, 2001; Stangor et al., 2001). It was therefore decided to base the consensus feedback on the participant's peer group, which consisted of fellow students who had previously completed the judgement task. It was hypothesised that participants would deem it both important and desirable to respond in a similar manner to their peer-group.

With the above information in mind, the first experiment in this chapter (Experiment 4) was designed to investigate the effect of fictitious social consensus feedback on levels of gender stereotype endorsement on the judgement task.

3.2 Experiment 4: Social consensus feedback

3.2.1 Introduction

In Experiment 4, participants again completed three blocks of stereotype judgement trials, with feedback provided in Block 2 only. However, unlike the studies of Chapter 2, feedback was now based solely on social norm information. This feedback ostensibly indicated the percentage of students in a previous study that agreed with the participant's judgement. However, although presented to participants as true and accurate feedback, in reality, it was

fictitious and manipulated so as to suggest that gender stereotype endorsement was very infrequent among the participant's peer group. In this way it was hypothesised that participants would modify their responses towards the perceived attitudes of their peer group and display lower levels of stereotype application in Block 2. It was further hypothesised that this improved performance would be maintained in Block 3 (despite the removal of the feedback), with participants investing continued effort to adapt their responding to the social norms they were presented with in Block 2. A battery of five questionnaires was also completed by participants in this study; these will be discussed further in Chapter 5.

3.2.2 Method

Participants

Thirty-six students (17 male, 19 female) from the University of Sussex took part in this experiment. Participants' ages ranged from 18 to 30 years (*M*: 19.61; *SD*: 2.81). They received either £6 or 4 course credits for taking part in the session which lasted approximately 45 minutes.

Materials & Design

The design of the experiment and materials used were identical to those of Experiment 1 (Section 2.2.2) but with two notable exceptions (1) the performance-related feedback was replaced with feedback based on fictitious social norm information, and (2) one further questionnaire was added; the Brief Fear of Negative Evaluation scale (BFoNE; Leary, 1983).

Social consensus feedback

The social consensus feedback consisted of a single sentence stating the percentage of students in a previous study run by the experimenter who agreed with the participant's judgement e.g. '*_% of previous students agreed with you'.* As mentioned above, this feedback was in fact fictitious and constructed so as to suggest that the vast majority of previous participants accepted stereotype incongruent word pairs as warranting *yes* responses (i.e. as being perfectly acceptable). In this way, the social feedback sought to endorse gender-fair responding and highlight any discrepancy between a participant's response and the peer group norm. For example, if a participant responded that both terms of a stereotype incongruent word pair (e.g. *carpenter/sister*) could not refer to one person, feedback indicated that a number between 2% and 5% of previous students agreed with this judgement. Conversely, if a

participant judged such a pairing as acceptable, feedback indicated that a number between 95% and 98% of previous students agreed with the participant's choice, thereby reinforcing their non-stereotypic responding. Note that the extreme, narrow range of feedback used was chosen so as to strongly and consistently convey that previous participants did not respond in a stereotyped manner.

A specific range of social consensus feedback was created for each of the three congruency conditions (see Appendix 7), with exact figures within this specified range counterbalanced across pairings (e.g. with the stereotype incongruent trials, the figure 95% was presented an equal number of times as 96%, 97% and 98% in response to correct judgements). Aside from the stereotype incongruent trials, feedback to word pairs in all other congruency conditions (stereotype congruent, neutral, definitionally matching and definitionally mismatching word pairs) was loosely based on real data, and intended to be typical of past response accuracy in Experiments 1-3 i.e. feedback strongly endorsed correct responses and rejected incorrect responses[44].

As in Experiment 1, the feedback was presented on-screen for 1000ms before ceding to the next trial (ITI = 500ms). However, instructions were adapted to inform participants that feedback indicated the percentage of participants in a previous experiment that agreed or disagreed with their responses.

Procedure

The experimental procedure largely matched that of Experiment 1 (Section 2.2.2), but entailed a more comprehensive debriefing session in which it was revealed to participants that the feedback information was entirely fictitious. Participants were then reassured that, in reality, stereotype biases occur much more frequently than the feedback suggested and that there was no evidence that they were stereotyping to a greater extent than their peers. Finally, unlike Experiment 1, the BFoNE was administered after the behavioural task (while the other questionnaires were administered before the behavioural task) as it was thought that answering questions about fear of negative evaluation could alert participants to the nature of the social feedback. This questionnaire data will be discussed further in Chapter 5.

[44]However, despite the fact that accuracy to male, definitionally mismatching word pairs was found to be relatively low in the previous experiments (likely due to the generic interpretation of certain male-specific terms e.g. host, landlord), feedback continued to strongly suggest that such terms should be interpreted according to their definitional gender i.e. as being male-specific. For instance, if a mismatch pairing such as 'host/mother' was judged as acceptable, feedback stated that only 0-2% of people agreed with this response.

3.2.3 Results

Data screening

As in previous studies, data for the neutral term *adolescent* were removed before analysis, resulting in a loss of 1.32% of the data.

Analysis

Response times below 150ms, and above 4,000ms were excluded from the analyses (representing 2.88% of the total data) along with times for all errors of judgement (representing a further 10.05%), totalling a loss of 12.93% of the data. Statistical analyses were conducted as described in Section 2.2.3.

Accuracy

A main effect of Stereotype bias was found, F_1 (1.27, 43.07) = 16.19, $p < .006$; F_2 (2, 32) = 11.45 $p < .001$, with higher accuracy to word pairs that contained a neutral role term (M = 96.3%), than those that contained male (M = 89.9%) or female stereotype-biased terms (M = 89.0%).

As expected, evidence of stereotyping was also revealed with a main effect of Congruency[45], F_1 (1.04, 35.24) = 16.30, $p < .001$; F_2 (2, 32) = 106.43, $p < .001$, driven by significantly lower accuracy to stereotype incongruent word pairs (M = 79.9%), than to stereotype congruent (M = 99.0%) and neutral pairs (M = 96.4%).

A significant effect of Block was found in the by-items analysis only, F_2 (1.73, 55.38) = 15.68, $p < .001$, with a 2.9% increase in accuracy across blocks. However this 2.9% increase did not lead to a significant effect of Block in the by-participants analysis, F_1 (1.19, 40.46) = 2.87, p = .092. As explained in Chapter 2, this variable pattern of results is due to the smaller number of participants than items within the respective analyses (with the standard errors of the condition means likely to be lower in the latter case, if the variances are roughly equal).

Importantly, a significant interaction of Block by Congruency emerged, F_1 (2.23, 75.96) = 3.08, p = .046; F_2 (4, 64) = 5.89, $p < .001$. As can be seen in Figure 3.1 below, this interaction was driven by a steady increase in accuracy of stereotype incongruent pairings across blocks, totalling a 6.14% increase from Block 1 to Block 3. Due to ceiling effects, much smaller improvements in accuracy of the neutral and congruent conditions were found across blocks (1.9% and 0.6% respectively), both of which had very high accuracy from the outset.

[45] i.e. an interaction of Stereotype bias by Kinship term gender.

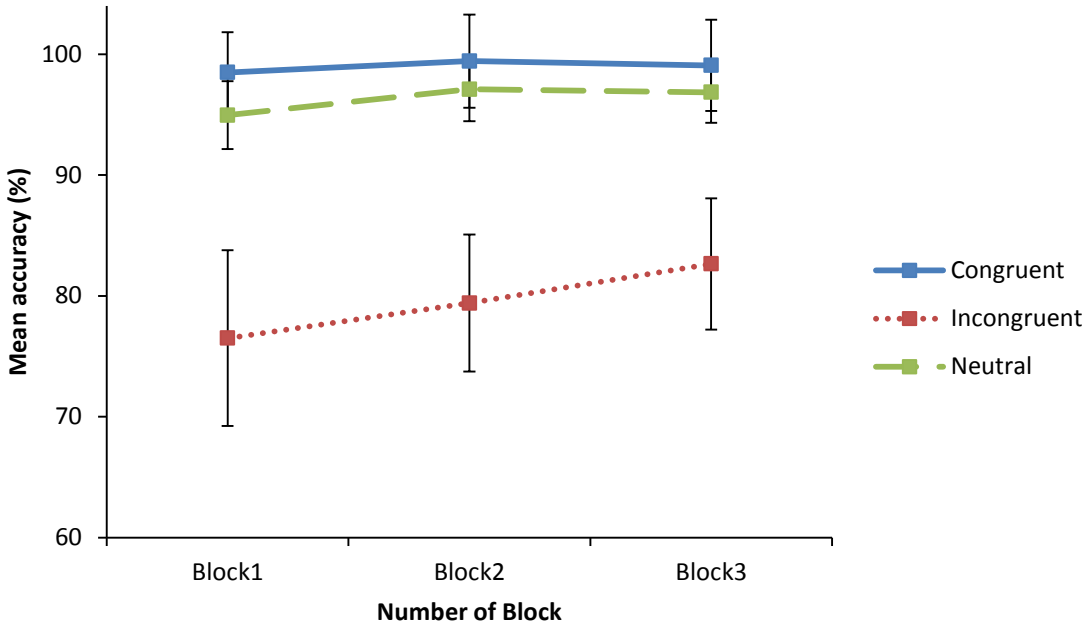


Figure 3.1. Experiment 4: Mean percentages of correct judgements to critical word pairs across blocks. The vertical axis begins at 60% while error bars indicate the 95% confidence intervals.

Paired samples *t*-tests[46] revealed that the aforementioned increase in accuracy to stereotype incongruent pairings across Blocks 1-3 was indeed a significant one, t_1 (35) = 2.09, p = .022, dz = .35; t_2 (23) = 4.26, p < .001, dz = .87. It is posited that this accuracy improvement is due to the social feedback manipulation in Block 2 of the judgement task, thus providing support for the use of this feedback as a useful stereotype reduction strategy. That said, a relatively small magnitude of effect was revealed, particularly in the by-participants analysis. By the end of the experiment, accuracy to stereotype incongruent word pairs remained significantly lower than to that of stereotype congruent pairings, t_1 (35) = 3.69, p = .001, dz = .62; t_2 (23) = 9.80, p < .001, dz = 2.0, and neutral word pairs, t_1 (35) = 3.83, p = .001, dz = .64; t_2 (30.91) = 7.59, p < .001, d = 2.73. It can therefore be concluded that, despite an increase in accuracy to stereotype incongruent word pairs, the social consensus feedback did not completely succeed in eliminating gender-biased responding.

Finally, a number of effects with Participant Gender emerged in the by-items analysis only. There was a main effect of Participant Gender, F_2 (1, 32) = 52.99, p < .001, with male participants showing higher accuracy than female participants overall (M = 93.8% vs. M =

[46] A one-tailed *t*-test was used for this comparison (as it was anticipated that performance on the incongruent pairings would improve after the feedback training) while all remaining differences were examined using two-tailed tests. This procedure was also followed for the RT data.

89.6% respectively). Next, a Participant Gender by Congruency interaction was revealed, F_2 (2, 32) = 39.73, p < .001, with both male and female participants displaying similar performance on stereotype congruent and neutral ratings, but with male participants showing much higher accuracy on stereotype incongruent pairings than female participants (M = 85.3% vs. M = 74.4% respectively). Finally, a Participant Gender by Stereotype bias interaction indicated that male participants consistently outperformed females, but to a lesser degree in response to neutral terms than the stereotype biased terms, F_2 (2, 32) = 4.92, p = .014. These results are in stark contrast to those of Chapter 2 in which female participants were repeatedly found to outperform males.

Response times

A main effect of Stereotype bias was found in the by-participants analysis only, F_1 (2, 62) = 6.36, p = .002; F_2 (2, 32) = 1.01, p = .375, with faster response times to word pairs that contained a neutral role term (M = 838ms), than those that contained male-biased (M = 890ms) or female-biased terms (M = 898ms).

There was also a significant effect of Kinship term gender (marginal in the F_1 analysis), F_1 (1, 34) = 4.03, p =.053; F_2 (1, 32) = 4.63, p = .039, with faster response times to word pairs that contained female kinship terms (M = 858ms) than male kinship terms (M = 893ms) overall. An interaction of Kinship term gender by Participant Gender was found in the by-participants analysis only, F_1 (1, 34) = 4.14, p = .050, with female participants responding faster to female kinship terms than male kinship terms (857ms vs. 946ms respectively, mean difference = 89ms), while male participants responded at an equivalently fast speed to both male and female kinship terms (840ms vs. 840ms respectively). A main effect of Participant Gender was found in the by-items analysis only, F_2 (1, 32) = 20.89, p < .001, with male participants faster on average at responding than female participants (M = 830ms vs. M = 875ms respectively).

A main effect of Block also emerged, F_1 (2, 68) = 24.22, p < .001; F_2 (2, 64) = 70.64, p < .001, with average response times very similar in Block 1 and Block 2 (937ms and 938ms respectively), but then decreased sharply in Block 3 (750ms). Contrasts also revealed a linear trend, F_1 (1, 34) = 30.11, p <.001; F_2 (1, 32) = 99.27, p < .001.

Next, a main effect of Congruency[47] was found, F_1 (1.67, 56.82) = 18.31, p < .001; F_2 (2, 32) = 12.47, p < .001, with fastest RTs recorded in response to stereotype congruent (M = 824ms)

[47] i.e. an interaction of Stereotype bias by Kinship term gender.

and neutral word pairs (*M* = 838ms), while RTs to incongruent pairings were considerably slower (*M* = 967ms).

Contrary to expectations, no interaction of Congruency by Block was found in the by-participants (F_1 (2.82, 95.70) = 1.80, p = .709) or by-items analyses (F_2 (3.78, 60.47) = 1.48, p = .222) respectively. Indeed, from the by-participants data displayed in Figure 3.2 below, it can be seen that RTs in each of the three congruency conditions produced a relatively similar pattern of results. The sharpest fall in RTs was found with the stereotype incongruent pairings, decreasing 263ms across Blocks 1-3 (versus 128ms for congruent pairings and 168ms for neutral pairings). It appears that the social feedback initially had little effect on speed of responding to incongruent pairings, but greatly aided subsequent performance in Block 3 (although, practice effects are also likely to have contributed to this decrease in RTs). This delayed impact of the feedback on RTs may be a result of the long sentence format of the consensus feedback (in contrast to the short performance-feedback presented in Chapter 2), as increased processing time may have been required before any effects of the training were evident. However, paired-tests reveal that this decrease in speed of response from Block 1 to Block 3 was significant, t_1 (35) = 3.95, p < .001, dz = .66; t_2 (23) = 9.22, p < .001, dz = 1.88, with a much greater magnitude of effect observed in the by-items analysis than the by-participants analysis.

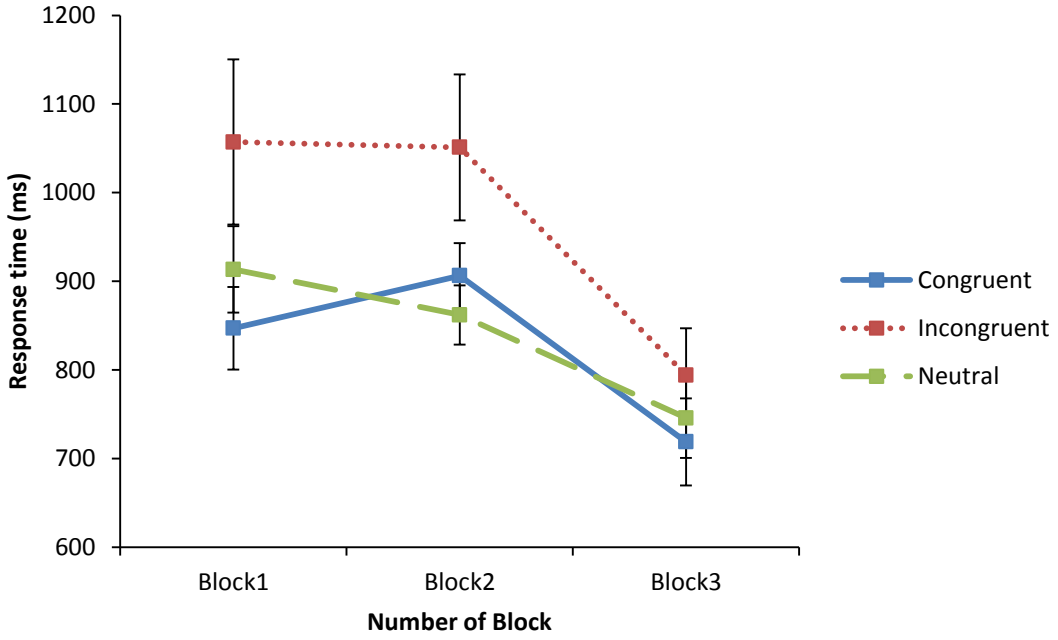


Figure 3.2. Experiment 4: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks. Error bars indicate the 95% confidence intervals

Figure 3.2 also reveals that RTs to stereotype congruent pairings rose slightly in Block 2 before then falling quite sharply in Block 3. Again, this pattern of responding may have resulted from longer processing requirements of the feedback provided. Finally, RTs to neutral pairings showed a more consistent, gradual decline across blocks. This pattern is unsurprising given that participants should have less difficulty responding to judgements with neutral role terms than those with a stereotype bias.

As RTs improved from Block 1 to Block 3 in all congruency conditions, this RT data provides provisional evidence for the use of social consensus feedback as a strategy for reducing the effects of gender stereotype activation. However, a marginal difference between the RTs of stereotype incongruent and neutral pairings remained in Block 3 (t_1 (35) = 1.75, p = .088, dz = .29; t_2 (44) = 4.07, p < .001, d = 1.23), along with a significant difference between RTs of stereotype incongruent and congruent pairings, t_1 (35) = 2.07, p = .046, dz = .35; t_2 (23) = 3.0, p = .006, dz = .61. Although their accompanying effect sizes are small, these significant differences reveal that the social feedback training did not fully eradicate the stereotyping effect as, ultimately, RTs to stereotype incongruent pairings were slower than in the other congruency conditions. Similarly, given the results of Experiment 2 in which participants were found to naturally get faster as the judgement task progressed, it is likely that any improvement in RTs across blocks in this experiment is, in some part, also due to practice effects.

Fillers - Accuracy

As was consistently found in Chapter 2, accuracy to the definitionally matching word pairs was higher than to the definitionally mismatching word pairs across this experiment (96% vs. 86% respectively).

Again, with the definitionally mismatching pairs, poorer accuracy to word pairs containing a male-specific role noun (M = 77%) as opposed to a female-specific role noun (M = 95%) was found. Once again it is hypothesised that this was due to the generic interpretation of several masculine terms such as *host* or *steward* that are in fact male-specific by definition.

Fillers - Response times

RTs to definitionally matching word pairs were faster than to definitionally mismatching pairs (1048ms vs. 1094ms respectively). Faster RTs to female word pairs over male word pairs in both the definitionally matching (999ms vs. 1048ms respectively) and mismatching cases (1044ms vs. 1145ms) were also found. This trend supports the accuracy data, with longer

processing of male pairings likely to reflect participants' reflection over terms that are masculine by definition but frequently used generically in reference to either or both sexes.

3.2.4 Discussion

Experiment 4 sought to investigate the influence of social consensus feedback on levels of gender stereotype application. Based on past research, it was hypothesised that a discrepancy between participants' responses and that of the perceived attitude of their peers would induce a feeling of discomfort, thus motivating participants to adapt their responding so as to mirror the perceived attitudes of their peer group (i.e. reduce stereotypic responding). While this approach of using social norm information as a strategy to reduce stereotype bias proved successful in the past (Puhl et al., 2005; Stangor et al., 2001), it had remained untested in the field of gender stereotyping.

Firstly, with the accuracy data, a significant 6.14% improvement in accuracy to stereotype incongruent word pairs across blocks was found, thus establishing the use of social consensus feedback as a useful stereotype-reduction strategy. However, this improvement is substantially lower than that reported in Experiment 1 (14.50%) in which performance-related feedback was provided to participants (a more comprehensive comparison of both experiments is provided in Section 3.7). Also, as in Experiment 1, accuracy to incongruent word pairs remained significantly lower than to stereotype congruent and neutral pairs by the end of the current study despite ample scope for further improvement.

The reaction time data tell a similar story. While response times to all congruency conditions decreased significantly from Block 1 to Block 3 (and most dramatically in the case of stereotype incongruent pairings), participants remained slower to respond to stereotype incongruent pairings than stereotype congruent or neutral pairs by the end of the experiment. Interestingly, RTs to incongruent pairings did not initially improve upon the introduction of the feedback in Block 2. Instead, it was when feedback was once again removed in Block 3 of the judgement trials that an acceleration of response times was evident. This pattern reveals that the social feedback had less of an immediate effect on responding than performance-related feedback where response times fell consistently across all three blocks. As this social feedback was conveyed in the form of a sentence, it is possible that participants took longer to process and digest the information, thus resulting in delayed changes to their patterns of responding.

Taken together, this accuracy and RT data reveal that presenting participants with social norm information is a useful means of attenuating the activation of automatic gender biases so as to

result in lower levels of stereotype application. Through reference to the fact that gender stereotyping was not tolerated among their peer group, it appears that participants were motivated to adapt their responding and conform to the perceived behaviour of their peers. Given the successful use of social norm information as a means of stereotype reduction towards other minority groups in the past (e.g. racial minorities and those suffering from obesity), Experiment 4 provides further support in favour of this strategy, but now in the domain of gender stereotypes. Nevertheless, it should be noted that the stereotype reduction effects reported above were relatively small and the feedback provided did not succeed in completely overcoming the impact of stereotype biases on responding, thus other stereotype-reduction strategies should also be considered in future research.

However, while it is hypothesised above that the social consensus feedback operated via social compliance mechanisms, a competing explanation for the success of this strategy is also possible. This issue is explored further in Experiment 5 below.

3.3 Experiment 5: Reverse social consensus feedback

3.3.1 Introduction

In Experiment 4, it was hypothesised that stereotype reduction was achieved through social compliance towards the perceived bias of a participant's peer-group. However, one further mechanism through which the social consensus feedback may have operated was by simply *alerting* participants to the issue of stereotype bias through the use of majority feedback, (either in support of or in opposition to a participant's judgements). This majority feedback may have simply reminded participants that nowadays males can do jobs typically held by women and vice versa – somewhat akin to the straightforward 'Correct'/'Incorrect' feedback of Chapter 2. Experiment 5 was therefore designed as a control experiment, aimed at differentiating between these two possibilities.

In order to successfully distinguish between the two options outlined above, the design of Experiment 5 remained identical to that of Experiment 4 but with one modification – the feedback to critical, stereotype incongruent word pairs was now centered on 50% (ranging from 35%-65%). This modified range of feedback was intended to suggest that people frequently *endorsed* stereotype biases, unlike Experiment 4 in which feedback implied that people rarely (2%-5% of the time) endorsed stereotypes. This form of feedback was now

termed reverse social consensus feedback (RSCF). The issue under investigation was whether people (a) "comply" with this RSCF by becoming more like their allegedly stereotyped peer group, and thus fail to reduce levels of stereotypic responding across blocks or (b) whether feedback alerts participants to the issue of stereotype biases and leads them by a relatively indirect route to accept counter-stereotypes as possible, thus successfully reducing levels of stereotypic responding across blocks.

More specifically, if participants conform to the feedback provided, and maintain stereotype biases following the provision of the social norm feedback in Block 2, no improvement in responding from Block 2 to Block 3 is anticipated. Conversely, if participants simply modify their behaviour once alerted to the issue of stereotype biases, an improvement in counter-stereotypic responding from Block 2 to Block 3 would be anticipated.

A pertinent issue for Experiment 5 was the range of feedback to present in response to stereotype incongruent pairings. Essentially, a much wider range of feedback responses was now deemed necessary than the range of 3% used in Experiment 4 (where it was conveyed that 2%-5% of past participants endorsed stereotyping, while 95% to 98% of past participants rejected stereotyping). Although such a narrow feedback range was previously appropriate, so as to strongly communicate that stereotyping was not supported, it was feared that this range would appear unrealistic if used in relation to stereotype endorsement i.e. it was not considered plausible to state that all previous students rejected stereotype incongruent pairings within any given 3% range, e.g. 50%-54%. A greater feedback range of 35%-65% (i.e. 30%) was consequently selected for the current study. Given that this chosen range was a relatively arbitrary decision, some doubt remained over how believable participants would find this feedback to be. A short pilot study was therefore conducted to assess the credibility of the feedback provided.

3.3.2 Pilot study 1

This pilot study was the first of two pilot studies which the participants completed one following the other. Each study lasted approximately 10 minutes and participants received 2 course credits for their participation[48]. In this first pilot study, 8 students (all female) were administered one block of judgement trials (as used in Experiment 4), but with the social feedback updated so as to suggest that past participants tended to endorse, rather than reject, stereotype biases. Participants were again informed that the feedback they would receive was

[48]The second pilot study was conducted for Experiments 8 and 9 and will be detailed further in Section 4.3.1.

indicative of the percentage of students in a previous experiment who agreed with their judgements.

As mentioned earlier, this newly constructed feedback was designed so as to centre on 50% for stereotype incongruent pairings (ranging from 35%-65% for both *yes* and *no* responses i.e. correct and incorrect judgements). Therefore, for half of these pairings, the feedback quite strongly indicated that stereotype biases were being endorsed (e.g. stating that a figure between 50%-65% of people had responded that the terms *bricklayer* and *aunt* could *not* refer to one person). For the other half of these stereotype incongruent pairings, the RSCF indicated that stereotype biases were being endorsed somewhat less often (for purposes of credibility), between 35% and 50%. However, even this lower range of stereotype endorsement was much greater than the endorsement portrayed in Experiment 4 (2%-5%). As before, exact figures that were used in conjunction with each of the stereotype incongruent pairings were randomly assigned within the specified range (of 35%-65%) but additionally distributed such that (a) no number appeared more than once and (b) the male and female incongruent pairings equally endorsed or rejected counter-stereotypic responding. Three distinct lists of these combinations were created (with the feedback values varied in each) and used randomly across participants. The feedback provided in response to the other congruency conditions was consistent with that outlined in Experiment 4; a full breakdown of the RSCF feedback is provided in Appendix 8.

In this pilot study, participants first completed a series of 8 practice trials (without feedback) to familiarise themselves with the judgement task. Once participants were satisfied with the instructions, they were then left alone in a quiet cubicle to finish the behavioural session. On completion, participants were instructed to call the experimenter who then asked them five questions about their experience of the task. The aim of these questions was to find out (1) what the participants believed the experiment to investigate (2) how believable participants found the RSCF to be, and (3) whether participants believed their responding was influenced by the feedback provided (see Appendix 9 for a list of the questions asked).

Question 3 was of most interest to the current study as participants rated how believable they found the RSF to be on a scale of 1 (believable) to 5 (unbelievable). On average participants judged it to be “quite believable” (M = 2, SD = 1.07). From the remaining questions, it was gleaned that participants could broadly identify the themes of the experiment (e.g. gender associations, stereotypes), yet all participants reported feeling influenced by the feedback they

received. The experimenter was satisfied with these findings on the plausibility of the feedback and Experiment 5 was next conducted with the same feedback parameters[49].

3.3.3 Method

Participants

Thirty-three students (17 female, 16 male) from the University of Sussex took part in this experiment. Participants' ages ranged from 18 to 29 years (*M*: 19.27; *SD*: 2.54). They received either £6 or 4 course credits for taking part in the session which lasted approximately 45 minutes.

Materials & Procedure

The materials used in this study were identical to those of Experiment 4, but with the fictitious social feedback updated according to the pilot study (Section 3.3.2). Similarly, all details of the procedure were identical to that of Experiment 4, aside from the debriefing. Participants were again informed of the aims of the experiment and reassured that the feedback provided was fictitious. However, it was further clarified that, although the feedback in this instance was expected to maintain the effects of stereotype biases, this experiment was in fact designed as a control condition for another study (Experiment 4) aimed at stereotype reduction, and that endorsement of gender stereotypes was not encouraged.

3.3.4 Results

Data screening

As in previous studies, data for the neutral term *adolescent* were removed before analysis, resulting in a loss of 1.32% of the data.

Analysis

Response times below 150ms, and above 4,000ms were excluded from analysis (representing 2.53% of the total data) along with times for all errors of judgement (representing a further

[49]The behavioural data from the judgement task in this pilot study was not analysed as the aim of the study was simply to ascertain whether participants found the fictitious feedback to be believable or not.

12.10%) totalling a loss of 14.63% of the data. Data trimming measures and two mixed ANOVAs were conducted as outlined in Section 2.2.3.

Accuracy

A main effect of Stereotype bias was revealed, F_1 (1.14, 35.33) = 15.20, $p < .001$; F_2 (2, 32) = 12.04, $p < .001$, with higher accuracy to word pairs that contained a neutral role term (M = 96.7%), than those that contained male-biased (M = 89.3%) or female-biased terms (M = 90.1%).

As anticipated, evidence of stereotyping was found with a main effect of Congruency[50], F_1 (1.04, 32.31) = 17.01, $p < .001$; F_2 (2, 32) = 60.68, $p <.001$. This effect was driven by significantly lower accuracy to stereotype incongruent word pairs (M = 80.9%) than to stereotype congruent (M = 98.25%) and neutral (M = 96.7%) pairings.

No main effect of Block was found in the by-participants analysis, F_1 (1.35, 41.89) = 1.91, p = .172, with accuracy rising only 1.6% across blocks. However, a significant effect of Block did emerge in the by-items analysis, F_2 (2, 64) = 6.07, p = .004.

Importantly, there was no evidence of a significant Block by Congruency interaction, F_1 (4, 124) = 0.31, p = .871; F_2 (4, 64) = 0.42, p = .800. The pattern of accuracy responding across blocks in each of the three conditions can be seen more clearly in Figure 3.3 below. It is apparent that there was no significant increase in accuracy of responses to stereotype incongruent word pairs from Block 1 to Block 3, t_1 (32) = 1.67, p = .105; t_2 (23) = 1.39, p = .178. Given the lack of improvement in accuracy scores across blocks, this data provides provisional evidence that participants complied with the perceived attitudes of their peer group, as opposed to attempting to overcome stereotype biases upon being alerted to the issue of stereotype bias through the feedback provided.

[50] i.e. an interaction of Stereotype bias by Kinship term gender.

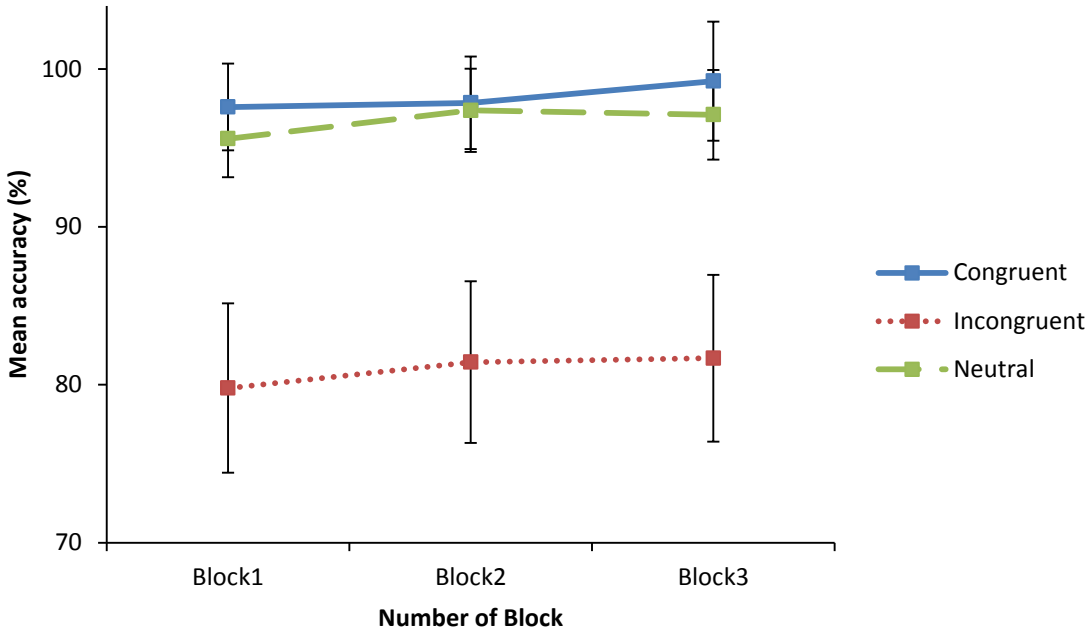


Figure 3.3. Experiment 5: Mean percentages of correct judgements to critical word pairs across blocks. Error bars indicate the 95% confidence intervals.

As in Experiment 4, a number of significant effects involving Participant Gender arose in the by-items analysis only. Firstly, there was a main effect of Participant Gender, F_2 (1, 32) = 83.93, $p < .001$, with male participants again achieving higher accuracy than females (94.5% vs. 89.6% respectively). A Participant Gender by Stereotype bias interaction was also found, F_2 (2, 32) = 7.0, $p < .001$ with male participants reaching higher levels of accuracy than female participants across all role nouns (although less so to those that were neutrally rated). Finally a Participant Gender by Congruency interaction was also revealed, F_2 (2, 32) = 18.53, $p < .001$, with both male and female participants displaying similar performance on stereotype congruent and neutral ratings, but with female participants showing much lower accuracy on stereotype incongruent pairings than male participants (M = 76.1% vs. M = 86.2% respectively). This superior male performance echoes that found in Experiment 4, which again contrasts sharply with the superior female performance consistently found in Chapter 2. Finally, the lack of equivalent significant effects in the by-participants analysis suggests greater levels of variance in this analysis relative to the by-items analysis.

Response times

A main effect of Block was found, F_1 (2, 62) = 23.17, $p < .001$; F_2 (2, 64) = 63.47, $p < .001$, with contrasts revealing a significant linear trend as response times decreased from Block 1 to Block 3, F_1 (1, 31) = 34.17, $p < .001$; F_2 (1, 32) = 118.89, $p < .001$.

A main effect of Congruency[51] was also revealed, F_1 (2, 62) = 21.84, p < .001; F_2 (2, 32) = 17.84, p < .001, with fastest response times to stereotype congruent word pairs (M = 760ms), followed closely by neutral (M = 777ms) and then incongruent pairings (M = 876ms). Importantly, there was again no interaction of Congruency by Block in either the by-participants (F_1 (4, 124) = .72, p = .578) or by-items analysis (F_1 (4, 64) = .96, p = .434), however, the pattern of by-participant responding across blocks is further examined in Figure 3.4 below.

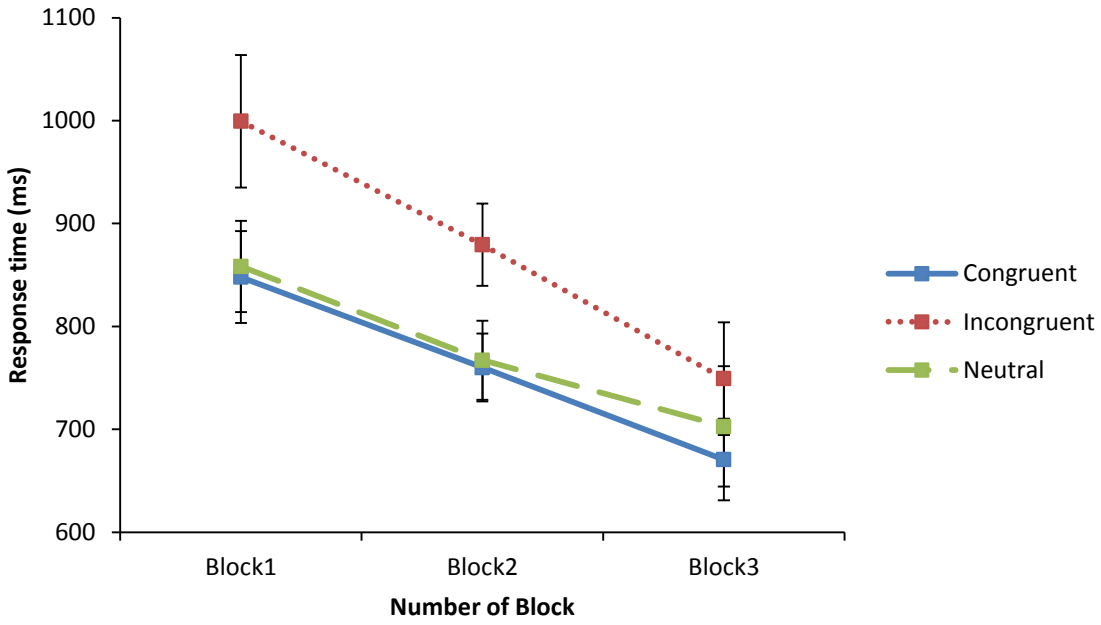


Figure 3.4. Experiment 5: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks. Error bars indicate the 95% confidence intervals.

Figure 3.4 reveals that RTs decreased across blocks in all three congruency conditions. Firstly, RTs to stereotype incongruent word pairs decreased significantly from Block 1 to Block 3, t_1 (32) = 5.27, p < .001, dz = . 92; t_2 (23) = 7.72, p < .001, dz = 1.58. Similarly, RTs to stereotype congruent and neutral word pairs decreased significantly across blocks, t_1 (32) = 5.08, p < .001, dz = .88; t_2 (23) = 7.60, p < .001, dz = 1.55 and t_1 (32) = 3.36, p = .002, dz = .59; t_2 (21) = 5.25, p < .001, dz = 1.12, respectively. The effect sizes reveal that a slightly larger magnitude of effect was found in the improvement across blocks to the incongruent word pairs, relative to the congruent and neutral pairings. Finally, despite significantly slower RTs to the stereotype incongruent pairings in Block 1 compared to RTs of both congruent (t_1 (32) = 3.92, p < .001, dz = .683; t_2 (23) = 8.76, p < .001, dz = 1.79)and neutral word pairs (t_1 (32) = 3.82, p = .001, dz = .67; t_2 (44) = 6.73, p < .001, d = 2.03.), only a significant difference between RTs of congruent

[51] i.e. an interaction of Stereotype bias by Kinship term gender.

and incongruent pairings remained by Block 3, t_1 (32) = 2.39, p = .023, dz = .42; t_2 (23) = 2.41, p = .025, dz = .49[52].

These improved RTs across blocks are relatively complicated to interpret in terms of the stated hypotheses. As in Experiment 2, the decrease may simply be due to practice effects, with participants consistently speeding up as the experiment progressed. On the other hand they may provide evidence that participants are improving at the judgement task upon being alerted to stereotype biases with the provision of feedback in Block 2. However, the latter of these two possibilities is deemed unlikely given a distinct lack of accompanying improvement in the accuracy data. Therefore, while the RT data do not directly provide support for the hypothesis that participants are conforming to perceived biases of their peers, when combined with the accuracy data it appears most likely that social compliance mechanisms are indeed the reason for which accuracy led to reduced stereotyping in Experiment 4.

Fillers - Accuracy

As expected, accuracy to the definitionally matching word pairs was higher than to the definitionally mismatching word pairs across the experiment (95.10% vs. 84.60% respectively). As has been consistently found in the previous studies, accuracy to definitionally mismatching pairs was poorer to word pairs containing a male-specific role name (M = 75.96%) than a female-specific role name (M = 93.23%). Once again it is hypothesised that this pattern of results is due to the generic interpretation of several of the definitionally masculine terms.

Fillers - Response times

As with the accuracy data, RTs to definitionally matching word pairs were faster than to definitionally mismatching pairs (872ms vs. 944ms respectively). Also, faster RTs to female word pairs over male word pairs were found in both the definitionally matching (838ms vs. 906ms respectively) and mismatching cases (907ms vs. 980ms respectively).

Again, these findings are in line with the accuracy data, as longer processing is likely to reflect participants' deliberation over terms which are masculine-specific by definition but frequently used in a more generic manner.

[52]No significant difference between RTs of incongruent and neutral word pairs was found by the end of the experiment, t_1 (32) = 1.16, p = .255; t_2 (44) = 1.19, p = .240.

3.3.5 Discussion

The aim of Experiment 5 was to establish the mechanism(s) through which social consensus feedback is likely to have succeeded as a stereotype reduction strategy in Experiment 4. While it was hypothesised that social compliance mechanisms were underlying the effects, it was alternatively possible that participants attempted to overcome stereotypic responding upon being alerted to the issue of stereotype biases through the variable feedback provided.

The results of Experiment 5 predominantly provide support for the first of these two proposals, i.e. that stereotype change resulted from social compliance mechanisms. This was concluded as accuracy of responses to stereotype incongruent pairings was not found to significantly improve across blocks, but remained in line with the perceived attitudes of participants' peers. Although RTs *were* found to significantly decrease across blocks (thus suggesting that participants were overcoming stereotypic responding upon being alerted to the issues of stereotype biases through the feedback provided), the lack of accompanying improvement in the accuracy data suggests that this pattern of results simply stemmed from practice effects.

Although accuracy to stereotype incongruent word pairs was not found to significantly increase across blocks, neither did it decrease. This pattern of results was not surprising and echoes data reported in past research that attitudes are more difficult to influence in a negative direction than positive. For example, Puhl et al. (2005, Experiment 1) asked participants to estimate the percentage of obese people who possess 10 negative and 10 positive traits. These authors report a significant increase in positive trait ratings after participants received social feedback indicating past students had responded in this direction. However, trait ratings did not change in the unfavourable feedback condition in which participants learnt that other students attributed obese people with more negative trait ratings than positive. Similarly, Stangor et al. (2001) report that attitudes towards racial minorities were easier to influence in a positive direction than negative (although they concede this may have been due to issues of social desirability).

In conclusion, Experiment 5 hints that social consensus feedback is likely to succeed as a stereotype reduction strategy through social compliance mechanisms as opposed to simple awareness of the issue of stereotype biases. However, before going further, it was important to ascertain whether Experiment 4 (in which participants received social feedback aimed at overcoming stereotype biases) actually resulted in significantly better performance than

Experiment 5 (in which participants received social feedback aimed at maintaining stereotype biases).

3.4 Experiments 4 and 5: Combined analysis

Data from Experiments 4 and 5 were next combined so as to more comprehensively examine the efficacy of social consensus feedback as a strategy for reducing levels of gender stereotype application. It was anticipated that accuracy and response times of judgements would be found to have improved to a greater extent across blocks in Experiment 4 relative to the control condition of Experiment 5.

Results

Analysis

The trimmed data from Experiments 4 and 5 were combined and both accuracy of judgements and response times were again analysed using mixed-design analyses of variance (ANOVAs), as described in Section 2.2.3. However, in the F_1 analyses, Experiment (Experiment 4 vs. 5) was further added as a between-subjects factor, while in the F_2 analyses it was added as a within-items factor.

Note that the findings reported below do not include effects that were revealed in both the individual experiment analyses (e.g. main effects of Block, Congruency etc.), but instead focus on the main interest of this combined analysis; accuracy of responses to critical trials across experiments, in particular to stereotype incongruent pairings.

Accuracy

Firstly, despite some variation in responding between the two experiments, the interaction of Congruency by Block by Experiment was significant in the by-items analysis only, F_1 (5.18, 292.1) = 1.78, p = .115; F_2 (3.58, 57.28) = 2.89, p = .035. From the by-participants pattern of responding to the critical trials across experiments in Figure 3.5 below[53], it can be seen that performance on stereotype congruent and neutral word pairs is consistently close to ceiling

[53] Note that the by-participants data was plotted here (despite significance in the by-items data) for reasons of consistency throughout the thesis and to mirror the more common practice in the literature of displaying by-participants data over by-items data.

level, while accuracy of stereotype incongruent word pairs is more variable across experiments.

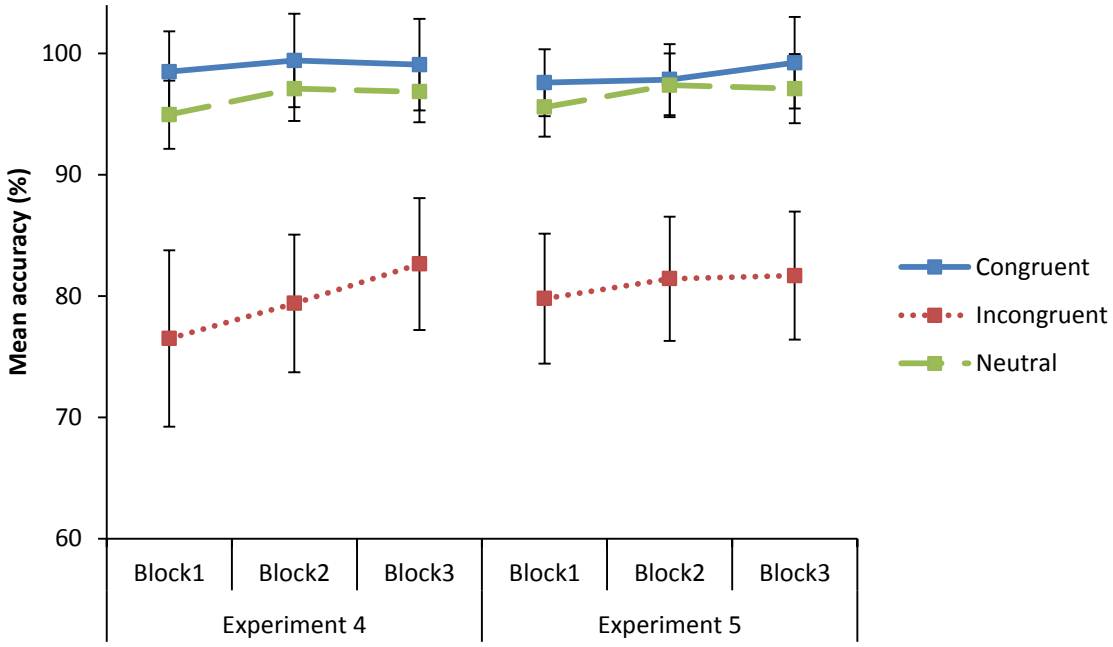


Figure 3.5. Mean accuracy to critical word pairs across blocks in Experiments 4 and 5. The vertical axis begins at 60% while error bars indicate the 95% confidence intervals.

In Experiment 4, accuracy to the incongruent pairings is seen to rise steadily across blocks (likely due to the provision of social feedback in Block 2) and as mentioned in Section 3.2.3, resulted in a significant improvement of 6.14% from Block 1 to Block 3 (p = .022). However, in Experiment 5, accuracy does not significantly increase across blocks (+1.89%), as participants were provided with reverse social feedback aimed at gender stereotype endorsement, as opposed to gender stereotype reduction (p = .105). Interpretation of results is somewhat complicated by the fact that final accuracy was very similar across experiments (due to Block 1 accuracy beginning higher in Experiment 5 than Experiment 4). However, the data suggests that performance to incongruent pairings improved to a greater extent across blocks in Experiment 4 due to social compliance mechanisms, than in the control condition in which participants were simply alerted to the issue of stereotype biases, but not encouraged to overcome them.

Response times

Next, a combined analysis of the RT data was carried out, with mean RTs to critical trials across experiments in the by-participants analysis depicted in Figure 3.6 below.

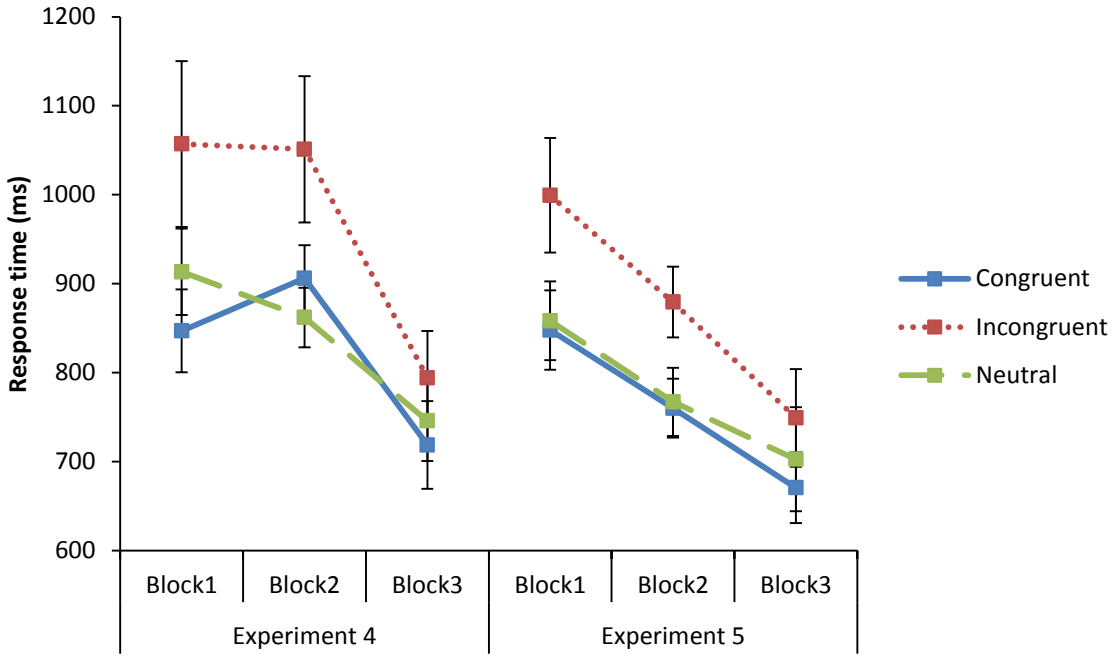


Figure 3.6. Mean response times (in milliseconds) to correct critical word pairs across blocks in Experiments 4 and 5. Error bars indicate the 95% confidence intervals.

Despite slight variation in the pattern of RT responding seen above, there was no evidence of a significant three-way interaction of Congruency by Block by Experiment, F_1 (7.84, 442.84) = 0.28, p = .975; F_2 (4, 64) = 0.06, p = .994. With the stereotype incongruent trials, it is apparent that RTs were slowest from the outset in Experiment 4. Indeed, even upon the introduction of the social feedback in Block 2, RTs were not found to immediately improve, however a sharp drop in RTs was then evident from Block 2 to Block 3. In Experiment 5, a more gradual decrease in RTs is found across the experimental blocks. As mentioned in the individual analyses, overall RTs decreased significantly across blocks in both experiments (263ms in Experiment 4 (p = .001) vs. 250ms in Experiment 5 (p < .001)). However, these means reveal that RTs did not improve to a greater extent across blocks when social norm information was presented to participants (Experiment 4) than when participants were simply alerted to the issue of stereotype biases (Experiment 5).

Discussion

This combined analysis of Experiment 4 and Experiment 5 has revealed that, despite a significant improvement in accuracy across blocks with the introduction of the social consensus feedback in Experiment 4, final accuracy performance was very similar across both experiments. This is attributed to the higher starting accuracy found in Experiment 5 relative to

Experiment 4. Exact reasons for this Block 1 difference remain unknown as there were no obvious differences between both samples of participants.

With the response time data, it emerged that RTs decreased significantly across blocks in both experiments, yet similar response times were found across experiments by Block 3. These results suggest that, on the whole, social consensus feedback is not an entirely successful strategy for overcoming activation of gender stereotypes, relative to the above control experiment at least. While weak effects of the feedback training were found on the accuracy data, no effects were revealed in the RT data. Future research could investigate contexts in which this social feedback is most effective, as success has previously been found when using this strategy as a means of reducing prejudice towards members of racial minorities and those suffering from obesity (e.g. Puhl et al., 2005; Stangor et al., 2001; Wittenbrink & Henly, 1996).

Overall, a detailed analysis of the use of feedback (both performance-related and social-norm related) as a means of reducing levels of gender stereotype application has been provided thus far. However, before progressing to an investigation of the use of counter-stereotype information in stereotype reduction in Chapter 4, one further experiment was designed, merging various elements of each of the studies carried out to date. This design was chosen so as to investigate whether two sources of feedback information can lead to greater stereotype reduction across blocks than one source.

3.5 Experiment 6: Accuracy and social feedback

3.5.1 Introduction

In the last of this series of studies aimed at reducing levels of gender stereotype application through the use of feedback, Experiment 6 sought to combine elements from the previous experiments, merging both accuracy information (as in Experiments 1 and 3) and social consensus information (as in Experiments 4 and 5) in the feedback provided. It was hypothesised that combining both of these sources of information into one feedback measure would lead to decreased stereotype application across blocks. However, Experiment 6 was somewhat exploratory in that it was not known (a) whether there would be a cumulative effect of providing both pieces of information, leading to greater levels of stereotype reduction than found in previous experiments, or (b) whether the direct correct/incorrect feedback already produces the maximum possible effect and that adding in the social feedback, which

seems to be weaker by itself, can't make any further improvement. Experiment 6 set out to investigate these issues.

Therefore, in this study, participants were again provided with feedback after responses in Block 2 of a three block judgement task, but with feedback comprised of both social and accuracy information, as will be described further below.

3.5.2 Method

Participants

Fifty monolingual, native English speakers (25 male, 25 female) from the student population of the University of Sussex took part in this experiment. Participants' ages ranged from 18 to 31 years (*M*: 21.22; *SD*: 3.20). They received either £5 or 4 course credits for taking part in the session which lasted approximately 45 minutes.

Materials

Materials were identical to those used in Experiment 4 but with accuracy information added to the social norm feedback. For example, where participants previously received feedback stating '*96% of previous students agreed with you'*, this was now adapted to state *'Yes and 96% of previous students agreed with you'.* Secondly, accuracy information was conveyed through the colour of the feedback text. All correct responses displayed feedback in blue text (the E-prime default for correct feedback responses), while all incorrect responses displayed feedback in red text (the E-prime default for incorrect feedback responses). All further details of the design and procedure of this experiment were as described in Experiment 4 (Section 3.2.2), but with instructions updated so as to include both accuracy and social information in the feedback examples[54].

3.5.3 Results

Data screening

[54] Note that, unlike the performance feedback of Chapter 2, the accuracy feedback in this experiment did not provide cumulative percentage scores, and responses stated *Yes* and *No* as opposed to *Correct* and *Incorrect* (so as to sound more natural).

As in previous studies, data for the role noun *adolescent* was removed, resulting in the loss of 1.32% of the data.

Analysis

Response times below 150ms, and above 4,000ms were identified and removed from analysis (representing 3.77% of the total data) as were times for all errors of judgement (representing a further 11.04%) totalling a loss of 14.81% of the data. Analyses were conducted as described in Section 2.2.3.

Accuracy

Analysis revealed a significant main effect of Stereotype bias, F_1 (1.35, 64.93) = 14.76, $p < .001$; F_2 (2, 32) = 8.38, $p = .001$, with accuracy of responses to neutral role nouns (M = 95.7%) higher than to male-biased (M = 91.4%) and female-biased (M = 90.9%) role nouns.

A significant interaction of Stereotype bias by Block was also found, F_1 (2.87, 137.66) = 5.87, p = .001; F_2 (4, 64) = 5.26, $p = .001$, with a greater improvement in accuracy to male and female-biased role nouns across blocks (+4.9% and +6.0% respectively) than to neutral role nouns (+1.6%). However, accuracy to neutral role nouns was higher from the outset, thus resulting in less scope for improvement than in the other stereotype conditions.

A significant interaction of Kinship term gender by Participant gender was also revealed, F_1 (1, 48) = 4.44, $p = .040$; F_2 (1, 32) = 4.75, $p = .037$, with female participants more accurate in response to word pairs containing female kinship terms than male kinship terms (M = 93.5% and 91.9% respectively), while accuracy of male participants was similarly high across both kinship terms (M = 92.9% to word pairs involving a male kinship term and M = 92.3% to those involving a female kinship term).

A significant main effect of Block was also found, F_1 (1.42, 67.90) = 10.14, $p < .001$; F_2 (2, 64) = 42.43, $p < .001$. Contrasts revealed this was a linear trend with a consistent increase in accuracy across blocks from 90.4% to 94.6%, F_1 (1, 48) = 12.20, $p = .001$; F_2 (1, 32) = 66.71, $p <$ .001.

There was also a significant effect of Congruency[55], F_1 (1.05, 50.15) = 17.29, $p < .001$; F_2 (2, 32) = 74.56, $p < .001$. As anticipated, accuracy to congruent and neutral word pairs was

[55] i.e. an interaction of Stereotype bias by Kinship term gender.

significantly higher (*M* = 98.15% vs. *M* = 95.70% respectively) than to incongruent pairings (*M* = 84.14%).

Finally, a significant interaction of Block by Congruency was also revealed, F_1 (2.08, 100.05) = 7.87, p = .001; F_2 (4, 64) = 21.11, p < .001. While accuracy improved across blocks in each of the three congruency conditions, the sharpest increase was found in response to stereotype incongruent pairings (+9.33% from Block 1 to Block 3), followed by the neutral pairings (+1.65%) and finally the congruent pairings (+1.55%). This pattern can be seen in more detail in Figure 3.7 below.

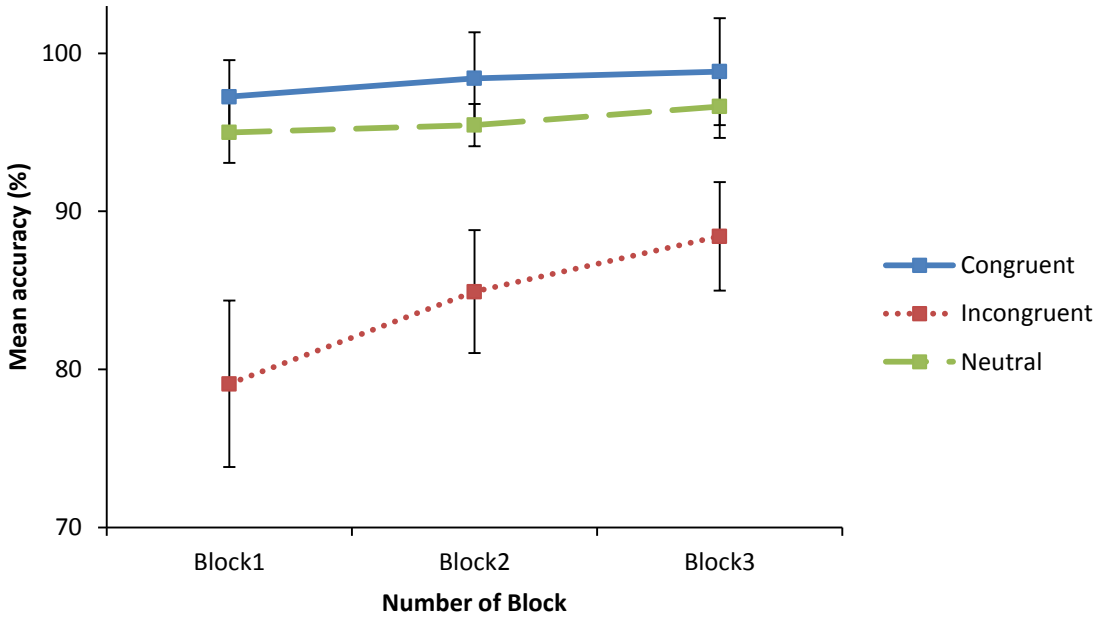


Figure 3.7. Experiment 6: Mean percentages of correct judgements to critical word pairs across blocks. Error bars indicate the 95% confidence intervals.

This Congruency by Block interaction was further examined using pairwise comparisons[56]. The aforementioned 9.33% improvement in accuracy of the incongruent pairings from Block 1 to Block 3 was indeed a significant one, t_1 (49) = 4.07, dz = .58, p < .001; t_2 (23) = 10.39, p < .001, dz = 2.12, confirming that combined social norm and accuracy feedback is an effective means of moderating the activation of stereotype biases associated with certain role nouns in English. However, while congruent items were responded to significantly more accurately than incongruent items in Block 1: t_1 (49) = 4.96, p < .001, dz = .70; t_2 (23) = 11.51, p < .001, dz =

[56] It was anticipated that accuracy to congruent items would be higher than to incongruent items in Block 1, and also that performance on these incongruent pairings would improve after the feedback training (i.e. from Block 1 to Block 3). Therefore, one-tailed *t*-tests were used for these comparisons while the remaining differences were examined using two-tailed tests.

2.35, this pattern also emerged in Block 3: t_1 (49) = 3.20, p = .002, dz = .45; t_2 (23) = 9.28, p < .001, dz = 1.89. Therefore, despite the improvement in accuracy of incongruent pairings from Block 1 to Block 3, accuracy ultimately remained poorer on these word pairs than to congruent pairings. Similarly, a significant difference between accuracy of incongruent and neutral pairings was found in Block 1, t_1 (49) = 5.0, p < .001, dz = .71; t_1 (29.78) = 9.89, p < .001, d = 3.62, and this difference was again significant in Block 3 after the social feedback training, t_1 (49) = 3.47, p = .001, dz = .49; t_1 (39.27) = 6.58, p < .001, d = 2.10.

Overall, while this combined social and accuracy feedback proved an effective means of reducing stereotyping to incongruent trials, final accuracy scores remained higher in the other congruency conditions, again highlighting scope for further improvement.

Response times

A main effect of Participant Gender was found in the by-items analysis only, F_2 (1, 32) = 45.94, p < .001, with male participants slower to respond than females (M = 861ms vs. M = 788ms).

A significant effect of Congruency[57] on the response times of judgements was also found, F_1 (2, 96) = 20.09, p < .001; F_2 (2, 32) = 11.10, p < .001, with fastest RTs found in response to congruent pairings (M = 798ms), followed by neutral (M = 837ms) and incongruent (M = 891ms) pairings respectively.

There was also a significant main effect of Block, F_1 (1.72, 82.42) = 45.60, p = .001; F_2 (1.70, 60.57) = 126.59, p < .001, with contrasts revealing a linear trend as response times decreased consistently across blocks, F_1 (1, 48) = 80.41, p < .001; F_2 (1, 32) = 273.14, p < .001.

[57] i.e. an interaction of Stereotype bias by Kinship term gender.

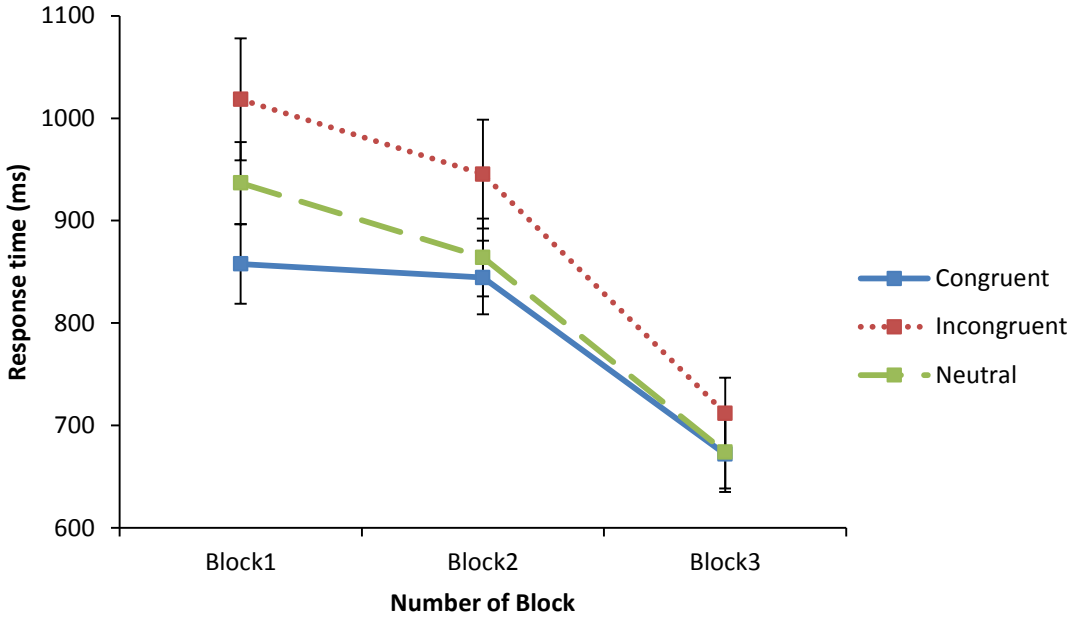


Figure 3.8. Experiment 6: Mean response times (in milliseconds) of judgements to correct critical word pairs across blocks. Error bars indicate the 95% confidence intervals.

A significant interaction of Block by Congruency was found in the F_1 analysis only, F_1 (4, 192) = 3.32, p = .012; F_2 (4, 64) = 2.08, p = .094, with a series of paired t-tests again performed to investigate this interaction further, according to the same criteria as the previous accuracy data.

It was first revealed that performance to stereotype incongruent pairings significantly improved from Block 1 to Block 3 (by 306ms), with RTs getting continuously faster as the task progressed, t_1 (49) = 7.67, p < .001, dz = 1.08; t_2 (23) = 8.81, p < .001, dz = 1.80. Indeed, while congruent items were responded to significantly faster than incongruent items in Block 1, t_1 (49) = 5.26, p < .001, dz = .74; t_2 (23) = 4.42, p < .001, dz = .90, this pattern was only marginally significant in Block 3, t_1 (49) = 1.83, p = .074, dz = .26; t_2 (23) = 1.80, p = .085, dz = .37. Similarly, a significant difference was found between RTs of incongruent and neutral pairings in Block 1, t_1 (49) = 3.12, p = .003, dz = .44; t_1 (44) = 3.48, p < .001, d = 1.05, but again this disappeared by Block 3 in the by-items analysis, t_1 (49) = 1.63, p = .108; t_2 (44) = 2.61, p = .012, d = .79, with RTs to incongruent pairings decreasing to fall in line with those of neutral pairings. Therefore, with this combined accuracy and social feedback information, RTs to incongruent word pairs were indeed found to fall dramatically across the experimental blocks, generally finishing just marginally slower than RTs in the other congruency conditions by Block 3.

Fillers - Accuracy

Performance on the filler trials was similar to findings reported in the previous experiments. Firstly, accuracy was high in the definitionally matching condition (*M* = 94.2%), yet once again fell in response to definitionally mismatching pairings, with an average of 83.3% correct judgements. This relatively poor performance again stems from low accuracy to certain definitionally masculine items, as participants tended to judge females as eligible to fulfil some 'male-specific' roles e.g. Block 1 accuracy to the terms *host* and *hero* was 10% and 22% respectively. However, an improvement was seen as the experiment progressed, with average accuracy to male, definitionally mismatching pairings rising from 71.7% in Block 1 to 78.9% in Block 3. Again, accuracy performance was much higher to female, definitionally mismatching pairings, with the average percentage correct in Block 1 at 90%, rising slightly to 91.9% by Block 3.

Fillers - Response times

As in previous studies, average RTs to the definitionally matching word pairs were faster than to the definitionally mismatching word pairs (*M* = 896ms vs. *M* = 1011ms respectively), with little difference between RTs to the male and female definitionally matching terms across blocks.

Again, average RTs to male role names in the mismatching condition were somewhat slower (*M* = 1060ms) than to female role names (*M* = 961ms). This pattern of results is likely to reflect the increased processing time taken by participants when deciding if certain terms (that are definitionally male, but generically used) can also be used in reference to females.

3.5.4 Discussion

The results of Experiment 6 provide support for the moderating effect of combined social-consensus and accuracy information on gender stereotype application. Accuracy of the word-pair judgements was seen to significantly increase across Blocks 1-3 in the stereotype incongruent condition. However, as in previous experiments, the use of this feedback strategy did not completely overcome the stereotype bias associated with certain role nouns in English. Response accuracy to stereotype congruent and neutral word pairs remained significantly higher than to stereotype incongruent pairings at the end of the experiment. The response time data were also found to significantly decrease across blocks. Moreover, final RTs to

stereotype incongruent pairings were found to be just marginally slower than to stereotype congruent and neutral word pairs by Block 3.

Despite these encouraging findings, further analysis was required to establish whether this combined accuracy and social feedback led to significantly greater stereotype reduction than the social feedback alone.

3.6 Experiments 4 and 6: Combined analysis

Data from Experiments 4 and 6 were next compared so as to more comprehensively examine whether the use of social *and* accuracy feedback in Experiment 6 resulted in significantly improved performance relative to Experiment 4, in which social feedback was presented alone. It was anticipated that performance would improve to a greater extent in Experiment 6 than in Experiment 4 owing to the extra information provided in the former study. More specifically, it was hypothesised that accuracy of stereotype incongruent word pairs in Experiment 6 would be significantly higher than accuracy of Experiment 4 by Block 3 while response times to stereotype incongruent word pairs in Experiment 6 would be significantly faster than those of Experiment 4 by Block 3.

Results

Analysis

The trimmed data from Experiments 4 and 6 were combined and both accuracy of judgements and response times (RTs) were again analysed using mixed-design ANOVAs, as described in Section 2.2.3. However, in the F_1 analyses, Experiment (Experiment 4 vs. 6) was further added as a between-subjects factor, while in the F_2 analyses it was added as a within-items factor.

Again, the findings reported below do not include effects that were revealed in both the individual experiment analyses but focus solely on the relationship of most interest i.e. the interaction of Experiment by Congruency by Block.

Accuracy

Contrary to expectations, a significant three-way interaction of Congruency by Block by Experiment was not found in either analysis, F_1 (2.26, 189.95) = 0.47, p = .647; F_2 (4, 64) = 0.92,

p = .456. As can be seen in Figure 3.9 below, patterns of by-participants responding were very similar across experiments, although accuracy to stereotype incongruent trials is the most variable condition across experiments.

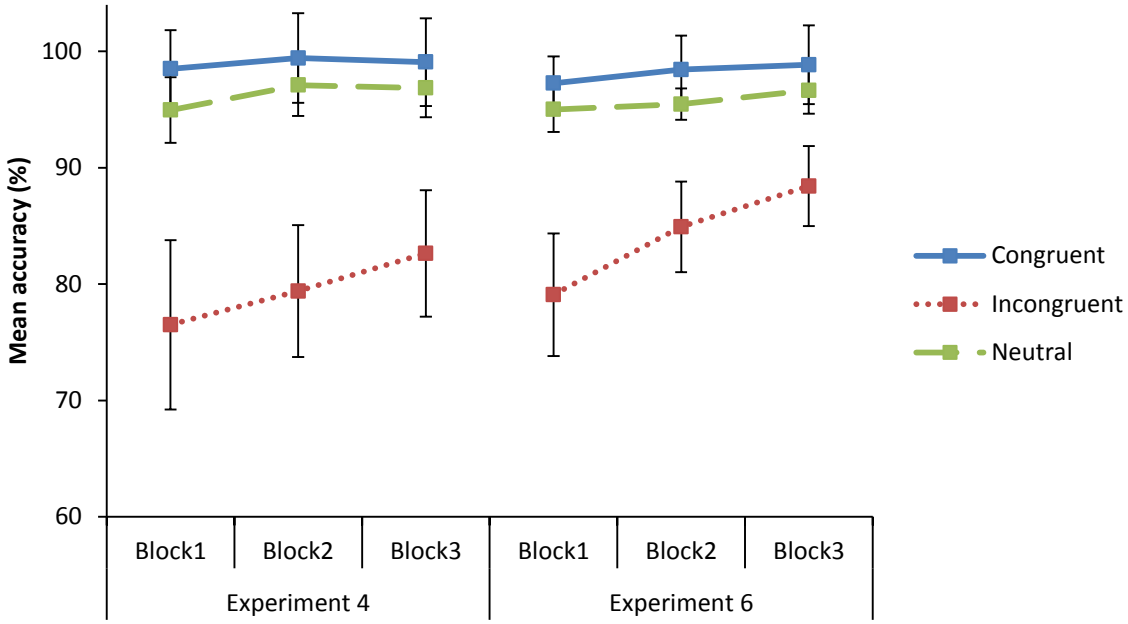


Figure 3.9. Mean % accuracy to critical word pairs across blocks in Experiments 4 and 6. The vertical axis begins at 60% while error bars indicate the 95% confidence intervals.

In Experiment 4, accuracy to the incongruent pairings is seen to rise quite steadily across blocks, after the provision of social feedback in Block 2. As mentioned earlier, this resulted in a significant increase of 6.14% across blocks (*p* = .022). However, a greater improvement is evident in Experiment 6, with accuracy increasing 9.33% across Blocks 1 to 3, an improvement which was highly significant (*p* < .001).

Next, to examine performance to the incongruent trials in more detail, a second Experiment (4 vs. 6) by Block (1 vs. 3) ANOVA was conducted on these trials only.

Despite means showing that accuracy increased to a greater extent following the combined feedback of Experiment 6 relative to the social consensus feedback presented alone in Experiment 4, only a marginally significant interaction of Experiment by Block was revealed in the by-items analysis, F_2 (1, 23) = 3.45, *p* = .076, while no interaction was found in the by-participants analysis, F_1 (1, 84) = 0.76, *p* = .386.

Next, two-tailed *t*-tests were used to examine Block 1 and Block 3 performance across experiments. Again, while no significant differences were found in the by-participants analysis

(Block 1: t_2 (84) = .371, p = .711; Block 3: t_2 (84) = 1.05, p = .297), the by-items data revealed a marginally significant difference between accuracy across experiments in Block 1, t_2 (23) = 2.03, p = .055, dz = .41, which became highly significant in Block 3, as accuracy in Experiment 6 rose to a greater extent than in Experiment 4, t_2 (23) = 4.99, p < .001, dz = 1.02. Although no significant effects were observed in the by-participant data (likely due to reasons of variability as explained in Section 2.2.3), this pattern of results in the by-items data indicates that combined social and accuracy feedback is a somewhat more effective means of reducing stereotyping on this judgement task, than social feedback alone.

Response times

With the response time data, there was again no evidence of a significant three-way interaction of Congruency by Block by Experiment, F_1 (3.77, 308.74) = 0.02, p = .998; F_2 (4, 64) = 0.23, p = .922. However, the F_1 pattern of results to critical pairings across blocks and experiments was next examined, and is displayed in Figure 3.10 below.

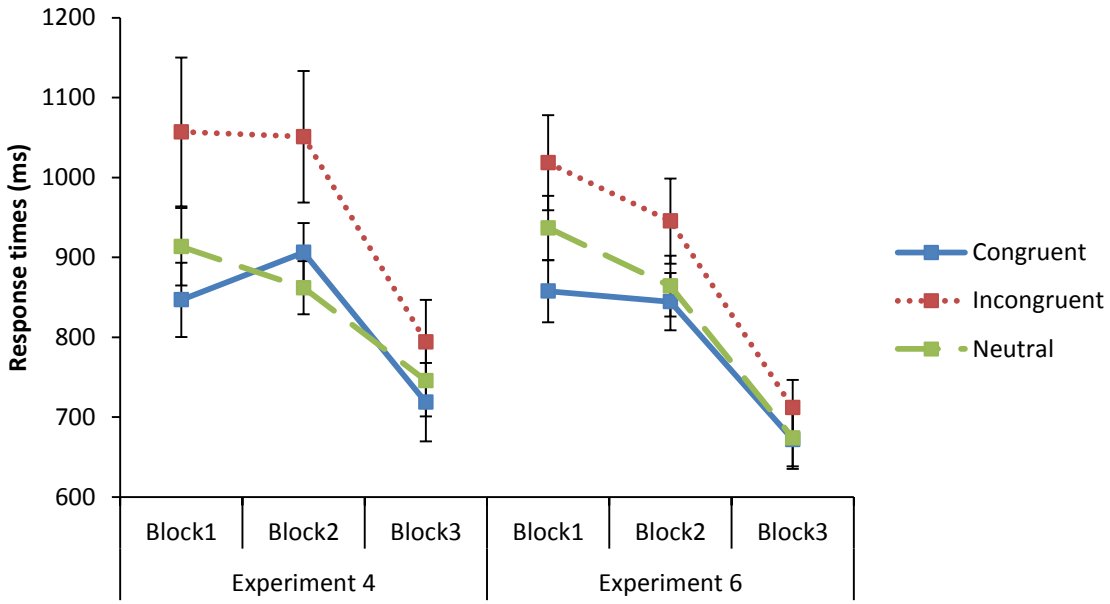


Figure 3.10. Mean response times (in milliseconds) to correct critical word pairs, across blocks in Experiments 4 and 6. Error bars indicate the 95% confidence intervals.

Figure 3.10 reveals that RTs across conditions in Experiment 6 are slightly faster by the end of the experiment, than those of Experiment 4. However, the pattern of responding to stereotype incongruent pairings across both experiments is strikingly similar; RTs were not found to

improve a great deal from Block 1 to Block 2, yet a sharp drop was then evident from Block 2 to Block 3.

Next, a second Experiment (4 vs. 6) by Block (1 vs. 3) ANOVA was conducted on the stereotype incongruent trials only so as to examine performance to these trials in more detail.

However, again there was no evidence of an Experiment by Block interaction, F_1 (1, 84) = 0.35, p = .555; F_2 (1, 23) = 2.20, p = .151. This pattern of results is not surprising given the similar RT pattern revealed across blocks and indicates there were no notable differences between RT responses to stereotype incongruent pairings before and after the respective feedback trainings in Experiment 4 and Experiment 6.

Discussion

Overall, trends in the accuracy and RT data hinted that performance to critical trials was significantly better when participants received combined social and consensus feedback (Experiment 6) as opposed to consensus feedback alone (Experiment 4). However, this combined analysis revealed only a marginally significant Block by Experiment interaction in the by-items accuracy data when analysing the stereotype incongruent trials. It can therefore be concluded that the two different feedback trainings used in Experiments 4 and 6 did not give rise to any striking differences in responding to stereotype incongruent pairings across blocks.

3.7 Performance feedback vs. Social consensus feedback

Thus far, Chapter 2 (investigating performance-related feedback) and Chapter 3 (investigating social consensus feedback) have provided an in-depth investigation into the efficacy and value of differing feedback strategies when used to overcome gender stereotype application. However, two issues remain unresolved (1) which of these two forms of feedback is most effective as a stereotype reduction strategy? and (2) is performance feedback more effective at overcoming stereotype biases when combined with social consensus information or when presented alone?

The first of these two questions was investigated by comparing performance on Experiment 1 (performance-related feedback) with Experiment 4 (social consensus feedback), while question

2 was examined by comparing performance on Experiment 1 (performance-related feedback) with Experiment 6 (combined accuracy and social consensus feedback).

The trimmed data from each of these experiments (Experiments 1, 4, and 6) were combined and both accuracy of judgements and response times (RTs) were again analysed using mixed-design ANOVAs, with Experiment (Experiment 1/4/6) included as a between-subjects/ within-items factor in the by-participants and by-items analyses respectively. However, as results pertaining to each of these experiments have been extensively outlined in previous sections, only interactions that shed light on the comparisons outlined above will be discussed.

Results

Accuracy

The analysis revealed a three-way interaction of Congruency by Block by Experiment that was marginally significant in the by-participants analysis and highly significant in the by-items analysis, F_1 (3.92, 265.59) = 2.08, p = .082; F_2 (8, 128) = 4.97, p < .001. The data from this by-participants analysis is displayed in Figure 3.11 below.

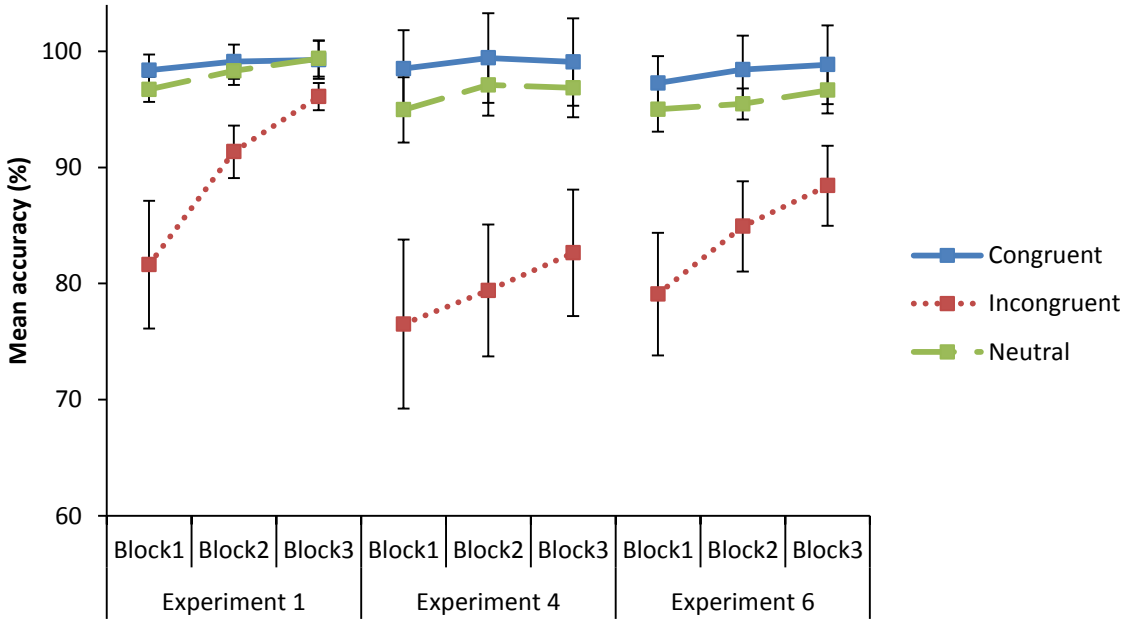


Figure 3.11. Mean percentages of correct judgements to critical word pairs across blocks in Experiments 1, 4 and 6. The vertical axis begins at 60% while error bars indicate the 95% confidence intervals.

Given the ceiling effects seen in response to stereotype congruent and neutral pairings across experiments, it is clear that this three-way interaction of Congruency by Block by Experiment stems from variable performance on the stereotype incongruent trials. Therefore, to investigate performance on these incongruent trials in more detail, a second set of ANOVAs were conducted examining (a) Experiment (1 vs. 4) by Block (1 vs. 3) performance and (b) Experiment (1 vs. 6) by Block (1 vs. 3) performance.

The comparison of Experiment 1 (performance feedback) and Experiment 4 (social consensus feedback) revealed a marginally significant interaction of Experiment by Block in the by-participants analysis but again a highly significant effect in the by-items analysis, F_1 (1, 85) = 3.68, p = .058; F_2 (1, 23) = 16.67, p < .001. With the by-participants data, no significant difference was found between incongruent performance accuracy in Block 1, t_1 (62.13) = 0.76, p = .453, yet a significant difference was found in Block 3, t_1 (38.79) = 2.83, p = .007, d = .91. This pattern of results reflects the greater improvement in accuracy across blocks in Experiment 1 relative to Experiment 4 (while taking into account variable Block 1 performance). In the by-items analysis, a similar pattern was found. Initially there was a marginally significant difference in Block 1 accuracy to incongruent pairings, t_2 (46) = 2.13, p = .039[58], d = .63, yet this became highly significant by Block 3, t_2 (33.08) = 6.84, p <.001, d = 2.38. These results are again reflective of the superior improvement in accuracy across blocks in Experiment 1 relative to Experiment 4.

Next, the comparison of Experiment 1 (performance feedback) and Experiment 6 (combined social consensus and accuracy feedback) revealed a non-significant interaction of Experiment by Block in the by-participants analysis, F_1 (1, 99) = 1.83, p = .179, yet this interaction was significant in the by-items analysis, F_2 (1, 23) = 19.65, p <.001. With the by-participants analysis, the mean scores revealed that Block 1 accuracy scores were very similar across experiments (mean difference = 2.53%[59]), while Block 3 accuracy was found to be 7.7% higher in Experiment 1 (96.1%) than Experiment 6 (88.4%). Indeed, a two-tailed, independent samples t-test revealed that this Block 3 performance was significantly different across the two experiments, t_1 (59.32) = 2.22, p = .031, d = .58; t_2 (46) = 5.69, p < .001, d = 1.68. These results again reflect the greater improvement in accuracy across blocks in Experiment 1 relative to Experiment 6 (while taking into account variable Block 1 performance), indicating that performance-related feedback is more effective as a means of stereotype reduction when presented on its own as opposed to when combined with social consensus information.

[58] This is above the Bonferroni-adjusted cut off of .007 for significance.
[59] This was not a significant difference, t_1 (99) = .46, p = .650; t_2 (46) = 1.08, p = .286.

Overall, these combined analyses reveal that performance-related feedback led to superior responding to stereotype incongruent word pairs than using social consensus feedback alone, or combined social and accuracy feedback as in Experiments 4 and 6 respectively.

Response times

Next, analysis of the response time data to critical trials revealed there was no significant interaction of Congruency by Block by Experiment, F_1 (7.80, 511.01) = 0.69, p = .70; F_2 (8, 128) = 0.52, p = .838. However, the by-participants pattern of responding is displayed in Figure 3.12 below so as to further investigate RTs across experiments.

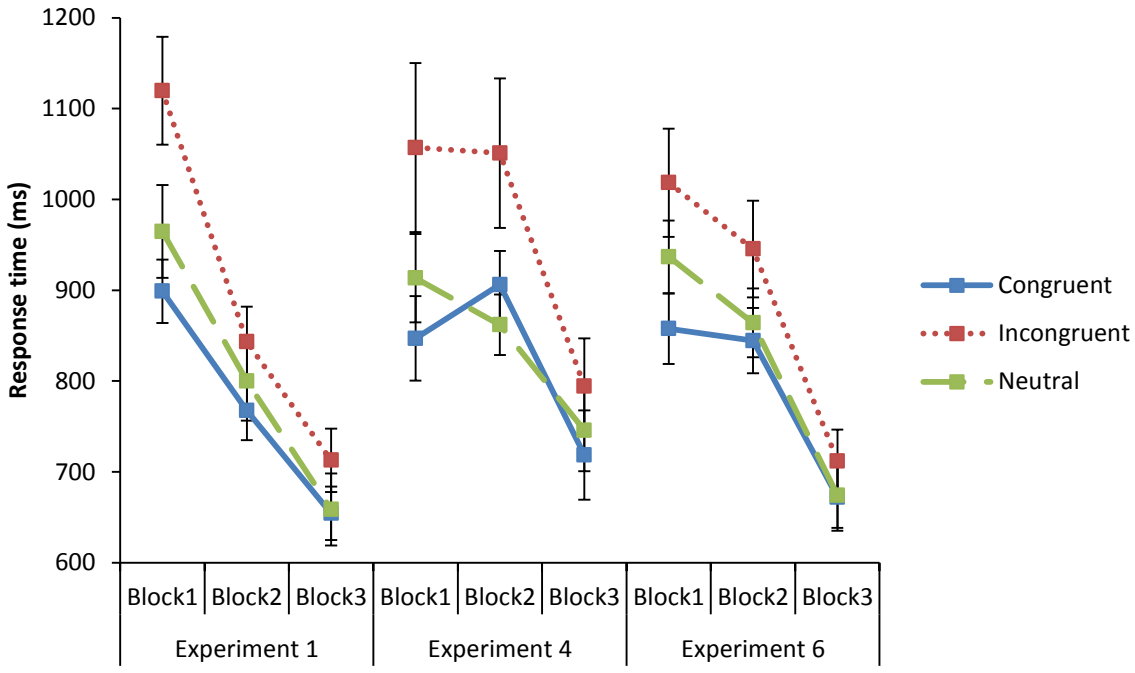


Figure 3.12. Mean response times (in milliseconds) to correct critical word pairs across blocks in Experiments 1, 4 and 6. Error bars indicate the 95% confidence intervals.

As with the accuracy data, a second set of ANOVAs were conducted on the incongruent trials, examining (a) Experiment (1 vs. 4) by Block (1 vs. 3) performance and (b) Experiment (1 vs. 6) by Block (1 vs. 3) performance. The comparison of Experiment 1 (performance feedback) and Experiment 4 (social consensus feedback) revealed a significant interaction of Experiment by Block, F_1 (1, 85) = 3.96, p = .050; F_2 (1, 23) = 14.22, p = .001, with quite variable performance across the two experiments, as displayed in Figure 3.12 above. This interaction is driven by the fact that response times were initially slower in Experiment 1 than in Experiment 4 (by 62ms, ps > .4), yet finished faster in Experiment 1 (by 82ms, ps > .1). Consequently, although final

Block 3 performance was not significantly different across experiments, the data trend suggests that Experiment 1 had the more effective feedback strategy of the two.
Finally, the comparison of Experiment 1 (performance feedback) and Experiment 6 (combined social consensus and accuracy feedback) revealed a marginally significant interaction of Experiment by Block in the by-participants analysis, F_1 (1, 99) = 3.26, p = .074, and a highly significant effect in the by-items analysis, F_2 (1, 23) = 8.29, p = .008. It appears this marginal interaction is driven by the slower Block 1 RTs of Experiment 1 (1120ms) than Experiment 6 (1018ms), with a mean difference of 102ms. This Block 1 difference across experiments was significant in the by-items data only, t_2 (46) = 2.56, p = .014, d = .75. In contrast, final RT scores differ by just 3ms. Therefore, while RTs reduced to a greater extent across blocks when participants received performance feedback as opposed to combined accuracy and social consensus information, final RTs were almost identical across these experiments.

Discussion

The final analyses outlined above sought to unite data from Chapters 2 and 3 in an effort to definitively establish which of the three feedback strategies (performance feedback, social consensus feedback, combined accuracy and social feedback) was most successful as a stereotype reduction strategy. Performance-related feedback stood out in this regard, achieving greater levels of stereotype accuracy than social consensus information when presented on its own (Experiment 4) and when combined with accuracy information (Experiment 6), by the end of the experiment. This pattern of results was largely mirrored in the RT data, with response times found to reduce to a greater extent across blocks in Experiment 1, yet final RTs were not significantly different across experiments. Reasons for which performance feedback was found to be a more effective stereotype reduction strategy than combined social and accuracy information remain unclear. However, it is possible that with the combined feedback, participants may have become confused or had trouble digesting both pieces of information. Future research is required to investigate this issue further.

Overall, these findings have important theoretical implications for the field of prejudice and stereotype reduction. With a growing number of studies turning to social norm information as a strategy for overcoming stereotypes, data from Chapters 2 and 3 reveals that stereotyping may be more successfully overcome by providing participants with accurate, performance-related information about their own personal stereotype tendencies. It appears that once people are alerted to the influence of stereotype biases on their behaviour they endeavour to

overcome this stereotype effect. While it is initially effortful for participants to succeed in reducing levels of gender stereotype application, this process can become more automated over time as participants become more familiar with, and accepting of, counter-stereotypical word pairs.

3.8 Chapter Discussion

As mentioned in Chapter 1, knowledge of the power of conformity resulting from landmark research by Milgram, Asch and Zimbardo remains under-used in laboratory-based stereotype-reduction research. Furthermore, in their meta-analytic review of the malleability of gender stereotypes, Lenton et al. (2009) posited a need for more research on how social motives may influence automatic gender stereotypes. Chapter 3 aimed to address these issues, investigating the efficacy of social-consensus feedback as a strategy for reducing gender stereotype application in a judgement task. While use of social norm information has previously proven successful as a strategy for overcoming prejudice in relation to racial minorities (Stangor et al., 2001) and those suffering from obesity (Puhl et al., 2005), it had remained untested in the field of gender stereotypes.

To begin with, Experiment 4 was designed as a preliminary investigation into the efficacy of social consensus feedback as a stereotype reduction strategy. Results indicated that both accuracy and response times improved significantly following the provision of the feedback, albeit to a lesser extent than in Experiment 1 where performance-related feedback was provided.

Next, Experiment 5 aimed to establish whether participants modified their behavioural responses upon the provision of social feedback as a result of compliance towards the perceived attitudes of their peer-group or simply due to an awareness of the topic of stereotype bias, incited through the provision of feedback. Although the RT data was ambiguous on this point, a lack of improvement in accuracy across blocks provides strong evidence for the former theory i.e. that it is indeed social compliance which drives behavioural change towards that of one's peer-group upon the provision of social consensus information.

However, in contrast to expectations, a comparison of Block 3 data of Experiments 4 and 5 revealed that final accuracy in Experiment 4 was not found to be significantly higher than in Experiment 5 (despite the fact that stereotypes were perceived to be rejected in the former and endorsed in the latter). These findings are likely to stem from the fact that starting

accuracy to incongruent pairings was lower in Block 1 of Experiment 4 than Block 1 of Experiment 5. The same pattern of results was found in the RT data.

Experiment 6 was designed as a final feedback-related study in which participants were provided with feedback based on two sources of information (accuracy and social consensus) as opposed to one. This combined feedback strategy was found to reduce levels of gender stereotype application across blocks, and to a greater extent than reported in Experiment 4 (social consensus information only), 9.33% vs. 6.15% respectively (although this difference was only significant in the by-items analysis). However, there was no significant difference in RT responding to incongruent items across these two experiments.

Finally, as mentioned earlier, performance on Experiment 1 (performance feedback) consistently outshone that of Experiment 4 (social consensus feedback) and Experiment 6 (combined social and accuracy feedback).

But how do these findings sit within the existent stereotype reduction literature?

The results of Experiment 4 fall in line with the claim of Prentice and Miller (1993) that participants will attempt to move their personal attitudes towards that of the perceived norm when they perceive their attitudes to be different from the normative attitude of their peer group. Furthermore, results provide support for Social Reality Theory (Hardin & Conley, 2000) and Group Norms Theory (Crandall et al., 2002; Kelman, 1958; Sherif & Sherif, 1953), which both posit that the human objective of social connection drives people to validate their experiences with others and to exhibit behaviours valued by admirable in-group members. Conversely, the findings in this chapter do not provide support for Deviance Regulation Theory (Blanton & Christie, 2003), which posits the rejection of perceived attitudes of peers as a means of self-definition.

However, while previous research has documented the successful use of social feedback in overcoming prejudice and stereotyping, the success of this strategy proved somewhat limited in Experiment 4, as accuracy increased a relatively small amount across blocks (6.14%). This limited success relative to past research may result from differences in study designs. For example, dependent variables in previous work include trait ratings towards members of racial groups (Experiment 1, Sechrist & Stangor 2001; Stangor et al., 2001), and those suffering from obesity (Puhl et al., 2005), seating distance from an African American confederate (Experiment 1, Sechrist & Stangor 2001), and a lexical decision task (Experiment 2, Sechrist & Stangor 2001). These past dependent variables were calculated one week after social consensus

feedback was administered to participants (as opposed to immediately in the studies presented in this chapter). While Experiments 4-6 may simply reflect overt agreement with one's social group despite no corresponding internal change in attitudes, the delayed assessment of dependent variables in these earlier studies suggests that learning about the beliefs of one's peer group may result in real changes to cognitive representations of certain social groups (Stangor et al., 2001). However, further research is required to establish whether this time difference in presenting the social feedback is a critical factor in attaining stronger stereotype reduction effects and could also illuminate the potential for long-term influences of this social feedback on gender stereotype application.

One further issue worth addressing in future studies of social consensus information relates to the plausibility of the information being presented. Although only a small number of participants indicated suspicion about the feedback information used throughout this chapter (informally to the experimenter when questioned about it after the study), data was not formally collected on this issue (aside from during the pilot study in relation to the RSCF of Experiment 5). Future researchers should take efforts to assess this aspect of norm presentation more thoroughly. Indeed, past research suggests that participants who indicate suspicion about conformity are less, as opposed to more, likely to exhibit conformity effects (Stricker, Messick, & Jackson, 1967); a finding which may potentially account for the relatively small magnitude of the social consensus training effect in this chapter.

Overall, evidence from Experiments 4-6 maintains that the use of social consensus feedback is a somewhat valuable means of stereotype reduction, particularly when combined with accuracy information. However, in line with Stangor and colleagues (2001), it is not proposed that interventionists aim to modify stereotypes outside of the laboratory using false information about the opinions of others. That said, in cases where individuals incorrectly assume that stereotypic beliefs are widely shared or over-estimate the negativity of stereotypes held by fellow group members, it is possible that providing them with accurate consensus information may be sufficient to generate stereotype change.

In conclusion, despite the varying success of all feedback strategies in reducing levels of gender stereotyping in Chapters 2 and 3, both accuracy and response time performance was consistently poorer than to stereotype congruent word pairs. Therefore, Chapter 4 investigated a different approach to overcoming the immediate activation of gender biases; the use of counter-stereotypes as a stereotype reduction strategy.

4. Counter-stereotypic strategies aimed at overcoming immediate activation of gender biases

4. 1 Introduction

While a great deal of literature has focused on the reduction of stereotyping through suppression of bias and avoidance strategies, often with unintended rebound effects (Macrae et al., 1994; Monteith et al., 1998), Chapter 4 explores how strengthening counter-stereotype associations in the cognitive network may work to reduce gender stereotyping.

Stereotype representations are thought to be made up of the strongest or most typical group associations, yet evidence of subtyping (i.e. creating new categories to account for unexpected information) suggests that counter-stereotype information may also be represented therein (Blair et al., 2001). For instance, information about female subtypes (e.g. businesswomen, female athletes) may be activated as part of stereotypical representations about women. Indeed, as discussed in Section 1.6.7, Blair and colleagues argue that as stereotypes and counter-stereotypes are often polar opposites, it is unlikely that they will be represented independently of one another. On the contrary, it is thought that as accessibility of one of these constructs is increased, the other is likely to decrease due to cognitive consistency and efficiency pressures. Ultimately, given the likely inter-dependency of stereotypic and counter-stereotypic information in cognitive representations, the studies in this chapter explore whether increasing the accessibility of counter-stereotypes can result in lower levels of stereotype activation. In fact, past researchers have successfully employed counter-stereotypes to this effect.

Blair et al. (2001) investigated counter-stereotype mental imagery as a means of moderating implicit gender stereotypes. Participants who were asked to imagine a strong woman (counter-stereotype condition) showed lower levels of automatic gender stereotype activation on a gender-related Implicit Association Test, relative to participants who imagined a weak woman (stereotype condition) (Experiment 2). Kawakami and colleagues (2000) also investigated the use of counter-stereotype information in their research aimed at reducing stereotyping towards racial groups and skin heads. They developed a non-stereotypic association training that involved presenting participants with either counter-stereotypic or stereotypic word pairs relating to the category of interest (e.g. race). Participants' task was then to affirm (i.e. say 'yes') and negate (i.e. say 'no') these counter-stereotypic and stereotypic pairings respectively. This training resulted in reduced levels of automatic stereotyping on a variety of measures

(previously described in Section 1.6.7). Combined, these stereotype reduction strategies reveal that directing a participant's attention to subtypes of a category and activating counter-stereotypical information are useful means of reducing stereotypes, which will be incorporated into the stereotype reduction strategies explored in this chapter.

As outlined in Chapter 1 (Section 1.6.4), two potential processes through which counter-stereotypes may lead to reduced stereotyping include (1) the bookkeeping process in which stereotypes are hypothesised to change slowly, through encountering numerous counter-stereotype exemplars of a particular category and (2) the conversion process in which stereotypes are thought to change more rapidly, upon encountering fewer, yet more striking counter-stereotype exemplars than postulated in the book-keeping process (Operario & Fiske, 2004). The experiments in this chapter explore both of these approaches to some extent. In Experiment 7 participants are asked to learn a list of counter-stereotype word pairs that they will subsequently be tested on. However, participants are not explicitly made aware that the pairings are counter-stereotypes. In this way, stereotype change is hypothesised to occur through a type of 'bookkeeping' process with change taking place gradually and subtly as more pairings are learnt. However, Experiment 8 employs a more striking counter-stereotype strategy in which participants are presented with pictures of men and women working in counter-stereotypical roles. These gender-salient pictures are hypothesised to bring about stereotype change through more direct and immediate conversion processes.

Finally, a concern when using counter-stereotype exemplars in research aimed at stereotype reduction also stems from work on subtyping. Essentially, subtyping processes can ensure that the original stereotype be protected and remain unchanged because new categories are formed to account for counter-stereotype information. On the other hand, it is also possible that stereotypes could be weakened and reduced with sufficient category variation and subtyping (Operario & Fiske, 2004). In this chapter, it is hypothesised that the latter of these two routes will be taken, as it is proposed that new and counter-stereotypical information should be incorporated into occupational stereotypes in order to change them. For example, repeated exposure to women working in typically male dominated roles *should* gradually alter perception biases towards these occupations. With this is mind, Experiments 7 and 8 both present participants with a relatively large number (24) of gender counter-stereotype exemplars in a bid to overcome occupational stereotypes.

4.2 Experiment 7: Counter-stereotype association learning

4.2.1 Introduction

The aim of Experiment 7 was to examine the role of counter-stereotype information as a potential moderator of spontaneous gender biases on stereotype application. Specifically, an 'Association Learning' (AL) task was employed in which participants learned a list of 24 counter-stereotypic word pairs, comprised of a stereotype-biased occupation and a male or female proper name (e.g. Farmer/Sarah or Beautician/David). Participants were given a set amount of time to learn this list followed by a short recall test to ensure that learning had occurred. Either side of this AL task, participants completed a block of judgement trials based on the paradigm of Oakhill et al. (2005) (previously described in Section 2.2.2). The aim of this AL training strategy was to strengthen counter-stereotypic associations about men and women in certain occupational roles so that these associations may then become more salient and potentially challenge the dominance of the stereotypic representation. By helping participants acquire the idea that males can occupy jobs typically held by females and females can occupy jobs typically held by males, it was hypothesised that participants may then become more flexible, mentally, when next confronted with stereotype incongruent pairings in the judgement task, and potentially display reduced levels of stereotype application. More specifically, it was anticipated that participants would exhibit higher levels of accuracy and shorter response times to stereotype incongruent pairings in Block 2 (following the counter-stereotype AL task), relative to Block 1 performance.

4.2.2 Method

Participants

Thirty-eight monolingual, native English speakers (16 male, 22 female) from the student population of the University of Sussex took part in this experiment. Participants' ages ranged from 18-28 (*M*: 20.39; SD: 3.17). They received either £6 or 4 course credits for taking part in the session, which lasted approximately 45 minutes.

Materials

Role nouns

As outlined in Section 2.2.2, 12 male-biased, 12 female-biased and 12 neutrally-rated role nouns were selected for the judgement task and each paired with 6 kinship terms. This resulted in 72 word pairs in each of the stereotype congruent, incongruent and neutral conditions producing a total of 216 critical items. However, unlike Experiments 1-6, these items were now divided into just two blocks of trials (as opposed to three or four in the previous studies), and separated by a training phase that did not involve the judgement trials. Each stereotyped role noun was therefore presented with 3 kinship terms in Block 1 and the remaining 3 in Block 2 (with exact kinship term use balanced across blocks). Overall, 36 stereotype congruent, 36 stereotype incongruent and 36 neutrally-rated pairings were presented in each block. Block presentation was also counter-balanced across participants to ensure that both blocks (as defined by content) appeared an equal number of times in the Block 1 and Block 2 position.

As before, role nouns for the behavioural task were selected from norms, based on the strength of their stereotype bias. However, with the addition of the counter-stereotype AL task, 12 supplementary male and female-biased role nouns were required in this study (neutral terms were not used in the AL task). All items were again selected from norms presented by Gabriel et al. (2008) and Kennison and Trofe (2003), and resulted in a set of 24 male-biased and 24 female-biased role nouns. These were then alternately divided between the judgement task and the AL task on the basis of their rating bias. In this way, it was ensured that both sections of the experiment contained items with an equally strong stereotype bias. A list of the role nouns used in the judgement task and the role nouns used in the AL task can both be found in Appendix 10.

Filler items

Filler trials were again constructed as outlined in Section 2.2.2, with the exception that 311 filler trials were now presented as opposed to 240 in previous experiments. The number of filler items used in this experiment was increased in an effort to further disguise the nature of the study, particularly by increasing the number of 'no' responses relative to 'yes' responses[60]. This resulted in 155 fillers in Block 1 and 156 in Block 2. In each block, 40 definitionally matching pairs were presented (20 male and 20 female) while 115/116 definitionally mismatching pairs were presented (57-59 male and female in each block)[61]. In the definitional

[60] As a reminder, all critical trials and all definitionally matching word pairs require a 'yes' response whereas only the definitionally mismatching pairs require a 'no' response.
[61] Therefore, Experiments 1-6 included 60 definitionally matching word pairs and 180 definitionally mismatching pairs versus 80 and 231 respectively in the current study.

matching condition, different role nouns were used across both blocks while in the definitional mismatching condition, the same role nouns were repeated across both blocks. A full list of the filler role nouns used can be found in Appendix 11.

Association Learning Task

In the AL task, participants were presented with a list of 24 counter-stereotypic word pairs, comprised of a male or female proper noun and a role noun of opposing gender bias.

E.g. Sarah Farmer

David Beautician etc.

It was anticipated that the majority of participants in the sample would be between the ages of 18-23 and so a search was conducted for popular British first names from around the time they would have been born. Statistics were found on the 10 most popular British first names for the years 1994 and 1984 (Merry, 1995)[62]. For each sex, the top ten names for the year 1994 were selected along with the most popular two from 1984[63]. This resulted in 12 male and 12 female first names, a full list of which is provided in Appendix 10. It was deemed that the participants would be highly familiar with common first names from this period through exposure over their lifetimes.

These first names were then individually paired with role nouns of a counter-stereotypic gender bias (also displayed in Appendix 10). Word pairs were formed based on the respective popularity and stereotype-bias ratings of the terms. For instance, the most popular male or female name was paired with the role noun of weakest stereotype bias. This procedure was followed so as to ensure that no one pairing would be considered more memorable than any other.

Participants were given 10 minutes to learn the word pairs before proceeding to a test phase in which their knowledge of the items was examined. Participants also had the option of proceeding to the test phase before this 10 minute time limit was reached, if they so wished. The test phase involved presenting participants with 24 pairings, 12 of which had appeared in the list they had learned and 12 of which had not, that they were asked to either accept (if previously learned) or reject (if new). Items that were incorrectly identified were reloaded into

[62]Names from 1984 were chosen so as to introduce some variance into the names selected while still being highly familiar to participants.

[63]However, the male name '*Jordan'* from 1994 was replaced (as it is also used as a female name) with a third name from 1984, '*Michael*'.

the program and reappeared later until all items were correctly identified. Therefore, if no errors were made, a participant was presented with 24 pairings. However, the higher number of pairings presented, the greater the number of mistakes that were made. The final number of pairings presented to a participant can be taken as a measure of how difficult the participant found it to learn the list, resulting in his/her 'learning score'.

Design

The design of the experiment was as described in Section 2.2.2 However, in this case the stereotype reduction manipulation was the use of an AL task between Blocks 1 and 2 of the judgement task.

Procedure

The procedure for this experiment was as outlined in Section 2.2.2, but with instructions changed to describe the AL task and accompanying test phase. Furthermore, 4 individual difference measures were administered to participants following the judgement task (as opposed to before the judgement task in previous studies). Again this experimental modification was made so as to ensure participants did not deduce the nature of the behavioural task through completion of these measures. The individual difference measures (the Implicit Association Test (IAT), the Bem Sex Role Inventory (BSRI), the Ambivalent Sexism Inventory (ASI), and the Modern Sexism Scale (MSS)) will be discussed further in Chapter 5.

4.2.3 Results

Data screening

The analyses reported below excluded data for word pairs that contained the neutral role noun 'adolescent', over concerns that low accuracy to word pairs involving this term (80.6 % in Block 1, 77% in Block 2) stemmed from considerations of age appropriateness over gender. In total, 1.32% of the data was removed for this reason.

Analysis

Response times below 150ms and above 4,000ms were excluded from analysis (representing 2.22 % of the total data) along with times for all errors of judgement (representing a further 8.84%) totalling a loss of 11.06% of the data. These data points were replaced, and analyses

conducted, as outlined in Section 2.2.3, but with the Block factor updated to contain two levels (Block 1 and Block 2) as opposed to three.

Learning Score

A learning score for each participant was calculated, which reflects performance on the AL test phase. The lowest possible score of 24 indicates optimal performance while higher scores reflect weaker performance (as items that were incorrectly identified were reloaded into the programme). The average learning score was 28.76 (*SD* = 5.81, range of 24-43) with 13 out of 38 participants achieving the optimal score of 24. The behavioural analyses were conducted with the data of all participants (despite some participants doing relatively poorly on the test phase) as the test phase required participants to reach a criterion of correctly responding to each pairing before proceeding. Indeed there was no significant correlation between accuracy of stereotype incongruent pairings in Block 2 and participants' learning scores (r = -.270, p = .101), suggesting that participants dissociated the two tasks.

Accuracy

The analysis revealed a marginal effect of Stereotype in the by-participants analysis, but a significant main effect in the by-items analysis, F_1 (1.48, 53.30) = 2.82, p = .083; F_2 (2, 32) = 4.45, p = .020, with higher accuracy to word pairs that contained a neutral role term (*M* = 94.6%), than those that contained male (*M* = 92.4%) or female stereotype-biased terms (*M* = 93.1%).

As anticipated, a highly significant main effect of Congruency[64] was found, F_1 (1.22, 43.73) = 17.99, p < .001; F_2 (2, 32) = 18.95, p < .001, with significantly lower accuracy to stereotype incongruent word pairs (*M* = 89.4%), than to congruent (*M* = 96.5%) and neutral (*M* = 94.6%) pairings.

Contrary to expectations, no main effect of Block was found in the by-participants analysis, F_1 (1, 36) = 0.64, p = .431, with accuracy increasing just 1% across blocks (92.9% to 93.9%). However, this increase was significant in the by-items analysis, F_2 (1, 32) = 9.38, p = .004. Similarly, a significant interaction of Congruency by Block was *not* found in the by-participants analysis yet this interaction was significant in the by-items analysis, F_1 (1.27, 45.83) = 1.03, p = .333; F_2 (2, 32) = 3.64, p = .038. Despite the lack of significant by-participant interaction, the

[64] i.e. an interaction of Stereotype bias by Kinship term gender.

pattern of responding across blocks in each of the critical conditions is displayed in Figure 4.1 below for further examination.

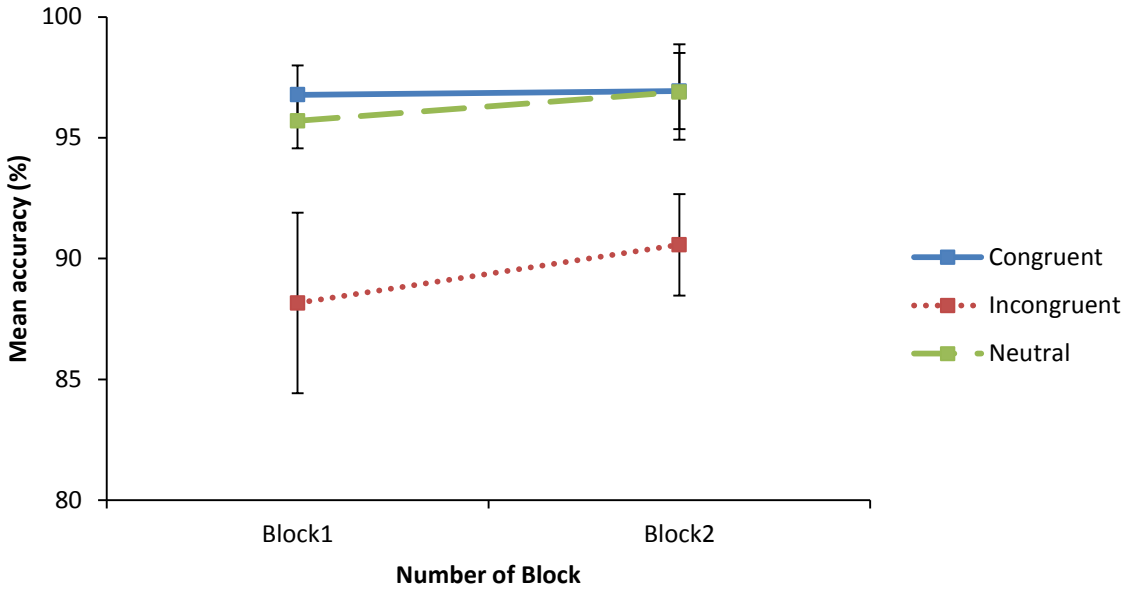


Figure 4.1. Experiment 7: Mean percentages of correct judgements to critical word pairs in Block 1 and Block 2. The vertical axis begins at 80% while error bars indicate the 95% confidence intervals.

A one-tailed, paired-samples *t*-test was conducted to investigate whether the above (2.41%) rise in accuracy across blocks to incongruent word pairs was a significant increase. Again, while no significant effect was found in the by-participants analysis, t_1 (37) = 1.04, p = .154, a significant difference was revealed in the by-items analysis, t_2 (23) = 3.13, p = .003, dz = .64. Overall, this relatively small increase in accuracy across blocks indicates that counter-stereotype AL task is not an efficient means of reducing levels of stereotype application on this judgement task. However, it should be noted that pre-training accuracy in Block 1 of this study (88.1%) is much higher than in previous experiments, a fact which has important implications for the potential of increasing accuracy across blocks. This issue will be further discussed in the Discussion (Section 4.2.4).

Finally, two effects involving Participant Gender also emerged in the by-items analysis. Firstly, there was a main effect of Participant Gender, F_2 (1, 32) = 53.79, p < .001, with female participants typically achieving higher levels of accuracy than male participants (95.9% vs. 91.6% respectively). Secondly, there was a significant interaction of Participant Gender by Block, F_2 (1, 32) = 4.89, p = .034, with females improving a greater extent across blocks (2.4%) than male participants (0.3%).

Response times

A main effect of Stereotype was found in the by-participants analysis only, F_1 (1.73, 63.67) = 8.45, p = .001; F_2 (2, 32) = 1.22, p = .309**,** with predictably faster response times to word pairs that contained a neutral role term (M = 788ms), than those that contained male-biased (M = 840ms) or female-biased terms (M = 817ms). These RT patterns were similar in the by-items analysis yet this effect was not significant, likely due to the fact that Stereotype bias was a within-subject but between-items factor[65].

A main effect of Congruency[66] was also revealed, F_1 (1.56, 56.20) = 12.47, p < .001; F_2 (2, 32) = 10.98, p < .001, with similarly fast response times to neutral (M = 778ms), and stereotype congruent word pairs (M = 780ms). However, response times to incongruent pairings were much slower (M = 857ms).

As with the accuracy data, there was no significant effect of Block in the by-participants analysis, F_1 (1, 36) = 0.46, p = .500, yet a marginal effect was observed in the by-items analysis, F_2 (1, 32) = 4.11, p = .051. However, RTs were found to actually *increase* 25ms across blocks in the by-items analysis (versus 20ms in the by-participants analysis) as opposed to decrease after the AL task.

There was no evidence of a significant Block by Congruency interaction in either analysis, F_1 (2, 72) = 0.27, p = .764; F_2 (2, 32) = 0.56, p = .580. However, to examine patterns of RT responding across blocks in greater detail, a graph of RT responding to critical word pairs (from the by-participants analysis) is provided in Figure 4.2 below.

[65] Mean RTs in the by-items analysis (1) neutral role nouns = 783ms (2) male-biased role nouns = 825ms (3) female-biased role nouns = 815ms.

[66] i.e. an interaction of Stereotype bias by Kinship term gender.

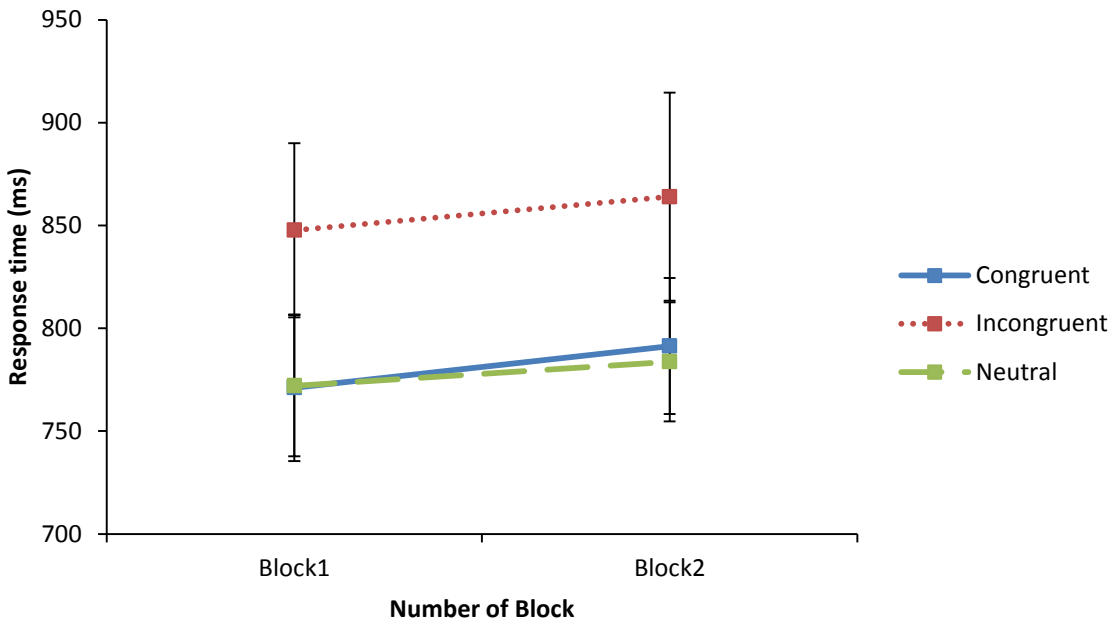


Figure 4.2. Experiment 7: Mean response times (in milliseconds) of correct judgements to critical word pairs across Block 1 and Block 2. The vertical axis begins at 700ms while error bars indicate the 95% confidence intervals.

On average, RTs increase slightly across blocks in all conditions as opposed to decrease. This pattern of data clearly suggests that the use of the counter-stereotypic AL task, did not successfully lead to faster responding to stereotype incongruent pairings on the judgement task. While reasons for this increase in RTs across blocks are unknown, it is likely that modifications made to the judgement task in this study (longer blocks of judgement trials and a cognitively intense training strategy) may have led to fatigue effects. These issues will be considered further in the Discussion (Section 4.2.4).

Participant Gender was also found to influence RT responding in this experiment. Firstly, there was a main effect of Participant Gender in the by-items data only, F_2 (1, 32) = 69.07, $p < .001$, with female participants faster at responding than males (760ms vs. 855ms respectively). There was also an interaction of Participant Gender by Block in the by-items data, F_2 (1, 32) = 7.30, $p = .011$, as male participants became 65ms slower to respond across blocks while female participants became 15ms faster at responding across blocks.

A significant interaction of Block by Congruency by Participant Gender was also found, F_1 (2, 72) = 5.09, $p = .009$; F_2 (2, 32) = 5.87, $p = .007$, with a graph of this interaction shown in Figure 4.3 below. This graph reveals an interesting difference in the pattern of responding to stereotype incongruent pairings for the male and female participants. Firstly, female

participants showed evidence of faster responding following the learning task, although this improvement was quite small (47ms). However, male participants displayed slower reaction times in Block 2 than Block 1 (109ms difference). These results reveal that the AL training did have some beneficial effects on the RTs of female participants, but did not for male participants.

As regards the stereotype congruent and neutrally rated pairings, the female participants again outperformed the males with faster response times across blocks. However, for both sexes, response times in these two conditions did not change much after the stereotype reduction training (just tending to increase slightly as opposed to decrease). On the whole, this pattern of responding again suggests that the AL task did not successfully lead to improved RT performance on this judgement task.

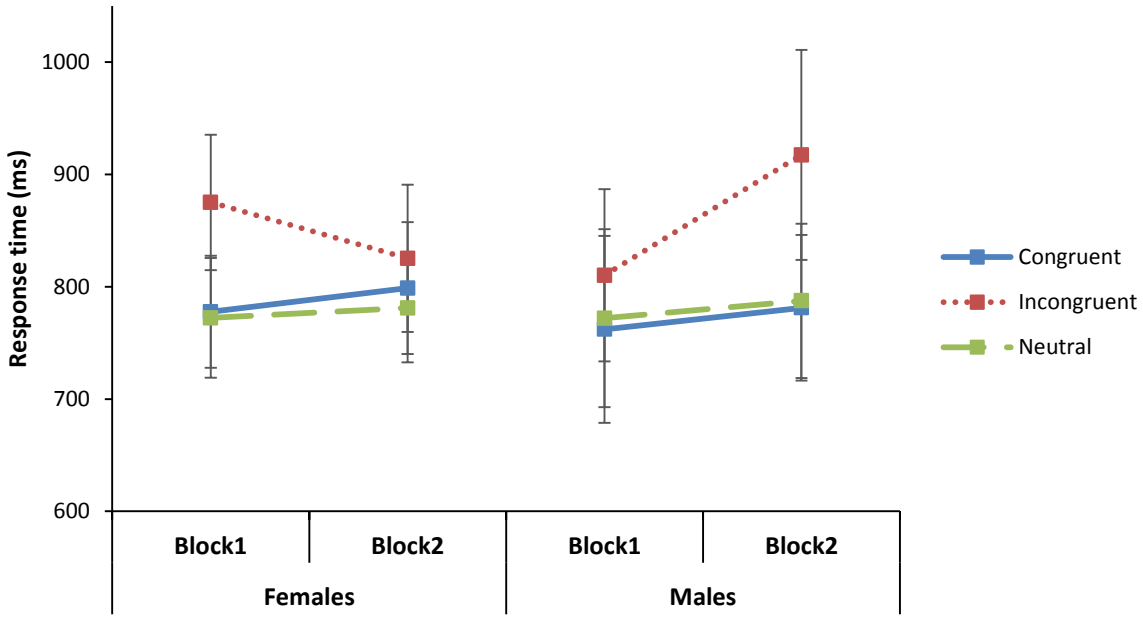


Figure 4.3. Experiment 7: Mean response times (in milliseconds) of correct judgements to critical word pairs across Block 1 and Block 2, for both female and male participants. Error bars indicate the 95% confidence intervals.

Finally, there was a significant interaction of Participant Gender with Kinship term gender, F_1 (1, 36) = 5.89, p = .020; F_2 (1, 32) = 17.95, p < .001. Female participants were found to respond faster to female kinship terms than male kinship terms (788ms vs. 820ms respectively), while male participants responded faster to male kinship terms than female (783ms vs. 826ms respectively).

Fillers - Accuracy

It was anticipated that accuracy of the filler trials would be high as these were gender unambiguous word pairs that should be responded to with relative ease. This was indeed found to be the case, but with slightly higher accuracy found to definitionally matching word pairs across blocks (91.5%) compared to definitionally mismatching word pairs (88.6%).

With these definitionally mismatching word pairs, lower accuracy was found to trials containing male-biased role nouns (*M* = 81.6%) as opposed to female-biased role nouns (*M* = 95.48%). As in previous chapters, this pattern was due to the generic interpretation of some definitionally masculine terms e.g. host, steward.

Fillers - Response times

Unlike the accuracy data, little difference was found between RTs to definitionally matching pairs (*M* = 916ms) and definitionally mismatching pairs (*M* = 922ms).

However, faster response times to female word pairs over male in both the definitionally matching (898ms vs. 934ms respectively) and mismatching cases (894ms vs. 950ms respectively) were observed. Again, this pattern is consistent with that seen in the accuracy data and reflects participants' deliberation over terms that are masculine by definition but are often used generically.

4.2.4 Discussion

Experiment 7 sought to reduce levels of stereotype application on a judgement task through the use of a counter-stereotype, association learning paradigm. However, a small yet significant increase in accuracy scores was found in the by-items data only, while no significant decrease in RTs across the experimental blocks was found. On the whole it can be concluded that the AL training was not an effective means of reducing gender stereotype application on the judgement task. That said, when comparing the results of this study with the previous experiments (1-6), some interesting findings emerge.

Firstly, it is worth noting that accuracy in Block 1 of these decision making trials (88.16%) was much higher than in all previous experiments, with average accuracy in Block 1 of Experiments 1-6 almost 8% lower (*M* = 80.25%, *SD* = 3.98%). It is clear that accuracy to stereotype incongruent word pairs had less scope for improvement across blocks in Experiment 7. Indeed,

despite improving only 2.41% from Block 1 to Block 2, final accuracy to these critical pairings (90.57%) was actually higher than final accuracy scores in Experiment 4 (social consensus feedback: 82.64%) and Experiment 6 (combined social consensus and accuracy feedback: 88.42%), and (unsurprisingly) the two control Experiments (2 and 5).

Similarly, with the response time data, average RTs to incongruent pairings in Block 1 of this study were much faster (848ms) than equivalent RTs in Block 1 of Experiments 1-6 (*M* = 1034ms, *SD* = 90ms). Again, this means that RTs had less scope for improvement across blocks in Experiment 7 and, as mentioned earlier, they actually increased (16ms) following the AL task as opposed to decreased. However, final RTs to the critical pairings at the end of this study were in fact slower than final RTs of all other studies aside from equalling Experiment 2 (a control).

Given the above findings, the question arises as to why Block 1 performance in this experiment differed so greatly from Block 1 performance of previous experiments. As there were no obvious differences between current and previous participant samples, the reasons for this remain unclear yet the present study did differ from the previous studies in two key respects:

(1) Experiment 7 contained more judgement trials than previous experiments, spread across two blocks as opposed to three i.e. before the counter-stereotypic training, participants in this study completed 223 judgement trials compared with 152 in the previous experiments. Although the number of experimental trials was increased in an attempt to further mask the experimental manipulation, it is possible that the modified design actually made the current participants *more* aware of the nature of the task and allowed them extra time to develop strategies that improved accuracy responding. Similarly, the increased number of trials in Block 1 of this study allowed participants to habituate to the judgement task and speed up their responding as the block progressed, thus providing a reasonable explanation for the unusually fast RT averages in Block 1. That said, Block 2 RTs were notably slow. It is possible that, when combined with the cognitively-intense AL training, the large number of trials in Block 1 may have led to fatigue effects in the participants. However, the above theories as to the impact of the increased number of judgement trials in this experiment are speculative and cannot be confirmed at this point.

(2) The second way in which Experiment 7 differed to previous work is that participants completed a battery of questionnaires (and an additional implicit association test) *after* the behavioural task in this study, as opposed to before. However, it seems unlikely that this procedural modification would have impacted responding on the behavioural task. Indeed, if

anything, it would have been hypothesised that Block 1 accuracy in Experiment 7 would be *lower* than in previous studies where participants may have benefitted from completing the questionnaires; for instance, by becoming more alert to the exact nature of the experimental task and consequently being more cautious in their responding or developing strategies to aid successful responding.

Overall, the results of this study are somewhat ambiguous. While the AL training task was not found to greatly reduce levels of stereotyping across blocks, final accuracy scores were higher than average (relative to the previous studies) while final RT scores were slower than average (relative to the previous studies). Further studies are required to clarify the underlying cause(s) of these results.

One final concern with the AL task employed in this study was that it was too subtle a manipulation to induce behavioural change. At no stage was it pointed out to participants that the word pairs they were learning were counter-stereotypes, or that this task was intended to aid performance on the behavioural task. In contrast, it was up to participants to recognise these pairings as counter-stereotypes and use this information to aid performance on the critical pairings. However, there is no direct evidence that participants picked up on this counter-stereotypical information and indeed participants may have entirely dissociated the two parts of the experimental session (i.e. the learning task and the judgement task). Indeed, there is some evidence that this occurred as some participants scored very highly on the learning task yet showed very low accuracy to incongruent pairings on the judgement task.

In contrast to the above AL training, many researchers theorise that changing knowledge structures must take place through experience and advocate the role of stereotype awareness in overcoming stereotypes (see Bargh, 1992, 1994). For this reason, it was decided to once again involve the use of counter-stereotypes in a stereotype reduction strategy, but using a more striking training paradigm than in Experiment 7. This next stereotype reduction approach is outlined in Experiments 8 and 9 below.

4.3 Experiments 8 & 9: Counter-stereotypic versus Stereotypic pictures as a strategy to overcome immediate activation of gender stereotypes.

Experiments 8 and 9 investigated exposure to pictures of people working in stereotypical versus non-stereotypical occupations as a means of gender-stereotype moderation. To do this,

24 pairs of pictures of men and women working in a variety of occupational roles were first required. Both copyright and non-copyright pictures were used, collected through a web search and from a picture database (www.masterfile.com) respectively. In the latter case, the pictures were differentially priced as outlined at http://www.masterfile.com/info/products/licensing-explained.html. Each picture pair depicted a man and a women working in the same role i.e. one picture was counter-stereotypical while the other was stereotypical.

As the pictures were sourced from different web locations, they varied in quality, orientation and size. Therefore, before presenting them to participants, the pictures were manipulated (cropped and re-sized) so as to be as clear as possible. When presented to participants they ranged in size from 228-459 pixels wide x 233-406 pixels high, (this large range in pixel numbers was due to the varying landscape versus portrait orientation of the pictures).

Before using these pictures in the behavioural studies, a pilot study was first conducted so as to evaluate (a) the similarity of the male and female versions of the pictures and (b) how realistic the pictures looked.

4.3.1 Pilot study 2

Nine students (all female) from the University of Sussex took part in this pilot study. It was the second of two pilot studies that the participants completed, one following the other[67]. Each study lasted 5-10 minutes and participants received 2 course credits for their participation.

In the picture study, each of the 24 picture pairs was presented to participants as 'pictures of men and women working in the same roles'. The participants' first task was to rate these on "how similar they were (ignoring gender and thinking about features such as the race, age, pose, facial expression of the people and the background)". Ratings were made on a scale ranging from 1 (very similar) to 6 (very dissimilar).

Next, participants judged how realistic they found the pictures to be – again ignoring gender and thinking about features such as the race, age, pose, facial expression of the people and the background. Each of the 24 picture pairs was again presented to participants but with all 48 pictures now rated individually on this realism measure, again on a 6-point scale ranging from 1 (very realistic) to 6 (very unrealistic).

[67]The first pilot study was conducted for Experiment 5 as described in Section 3.3.2. However, one participant completed the picture pilot study only.

Results and Conclusion

Similarity

The mean similarity rating across picture pairs was 2.20 (*SD* = 0.61), thus falling between the points of moderately similar (2) and mildly similar (3). Two of the pairings received particularly low ratings of between 3 and 4 (i.e. mildly similar and mildly dissimilar); these were Electrician (*M* = 3.22, *SD* = 1.72) and Solider (*M* = 3.33, *SD* = 0.87). However, when participants were informally questioned on their ratings, they revealed that a lack of clearly dissimilar pictures in the task forced them to become stricter in their judgements than was anticipated by the experimenter. For this reason, and combined with the fact that the ratings still fell on the side of more similar than dissimilar, it was deemed that all pictures were suitable for use in the main experiments (pending realism scores).

Realism

The mean rating of how realistic a picture looked was 1.94 (*SD* = 0.55), thus falling between the points of very realistic (1) and moderately realistic (2). Indeed, all pictures were rated as being more realistic than unrealistic, although the picture of the male fortune teller was closest to this boundary (*M* = 3.22, *SD* = 1.56). Again, in light of the fact that no clearly unrealistic pictures had been included as a baseline reference for participants (thus potentially leading them to be stricter in their judgements), it was decided not to replace this picture.

Ultimately, all pictures were kept for the experimental task as none were rated as being more unrealistic than realistic or more dissimilar than similar. These 48 pictures are presented in Appendix 12.

4.4 Experiment 8: Counter-stereotypic pictures as a strategy to overcome immediate activation of gender stereotypes.

4.4.1 Introduction

As mentioned in Chapter 1 (Section 1.9), Macrae and Bodenhausen (2000) identified that there is an over-reliance on verbal category labels in research investigating the process of category activation. They caution that this is problematic as in reality people are complex stimuli that can be classified by perceivers along multiple dimensions. Consequently, it cannot be assumed

that the processing of verbal labels equates to the processes involved in person perception. In an attempt to address this issue of over-reliance on verbal labels in stereotype research, Experiment 8 employed the use of pictures of real people working in a variety of counter-stereotypic social roles as part of the experimental design.

More specifically, as in previous experiments, participants were first given a block of judgement trials in which they had to quickly decide whether or not two terms could be used to refer to one person. Next, participants completed a counter-stereotype picture task before doing one final block of judgement trials. However, unlike Experiment 7, these two blocks did not contain more trials than in Experiments 1-6. Instead participants completed just two of the three (152 word pair) blocks used in Experiments 1-6 so as to make the current experiment more directly comparable with the previous studies.

In the picture task participants were presented with 24 pictures of people working in counter-stereotypical roles. The participants' task was to answer a set of four questions on each picture relating to the character's supposed earnings, leisure activities, job satisfaction and personal life. It was theorised that attending to such information would lead to deeper processing of both the character presented and counter-stereotypical job this person was depicted as holding, thus activating the participants' world knowledge that women can do jobs typically held by men and men can do jobs typically held by women. In this way exposure to the counter-stereotype pictures should encourage participants to accept stereotype incongruent word pairs as possible in the subsequent judgement task.

It was hypothesised that participants would initially respond more slowly and less accurately to trials of stereotype incongruent word pairs (e.g. *nurse/father*) than to stereotype congruent word pairs (e.g. *nurse/mother*) in Block 1. However, following the picture training, it was hypothesised that the processing cost associated with the stereotype incongruent condition in Block 1 would be attenuated and lead to higher accuracy and faster reaction times to the critical trials in Block 2.

4.4.2 Method

Participants

The participants were 30 monolingual, native English speakers (14 male, 16 female) from the student population of the University of Sussex. Participants' ages ranged from 18 to 37 years

(*M*: 20.27; *SD*: 4.12). They received either £6 or 4 course credits for taking part in the session, which lasted approximately 45-60 minutes.

Materials

Word pairs for the judgement task

The word pairs used were identical to those used in Experiments 1-6. However, as two blocks of trials were now administered (instead of three), each participant saw 304 word pairs as opposed to 456. Nevertheless, use of the three original blocks was counter-balanced so that the original 456 pairs appeared an equal number of times across participants. All further word pair presentation details followed the protocol outlined in Section 2.2.2.

Picture task

In this picture task, 24 pictures depicted either a man or a woman working or situated in a counter-stereotypical job environment. Half of the pictures depicted people working in roles that were also mentioned in the judgement task and half depicted 'new' role terms that the participants had not yet been exposed to (6 male and 6 female stereotypical terms in each case)[68].

When displayed on-screen, the pictures were accompanied by a sentence. This sentence always took the same format – an introduction of the character in the picture, followed by a statement of what their job was e.g. *This is Rebecca. She is a bricklayer* or *This is Christopher. He is a make-up artist*. The first names presented were identical to those used in Experiment 7 i.e. they were a selection of popular baby names from the time that most of the participants were born, and that the participants were likely to be highly familiar with. Upon presentation of a picture and the accompanying sentence, participants were required to answer (the same) 4 questions on each picture relating to the characters' probable salary, leisure activities, job satisfaction and lifestyle in a booklet provided. An example booklet page is provided in Appendix 13. Note that, three different experimental scripts were created with the pictures presented in a different, yet fixed, order in each. Three response booklets could then be prepared that matched the picture presentation order of the various scripts.

[68] It was hypothesised that the role terms that appeared in the picture booklet would achieve a higher level of stereotype incongruent accuracy and lower response times in Block 3 of the judgement task than those that did not appear in the picture booklet (as the pictures explicitly depicted a person of counter-stereotypical gender fulfilling the role). However, it was found that accuracy to both sets of terms was identical (at 88%) while RTs were actually slower for the role nouns that had previously appeared in the booklet (*M* = 731ms) compared to those that didn't (*M* = 686ms), although this difference was not significant, $t(22) = 1.34$, $p = .193$.

Design & Procedure

The experimental design was identical to that described in Section 2.2.2 but with the stereotype reduction training now based on counter-stereotypical pictures. Instructions were updated so as to explain the procedure for the picture task and participants were provided with an appropriate booklet. Participants were also administered one questionnaire (the ASI) to complete at the end of the experiment; the results of which will be discussed further in Chapter 5. Finally, students were thanked for their time and fully debriefed as to the aims of the study.

4.4.3 Results

Data screening

As in previous studies, the analyses reported below excluded data for word pairs that contained the neutral term 'adolescent', due to concerns that low accuracy to word pairs involving this term (45% in Block 1, 67% overall) stemmed from considerations of age appropriateness over gender. In total, 1.32% of the data was removed for this reason.

Analysis

Response times below 150ms, and above 4,000ms were excluded from analysis (representing .92% of the total data) along with times for all errors of judgement (representing a further 10.88%), totalling a loss of 11.8% of the data. These data points were replaced, and analyses conducted, as outlined in Section 2.2.3, but with the factor Block updated to contain just two levels (Block 1 and Block 2).

Accuracy

Analysis revealed a main effect of Stereotype, F_1 (1.67, 46.67) = 6.27, p = .006, F_2 (2, 32) = 9.59, p = .001, with higher accuracy to word pairs that contained a neutral role term (M = 94.3%), than those that contained male (M = 88.2%) or female-biased terms (M = 89.2 %).

A main effect of Block was also found, F_1 (1, 28) = 6.90, p = .014; F_2 (1, 32) = 17.73, p < .001, driven by a 3.5% increase in accuracy of critical pairings from Block 1 (88.8%), to Block 2 (92.3%).

As anticipated, there was also a main effect of Congruency[69], F_1 (1.18, 33.08) = 14.76, p < .001; F_2 (2, 32) = 67.55, p < .001, with significantly lower accuracy to stereotype incongruent word pairs (M = 79.80%), than to congruent (M = 97.15%) and neutral (M = 94.35%) pairings.

Importantly, an interaction of Congruency by Block was also found, F_1 (1.39, 38.89) = 8.93, p = .002; F_2 (2, 32) = 22.00, p < .001. This interaction is driven by a substantial 9.87% increase in accuracy towards stereotype incongruent pairings across blocks, as can be seen more clearly in Figure 4.4 below[70]. Accuracy to neutral and stereotype congruent pairings was high from the outset, with little room for improvement across blocks.

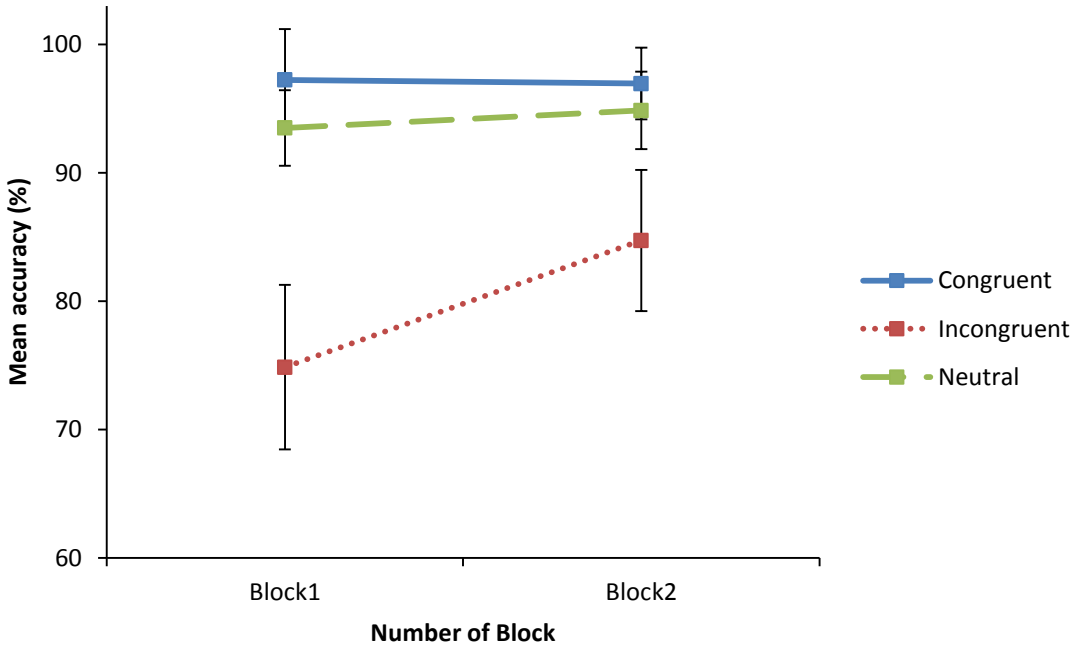


Figure 4.4. Experiment 8: Mean percentages of correct judgements to critical word pairs in Block 1 and Block 2. The vertical axis begins at 60% while error bars indicate the 95% confidence intervals.

The above-mentioned increase in accuracy to stereotype incongruent pairings across blocks was indeed a significant one, t_1 (29) = 3.33, p = .002, dz = .61; t_2 (23) = 5.70, p < .001, dz = 1.16, revealing the efficacy of the counter-stereotypic picture task as a gender stereotype reduction strategy.

[69] i.e. an interaction of Stereotype bias by Kinship term gender.
[70] Note that Block 1 accuracy in this experiment (74.86%) is more similar to Block 1 accuracy of Experiments 1-6 (80.25%) than Experiment 7 where it was particularly high (88.1%).

However, despite this improvement across blocks, accuracy to stereotype incongruent word pairs remained significantly lower than accuracy to stereotype congruent, t_1 (29) = 3.10, p = .004, dz = .57; t_2 (23) = 9.56, p < .001, dz = 1.95, and neutrally-rated word pairs, t_1 (29) = 2.60, p = .015, dz = .47; t_2 (44) =, 6.65, p < .001, d = 2.00, by the end of the experiment. Thus, this picture training did not completely succeed in eradicating the immediate effects of stereotype bias in this judgement task.

Next, an interaction of Participant Gender with Kinship term gender was revealed, F_1 (1, 28) = 5.27, p = .029; F_2 (1, 32) = 5.16, p = .030. Female participants displayed marginally higher accuracy in response to female kinship terms (88.5%) as opposed to male kinship terms (86.6%) while male participants displayed the opposite pattern, showing greater accuracy in response to male kinship terms (94.4%) than female kinship terms (92.8%). Male participants were also found to be more accurate than female participants to kinship terms overall, with an average of 93.6% accuracy, compared to 87.6% accuracy for the female participants (unsurprisingly as reasons for this superior male performance remain unknown as (sex aside) there were no obvious differences between the male and female samples).

Finally, a number of further effects involving Participant Gender emerged in the by-items analysis only. To begin, a main effect of Participant Gender was revealed, F_2 (1, 32) = 104.01, p < .001, with male participants achieving much higher levels of accuracy than female participants overall (93.6% vs. 87.5%). There was also a highly significant interaction of Participant Gender by Congruency, F_2 (2, 32) = 8.08, p = .001. While male participants outperformed females in each of the three congruency conditions, this difference was most apparent in response to stereotype incongruent pairings where male participants achieved an average accuracy score of 85.3% while female participants reached only 75.0%. This pattern of results was not anticipated and there is no ready explanation for why female participants scored much lower than in previous experiments. Finally, there was a Participant Gender by Block interaction, F_2 (1, 32) = 4.92, p = .034, with the accuracy of male participants increasing 2.4% across blocks, compared to 4.8% for female participants (although the females had more scope for improvement from Block 1). That said, the accuracy of females was still lower than that of the males overall.

Response times

A main effect of Stereotype was found in the by-participants analysis, along with a marginally significant effect in the by-items analysis, F_1 (2, 56) = 5.50, p = .007; F_2 (2, 32) = 3.00, p = .064, with predictably faster response times to word pairs that contained a neutral role term (M =

828ms), than those that contained male-biased (*M* = 850ms) or female-biased terms (*M* = 889ms).

A main effect of Block was also revealed, F_1 (1, 28) = 15.50, $p < .001$; F_2 (1, 32) = 97.60, $p < .001$, with response times decreasing 143ms from Block 1 to Block 2 (927ms vs. 784ms respectively).

There was a main effect of Congruency[71], F_1 (1.33, 37.12) = 12.31, $p < .001$; F_2 (2, 32) = 11.62, $p < .001$, with fastest RTs observed in response to stereotype congruent word pairs (*M* = 815ms), followed by neutral (*M* = 829ms) and incongruent pairings respectively (*M* = 920ms).

Importantly, a significant interaction between Block and Congruency also emerged, F_1 (2, 56) = 4.87, $p = .011$; F_2 (2, 32) = 5.27, $p = .010$. As can be seen in Figure 4.5 below, RTs decreased across all conditions from Block 1 to Block 2, with the greatest reduction found in response to stereotype incongruent pairings (225ms). This was found to be a significant improvement across blocks, t_1 (29) = 4.23, $p < .001$, $dz = .77$; t_2 (23) = 7.89, $p < .001$, $dz = 1.61$. Furthermore, by the end of the experiment, there was no significant difference between RTs to stereotype incongruent and stereotype congruent, t_1 (29) = 1.59, $p = .122$; t_2 (23) = 1.65, $p = .112$, or neutral pairings, t_1 (29) = 1.36, $p = .183$; t_2 (44) = 1.41, $p = .167$. Overall, the RT data provides further strong support for the use of counter-stereotypical pictures as an effective stereotype-reduction strategy.

[71] i.e. an interaction of Stereotype bias by Kinship term gender.

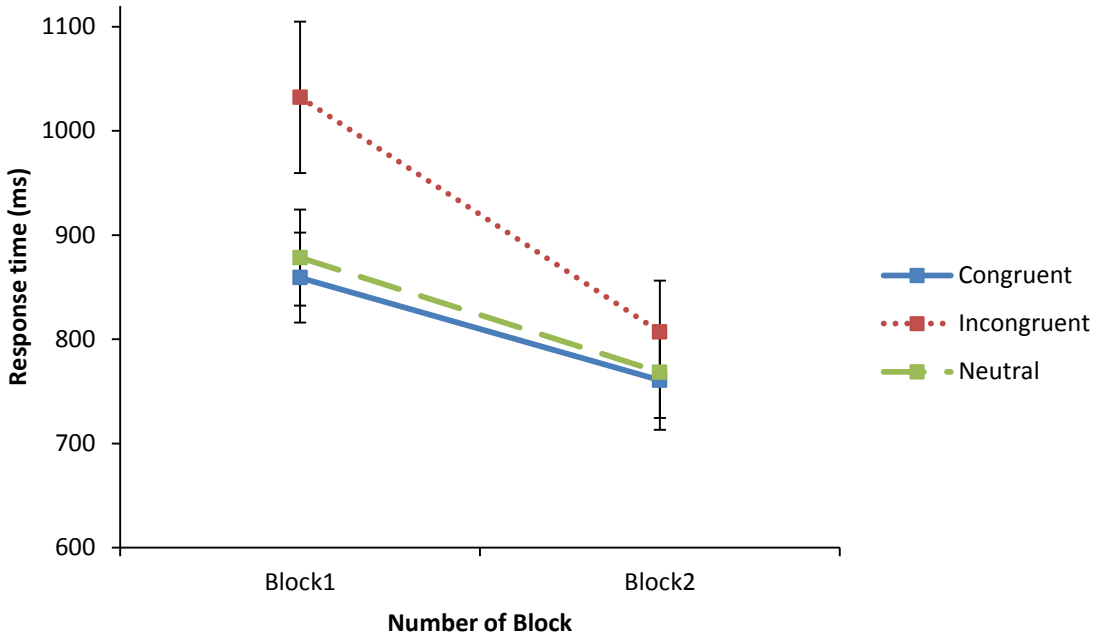


Figure 4.5. Experiment 8: Mean response times (in milliseconds) of correct judgements to critical word pairs across Block 1 and Block 2. Error bars indicate the 95% confidence intervals.

A main effect of Participant Gender was also observed, F_1 (1, 28) = 5.64, p = .025; F_2 (1, 32) = 93.24, p < .001, with male participants typically much slower to respond than female participants (925ms vs. 786ms). Similarly, an interaction of Participant Gender with Kinship term gender was also found, F_1 (1, 28) = 14.23, p = .001; F_2 (1, 32) = 9.09, p = .005. Again, female participants responded faster to female kinship terms over male kinship terms (765ms vs. 807ms respectively), while male participants responded faster to male kinship terms over female kinship terms (892ms vs. 958ms respectively). These means reveal that female participants responded faster than male participants to both kinship terms (786ms vs. 925ms), a finding which may have contributed to the lower accuracy scores achieved by females relative to the males. It seems likely that accuracy performance may have deteriorated for the sake of faster responding.

Fillers - Accuracy

Performance on the fillers was again somewhat variable, with an average of 93.0% accuracy on definitionally matching word pairs versus an average of 87.28% on definitionally mismatching word pairs across the experiment.

As in previous studies, accuracy of responses to definitionally mismatching word pairs was lower to trials involving definitionally male terms (M = 80.61%) than accuracy in response to

trials involving definitionally female terms (*M* = 93.94%). Again, this is thought to result from the generic interpretation of certain terms such as *host* or *steward* which have female-specific counterparts and which could, therefore, be taken as male specific.

Fillers - Response times

Average response times to definitionally matching word pairs were found to be faster than those to definitionally mismatching word pairs (888ms vs. 950ms respectively). Response times were also faster in response to female pairings over male in both the definitionally matching (862ms vs. 914ms respectively) and mismatching cases (910ms vs. 989ms respectively).

These findings support the accuracy data, with longer processing of male mismatching pairs likely to reflect participants' deliberation over terms that are masculine by definition but often used generically in reference to both sexes.

4.4.4 Discussion

Overall, Experiment 8 provides preliminary evidence for the use of counter-stereotypical pictures as an effective strategy for reducing the immediate activation of gender stereotypes when gender-biased role terms are read. Both accuracy and reaction times to stereotype incongruent word pairs significantly improved from Block 1 to Block 2 following the counter-stereotypic picture task. While accuracy remained significantly lower to these incongruent pairs than the stereotype congruent and neutral pairings in Block 2, RTs had improved to a similar level across all three word pair conditions.

It is hypothesised that exposure to the counter-stereotypical pictures triggered participants' world knowledge that, although there is a strong gender bias associated with certain social roles in society, nowadays both men and women can and *do* fulfil these roles. The activation of this knowledge is then thought to have helped participants overcome stereotype application in the second block of judgement trials.

It is worth noting that Block 1 accuracy was somewhat lower than the average Block 1 accuracy found in Experiments 1-6 (3.85% difference), thus leaving more scope for improvement across blocks in the current study. However, the average Block 1 RT to the incongruent pairings was quite similar to the average of Experiments 1-6, (42ms faster in the current study).

Overall, before accepting this picture training as a successful means of stereotype reduction, a control condition against which these results could be compared was required so as to verify that the counter-stereotype manipulation of Experiment 8 was indeed the reason for the improved task performance in Block 2. This led to the conception of Experiment 9 in which gender *stereotypical* pictures replaced the counter-stereotypical pictures in the picture task. This design was chosen so as to maintain as opposed to weaken the gender biases associated with many occupational terms in English.

4.5 Experiment 9: Stereotypical pictures; control condition

4.5.1 Introduction

By providing participants with pictures of people working in gender stereotypical roles, Experiment 9 sought to reinforce participants' world knowledge that women are typically associated with a certain set of roles (e.g. beautician, florist), and men are typically associated with another set (builder, farmer). The experimental design was exactly as outlined in Experiment 8, but with the counter-stereotypical pictures replaced by stereotypical pictures. The rationale for Experiment 9 was that attending to these gender-stereotypical pictures would lead to deeper adherence to gender biases in the judgement task. Therefore, if there was no improvement in response to stereotype incongruent trials from Block 1 to Block 2, it could be confidently assumed that the reduction in stereotype bias across blocks in Experiment 8 was indeed due to the presentation of counter-stereotypical pictures.

As in Experiment 8, it was hypothesised that participants would initially respond more slowly and less accurately to trials of stereotype incongruent word pairs (e.g. *nurse/father*) than to stereotype congruent word pairs (*nurse/mother*) in Block 1. However, unlike Experiment 8, it was hypothesised that the processing cost associated with the stereotype incongruent condition in Block 1 would not be attenuated in Block 2 following presentation of the stereotype congruent pictures.

4.5.2 Method

Participants

The participants were 34 monolingual, native English speaking students (19 female, 15 male) from the University of Sussex. Participants' ages ranged from 18 to 32 years (*M*: 21.23; *SD*: 4.53). They received either £6 or 4 course credits for taking part in the session which lasted approximately 45-60 minutes.

Materials

The same materials and instructions were employed as in Experiment 8, aside from a different set of pictures (and accompanying booklets) used for the picture task. The pictures all depicted men and women working in a stereotypical job environment and were accompanied by a sentence introducing the character and stating their job e.g. *This is Rebecca. She is a make-up artist* or *This is Christopher. He is a bricklayer*. Note that the stereotypic pictures used in this study were previously rated for similarity and realism in the pilot study (Section 4.3.1)

Design & Procedure

The design and procedure were identical to those outlined in Experiment 8 (Section 4.4.2), but with participants answering questions on pictures of people working in stereotypical roles as opposed to counter-stereotypical roles.

4.5.3 Results

Data screening[72]

Response times below 150ms, and above 4,000ms were excluded from analysis (representing 1.77% of the total data) along with times for all errors of judgement (representing a further 12.85%), totalling a loss of 14.61% of the data. These data points were replaced as described in Section 2.2.3, while accuracy and response times of judgements were again analysed as outlined therein, but with the factor Block updated to contain just two levels (Block 1 and Block 2).

Accuracy

A main effect of Stereotype was found although this was just marginally significant in the by-items analysis, F_1 (1.30, 41.61) = 7.81, p = .004; F_2 (2, 33) = 3.10, p = .059, with greater accuracy to neutral role nouns (M = 93.1%), than male-biased (M = 90.7%) or female-biased terms (M = 88.8%).

[72]Note the neutral term 'adolescent' was replaced with the neutral term 'swimmer' in this experiment, therefore data pertaining to all neutral items were included in the analysis.

A main effect of Congruency[73] was also revealed, F_1 (1.03, 33.07) = 12.47, p = .001; F_2 (2, 33) = 55.04, p < .001, with significantly higher accuracy to stereotype congruent (M = 97.0%) and neutral (M = 93.1%) word pairs, than to stereotype incongruent pairings (M = 83.3%).

However, no significant effect of Block was found, F_1 (1, 32) = 0.89, p = .351; F_2 (1, 33) = 0.67, p = .417, with accuracy increasing just 0.5% across the 2 blocks (Block 1 M = 90.6% vs. Block 2 M = 91.1%). Importantly, there was no significant interaction of Congruency by Block, F_1 (2, 64) = 1.05, p = .357; F_2 (2, 33) = 0.74, p = .490, with responding across conditions shown in Figure 4.6 below.

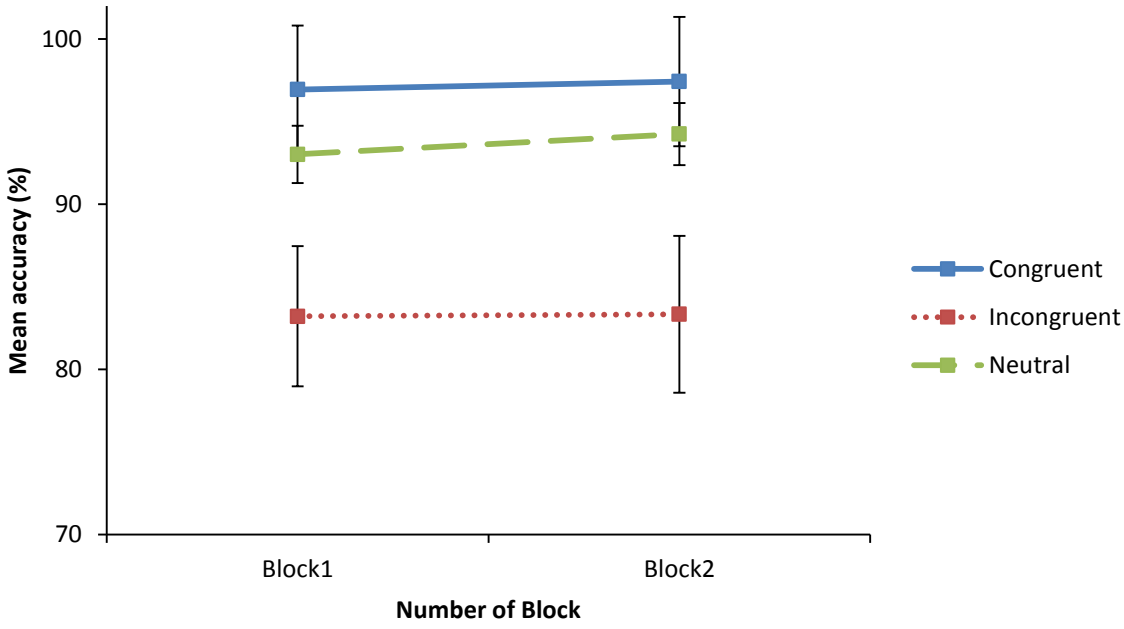


Figure 4.6. Experiment 9: Mean percentages of correct judgements to critical word pairs in Block 1 and Block 2. Error bars indicate the 95% confidence intervals.

Accuracy of stereotype incongruent pairings failed to significantly increase across the blocks, ((+.26%), t_1 (33) = 0.10, p = .918; t_2 (23) = 0.15, p = .880), suggesting that the stereotypical picture training did indeed maintain stereotype biases. However, it is worth nothing that Block 1 accuracy to incongruent pairings in this study was considerably higher than Block 1 accuracy to incongruent pairings in Experiment 8 (83.21% vs. 74.86% respectively), thus leaving less scope for improvement in the current analysis. This issue will be discussed further in Section 4.5.4.

[73] i.e. an interaction of Stereotype bias by Kinship term gender.

Accuracy also remained significantly poorer to stereotype incongruent pairings than to neutral (t_1 (33) = 3.718, p = .001, dz = .64; t_2 (23) = 6.70, p < .001, dz = 1.37) and stereotype congruent pairings (t_1 (33) = 3.32, p = .002, dz = .57; t_2 (23) = 8.64, p < .001, dz = 1.76) by the end of the experiment.

Finally, two effects involving Participant Gender were found in the by-items analysis only. There was a main effect of Participant Gender, F_2 (1, 33) = 165.93, p < .001, with female participants achieving much higher levels of accuracy than male participants overall (94.9% vs. 86.5%). There was also a Participant Gender by Congruency interaction, F_2 (2, 33) = 22.14, p < .001, with female participants outperforming males in each of the congruency conditions, but particularly in response to incongruent pairings (89.15% vs. 75.85% respectively). In contrast to Experiment 8, it was now females who outperformed males in accuracy performance. The reason(s) for this contrasting performance between both sexes remain(s) unclear as again there were no obvious differences between the two samples.

Response times

A significant main effect of Congruency[74] was found, F_1 (2, 64) = 15.18, p < .001; F_2 (2, 33) = 7.22, p = .003, with fastest response times to stereotype congruent word pairs (M = 817ms), followed by neutral (M = 862ms) and incongruent pairings respectively (M = 920ms).

A main effect of Block was also observed, F_1 (1, 32) = 14.56, p = .001; F_2 (1, 33) = 130.93, p < .001, with average response times decreasing 128ms from Block 1 to Block 2.

As with the accuracy data, there was no evidence of a significant Congruency by Block interaction, F_1 (1.82, 58.12) = 0.38, p = .663; F_2 (2, 33) = 0.01, p = .988. As can be seen in Figure 4.7 below, RTs were found to decrease to a similar extent across blocks in all three congruency conditions. While these improvements were each statistically significant (p < .03), they are taken as evidence for task habituation, as participants got progressively faster at responding to all critical word pairs as the task progressed, yet there was no equivalent increase in accuracy performance across critical trials.

A significant difference between RTs to stereotype incongruent and congruent pairs still remained at the end of the experiment, t_1 (33) = 2.90, p = .007, dz = .50; t_2 (23) = 2.82, p = .010, dz = .57, and also between stereotype incongruent and neutral pairings, t_1 (33) = 2.15, p = .039, dz = .37; t_2 (23) = 2.13, p =.044, dz = .43.

[74] i.e. an interaction of Stereotype bias by Kinship term gender.

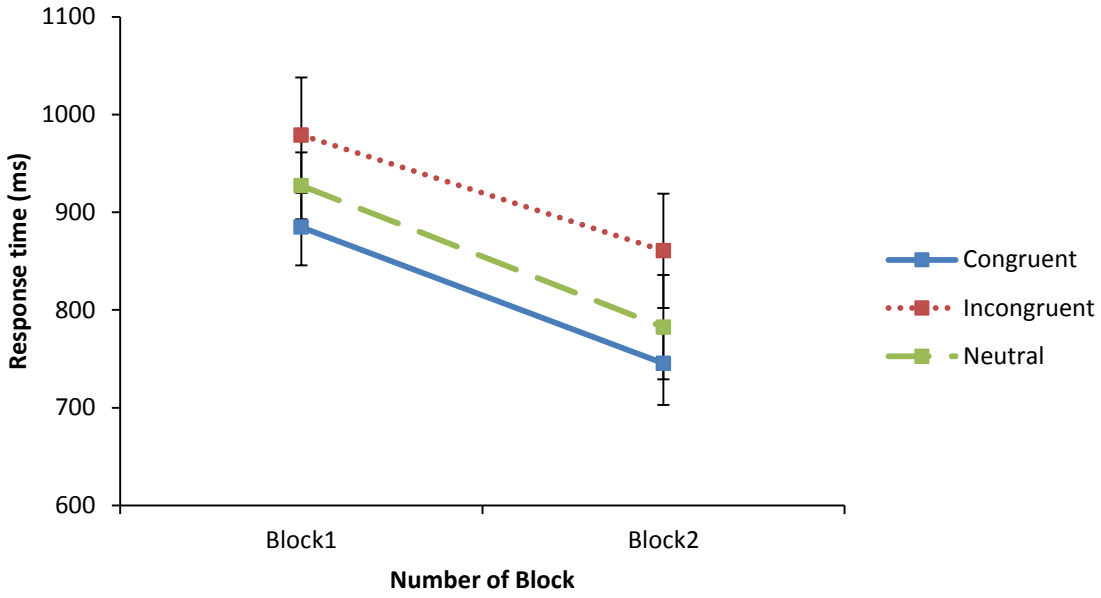


Figure 4.7. Experiment 9: Mean response times (in milliseconds) of correct judgements to critical word pairs across Block 1 and Block 2. Error bars indicate the 95% confidence intervals.

An interaction of Kinship term gender by Participant Gender also emerged, F_1 (1, 32) = 8.17, p = .007; F_2 (1, 33) = 8.32, p = .007, with female participants faster when responding to female kinship terms as opposed to male kinship terms (796ms vs. 855ms respectively, 59ms difference). However male participants were faster at responding to male kinship terms than female kinship terms, (885ms vs. 936ms respectively, 51ms difference), although they were slower than females at both.

There was also a main effect of Participant Gender in the by-items analysis, F_1 (1, 32) = 0.71, p = .406; F_2 (1, 33) = 19.34, p < .001, with male participants slower at responding than female participants overall (880ms vs. 815ms respectively).

Finally, a significant three-way interaction of Block by Congruency by Participant Gender was found in the by-participants analysis, F_1 (1.82, 58.12) = 4.21, p = .023; F_2 (2, 33) = 1.62, p = .214. To investigate this interaction in more detail, the mean difference between RTs of Block 1 and 2 across conditions for both males and females were examined. The results of this analysis can be seen in Figure 4.8 below. Positive data indicates male participants responded more slowly than females, while negative data indicates female participants responded more slowly than males.

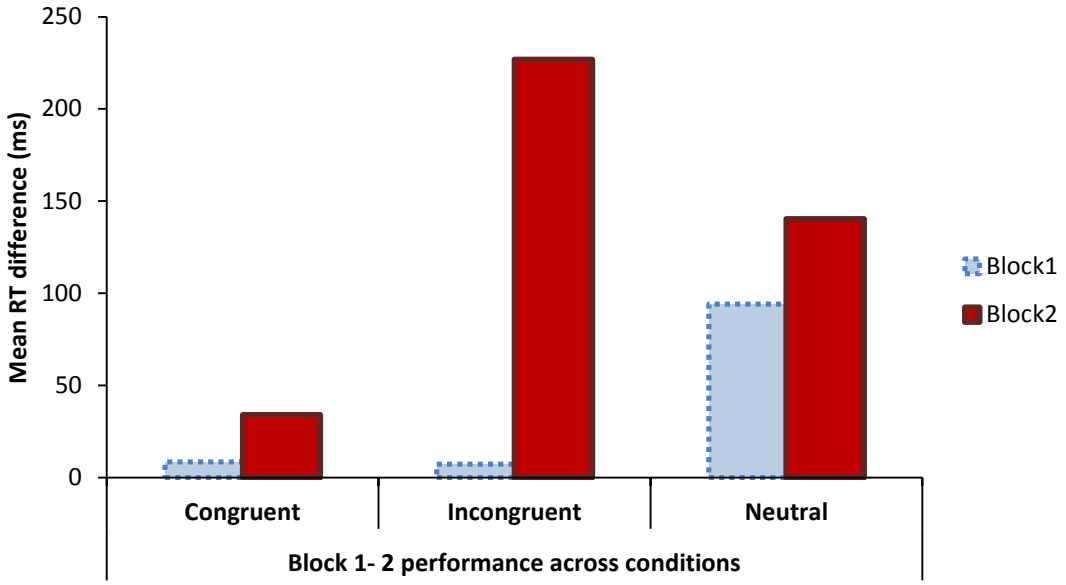


Figure 4.8. Experiment 9: Mean RT difference (in milliseconds) between Block 1 and Block 2 performance of female and male participants in response to critical word pairs.

The positive data displayed in Figure 4.8 illustrates that male participants were slower at responding than females across all conditions. While both sets of participants had quite similar response times to stereotype congruent word pairs, male participants were on average 117ms slower than females to respond to neutral word pairs across blocks. Response times of both sexes differed most radically in response to stereotype incongruent word pairs. In Block 1, male participants were just 7ms slower at responding to incongruent word pairs than female participants. However, as the behavioural task continued, female participants quickened their pace of responding and revealed much faster RTs than male participants in Block 2 (227ms difference). Again, there is no obvious reason for which male participants were much slower at responding in this study than female participants, yet combined with the accuracy data, it appears they are having much more difficulty with the incongruent pairings than the females.

Fillers - Accuracy

Performance on the definitionally matching word pairs revealed a high mean accuracy score of 93.6% across blocks, with similar performance on both the male (94.2%) and female pairings (92.9%).

However, performance deteriorated on the definitionally mismatching word pairs (*M* = 83.7%). Average accuracy to definitionally female pairings was high at 91.3%, but dropped to 76.2% for

the definitionally male pairings. Again, it is hypothesised that this difference in accuracy performance is due to the generic interpretation of certain male terms that are in fact male-specific by definition.

Fillers - Response times

The response time data tell a similar story to the accuracy data. Reaction times to both male and female definitionally matching word pairs were quite similar (926ms for the male versus 880ms for the female pairs) with an average RT of 903ms across blocks.

Average RTs in the defintionally mismatching condition were slower, at 982ms. Female mismatching pairings were responded to faster (943ms) than male mismatching pairings (1022ms) in general, again thought to reflect participants' indecision over certain male terms that may be used generically despite their gender specific definitions.

4.5.4 Discussion

This control experiment sought to maintain the stereotypical gender bias associated with certain role terms in English, by presenting participants with pictures of men and women working in gender *stereotypical* roles. The hypothesis that accuracy performance to the stereotype incongruent word pairs would not improve across blocks in the judgement task was indeed supported. However, response times to stereotype incongruent pairings *were* found to speed up across blocks in all congruency conditions. This pattern of results suggests that participants were benefitting from a practice effect and naturally speeding up at the task as it progressed. However, while RTs in the current experiment improved consistently across all conditions, RTs to the stereotype incongruent pairings in Experiment 8 decreased more sharply, with final response times in line with those of stereotype congruent and neutral pairings.

One point worth noting across Experiments 8 and 9 is the contrasting performance levels of male and female participants in both studies. However, as there were no obvious differences between both participant samples, the reasons for these discrepancies remain unclear.

As accuracy of stereotype incongruent trials did not significantly increase across blocks in Experiment 9, it is concluded that this stereotype-consistent picture manipulation did not help participants to overcome gender stereotype biases. Moreover, it appears that the counter-stereotype manipulation in Experiment 8 was indeed the reason for the improved performance evident in Block 2 of the relevant judgement trials. However, as mentioned

earlier, Block 1 performance to incongruent word pairs differed considerably across experiments (particularly in the accuracy data), thus affecting the scope for further improvements. Therefore, one final combined analysis of Experiments 8 and 9 was required in order to more definitively establish whether the counter-stereotype manipulation led to significantly better task performance than the stereotype manipulation.

4.6 Experiments 8 and 9: Combined analysis

Rationale & Hypotheses

Data from Experiments 8 and 9 were combined so as to investigate whether average performance on stereotype incongruent pairings was indeed better in Experiment 8 (counter-stereotypical pictures) compared to Experiment 9 (stereotypical pictures). More specifically, it was anticipated that accuracy of stereotype incongruent word pairs in Experiment 8 would be higher than in Experiment 9, while response times to stereotype incongruent word pairs in Experiment 8 would be faster than those of Experiment 9, in the judgement task.

Results

Analysis

The trimmed data from Experiments 8 and 9 were combined and both accuracy of judgements and response times were again analysed using two, mixed-design ANOVAs, as described in Section 2.2.3 However, in the F_1 analyses, Experiment (Experiment 8 and 9) was further added as a between-subjects factor, while in the F_2 analyses it was added as a within-items factor.

The main focus of this combined analysis was to examine performance to critical trials across experiments, in particular to stereotype incongruent pairings. For this reason, the findings reported below do not include effects that were found in both the individual experiment analyses (e.g. main effects of block, congruency, role noun gender etc.), but instead focus on the effects of particular interest.

Accuracy

Firstly, a significant interaction of Block by Experiment was found, F_1 (1, 60) = 4.36, p = .041; F_2 (1, 32) = 15.10, p < .001, with average accuracy increasing to a greater extent across blocks in Experiment 8 (+3.5%), than in Experiment 9 (+0.5%).

Importantly, there was also an interaction of Congruency by Block by Experiment, F_1 (1.50, 89.81) = 7.89, p = .002; F_2 (2, 32) = 12.48, p < .001, with the by-participants pattern of responding displayed in Figure 4.9 below. It is clear that this interaction was primarily driven by performance on stereotype incongruent pairings, as accuracy to these pairings increased 9.87% across blocks in Experiment 8, compared to just 0.12% in Experiment 9.

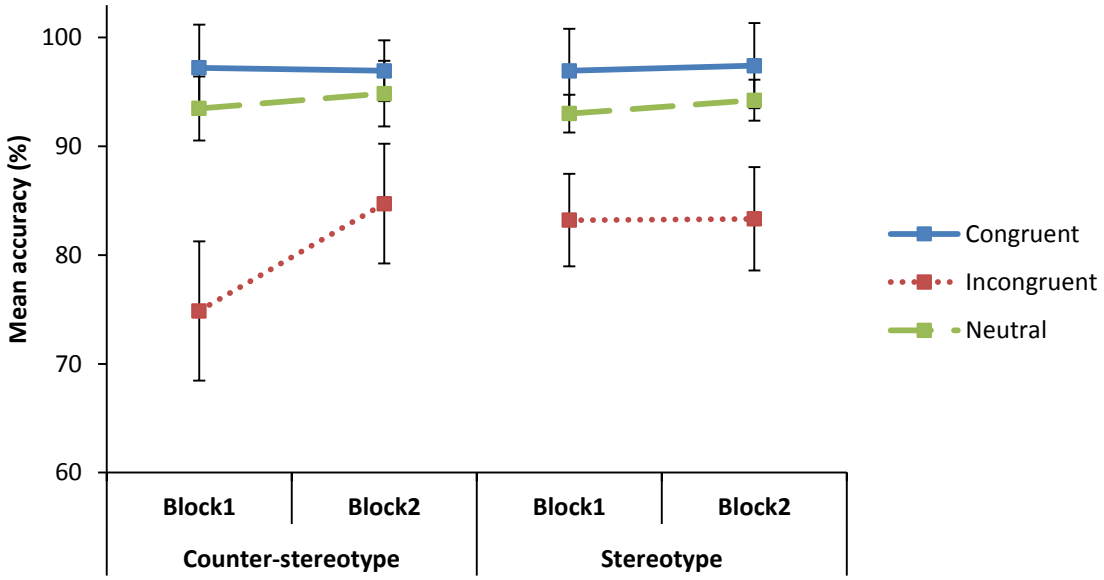


Figure 4.9. Mean percentages of correct judgements to critical word pairs across blocks in both picture experiments. The vertical axis begins at 60% while error bars indicate the 95% confidence intervals.

As mentioned in the individual experiment analyses, accuracy to incongruent pairings was found to rise significantly across blocks in Experiment 8 (p = .002) but not in Experiment 9 (p > .9). However, Figure 4.9 again illustrates that average Block 1 accuracy to incongruent word pairs in Experiment 8 was much lower than in Experiment 9 (7.1% difference) thus allowing much more scope for improvement in the former study. This Block 1 difference was not significant in the by-participants analysis, t_1 (62) = 1.17, p = .248, yet was highly significant in the by-items analysis, t_2 (23) = 4.89, p < .001.

Next, so as to more closely examine performance on the stereotype incongruent pairings across experiments, a second ANOVA was conducted on these incongruent pairings alone. Importantly, a Block by Experiment interaction was revealed, F_1 (1, 62) = 10.19, p = .002; F_2 (1, 23) = 28.27, p < .001**,** suggesting that accuracy of responding across blocks differed as a function of Experiment. As mentioned above, Block 1 accuracy was significantly different

across experiments in the by-items analysis, however no such difference emerged in Block 2 comparisons in either the by-participants or by-items analysis respectively, t_1 (62) = .20, p = .844, t_2 (23) = 1.04, p = .311.

Overall, while accuracy improved a greater amount in Experiment 8 than 9, it is not clear how the different picture strategies would have affected Block 2 performance if initial performance had been more similar. The reason(s) for such Block 1 accuracy differences across experiments remain(s) unknown as both experiments were identical up until the picture task (between Block 1 and Block 2 of the judgement trials), and there were no discernible differences between the participant samples (despite the variable male and female performance across experiments).

Returning to the first combined analysis (involving all three congruency conditions), a marginal interaction of Experiment by Participant Gender was found in the by-participants analysis yet this effect was highly significant by-items, F_1 (1, 60) = 3.90, p = .053; F_2 (1, 32) = 327.32, p < .001, with females outperforming males when shown stereotypical pictures (95.2% vs. 87.5%) while male participants outperformed females when shown counter-stereotypical pictures (93.6% vs. 86.5%).

Also, an interaction of Experiment by Participant Gender by Congruency was found in the by-items analysis only, F_2 (2, 32) = 21.38, p < .001, with females achieving particularly high accuracy to incongruent pairings in Experiment 9 compared to males (89.2 vs. 75.9, +13.3%), while the opposite pattern was found in Experiment 8 (75.0% vs. 85.3, -10.3%). However, accuracy to stereotype congruent and neutrally rated role names was similarly high across both experiments and participant genders.

Response times

An interaction of Congruency by Block by Experiment was found with the RT data in the by-participants analysis, F_1 (1.82, 108.94) = 4.15, p = .021, yet there was just a trend in this direction in the by-items data, F_2 (2, 32) = 2.45, p = .102. The former interaction is driven by variable performance across experiments, most notably to stereotype incongruent trials. This pattern of responding can be seen in Figure 4.10 below.

As previously stated, RTs to incongruent pairings were found to decrease significantly across blocks in both experiments. However, the relative decrease of RTs across blocks was greater in Experiment 8 (225ms, p < .001) than Experiment 9 (118ms, p = .009).

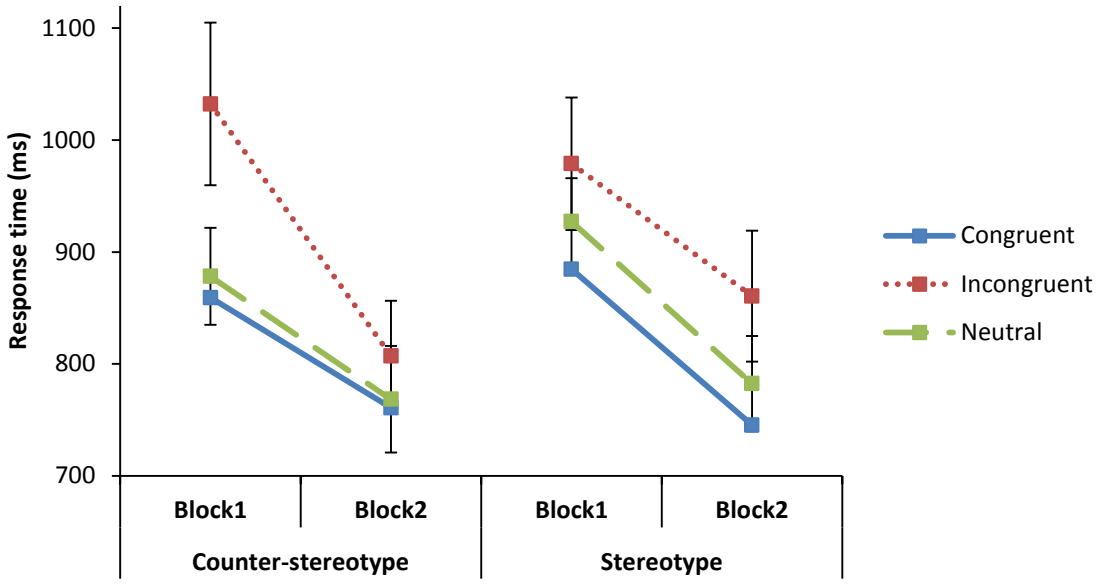


Figure 4.10. Mean response times to critical word pairs across blocks in both picture experiments. The vertical axis begins at 700ms while error bars indicate the 95% confidence intervals.

As with the accuracy data there was an interaction of Experiment by Participant Gender but in the by-items analysis only, F_2 (1, 32) = 17.52, $p < .001$, with females faster at responding than males when they were shown both the stereotypical pictures (815ms vs. 881ms, mean difference = 66ms) and the counter-stereotypical pictures (931ms vs. 773ms, mean difference = 158ms), but particularly so in the latter analysis (despite males achieving higher accuracy in this experiment).

Next, a second ANOVA was conducted on the stereotype incongruent pairings alone so as to more closely examine performance on these critical trials.

A significant interaction of Block by Experiment was not revealed in the by-participants analysis yet was significant in the by-items analysis, F_1 (1, 62) = 2.06, $p = .156$; F_2 (1, 23) = 4.86, $p = .038$, suggesting that RTs to critical pairings differed across blocks as a function of Experiment in this latter analysis only. Overall, while Block 1 RTs were not significantly slower in Experiment 8 than Experiment 9 in the by-participants analysis, t_1 (62) = 0.67, $p = .504$, this effect was significant by-items, t_2 (23) = 2.69, $p = .013$ (Experiment 8: 1042ms vs. Experiment 9: 971ms respectively). However, final Block 2 RTs were not significantly different across experiments in either analysis, t_1 (51.53) = 0.69, $p = .494$; t_2 (23) = 0.16, $p = .871$. Therefore, despite the greater RT reduction across blocks when participants were presented with

counter-stereotype pictures as opposed to stereotype consistent pictures, final RTs were not significantly different across experiments.

Combined analysis: Conclusion

The combined analysis of Experiments 8 and 9 sought to more comprehensively compare performance to critical incongruent trials across both counter-stereotypical and stereotypical picture conditions. Firstly, accuracy of participants in Experiment 8 (counter-stereotype pictures) improved to a much greater extent across blocks than those in Experiment 9 (stereotype pictures). However, this effect is driven by variable Block 1 performance while final accuracy is similar across both studies. With the response time data, it was found that responses became faster in Block 2 of both experiments (relative to Block 1), irrespective of the pictures presented. While, the magnitude of this decrease was greater when participants received the gender counter-stereotypical pictures as opposed to the gender stereotypical pictures, there was no significant difference between Block 2 RTs across experiments.

Notably, the combined analysis further revealed that female participants outperformed males in Experiment 9 (in both accuracy and response times), while males outperformed females in Experiment 8 (in accuracy only). These results suggest that female participants are perhaps generally more attuned to stereotype biases than males, yet males respond better when training is provided to improve performance in this area while, for some reason, female performance deteriorates. Further research may shed light on these sex differences in performance.

Overall, this combined analysis suggests that the use of counter-stereotype pictures is a relatively useful means of moderating the effects of immediate gender activation in the judgement task. However, given the variable Block 1 performance across experiments, strong conclusions cannot be drawn on this data.

4.7 Picture Booklets: analysis of participant responses

As previously mentioned, aside from simply presenting participants in Experiments 8 and 9 with 24 pictures of people working in counter-stereotypical and stereotypical roles respectively, participants were also asked to answer four questions on each of these pictures

in a booklet provided. The primary purpose of answering these questions was to focus the participants' attention on the pictures presented, and notably the job that each person was doing. The four questions that participants answered about each picture investigated the hypothesised (1) earnings: *How much do you think [insert character name] earns each year?* (2) job satisfaction: *How satisfied do you think [he/she] is with [his/her] job?* (3) lifestyle: *What are [his/her] leisure activities?* and (4) personal life of the characters presented: *Briefly describe [his/her] personal life.* However, as the pictures gave away relatively little about the life or job of the character (merely depicting them in a specific work environment), participants had little information upon which to base their responses to each of the questions. Combined with the behavioural data, the responses to these questions were therefore (a) another way of establishing that participants *do* indeed form occupational stereotypes about men and women working in different roles and (b) a window to the stereotypical views that participants hold about people in these different roles.

As will be described in more detail shortly, two of the questions above had categorical response options (earnings and job satisfaction) as participants made responses along a Likert scale. However, two of the questions were open-ended (lifestyle and personal life) and thus varied greatly in the level of detail and content provided by participants. In order to get an overview of themes running throughout these latter responses, six broad categories were identified according to which the responses were subsequently rated by two independent raters (this process will be later discussed with the analyses of questions 3 and 4). This rating procedure allowed for a statistical analysis of the subjective responses provided by participants in questions 3 and 4.

As a reminder, the pictures presented to participants depicted a combination of males working in female-typical roles, females working in male-typical roles (Experiment 8), males working in male-typical roles, females working in female-typical roles (Experiment 9). Averages responses to each of these categories were compiled, with eight comparisons then carried out:

1. Men in male stereotypical roles vs. Men in male counter-stereotypical roles
2. Women in female stereotypical roles vs. Women in female counter-stereotypical roles
3. Men in male stereotypical roles vs. Women in female stereotypical roles
4. Men in male counter-stereotypical roles vs. Women in female counter-stereotypical roles
5. Men in male counter-stereotypical roles vs. Women in female stereotypical roles

6. Women in female counter-stereotypical roles vs. Men in male stereotypical roles
7. All men vs. All women (across both stereotypical and counter-stereotypical roles)
8. All stereotypical vs. All counter-stereotypical roles (across both men and women)

These eight *t*-tests were chosen as, together, they cover all possible comparisons across the stimuli[75]. Responses to each of the four booklet questions are outlined in turn below, with mean scores and significance levels for each of the above comparisons also provided in Table 4.1.

[75] Two-tailed, independent samples *t*-tests were conducted for comparisons involving different pictures (comparisons 1, 2, 3, and 4 below) while two-tailed, related samples *t*-tests were used for comparisons involving pictures of the same occupations (comparisons 5, 6, 7 and 8 below).

Table 4.1. Booklet analyses: Mean scores and significance levels of comparisons across groups

	MS v MCs	FS v FCs	MS v FS	MCs v FCs	MCs v FS	FCs v MS	S v Cs	M v F
Earnings	3.72/3.08 *p* = .107	2.65/3.76 *p* = .008	3.72/2.65 *p* = .012	3.08/3.76 *p* = .081	3.08/2.65 *p* < .001	3.76/3.72 *p* = .432	3.42/3.19 *p* < .001	3.40/3.21 *p* = .005
Job Satisfaction	2.11/2.09 *p* = .953	2.14/2.11 *p* = .907	2.11/2.14 *p* = .889	2.09/2.11 *p* = .939	2.09/2.14 *p* = .570	2.11/2.11 *p* = .933	2.12/2.10 *p* = .657	2.10/2.12 *p* = .589
Leisure: **Male/Female**	1.43/1.95 *p* < .001	1.59/2.05 *p* < .001	1.43/1.59 *p* = .007	1.95/2.05 *p* = .343	1.95/1.59 *p* = .014	2.05/1.43 *p* < .001	1.51/2.00 *p* < .001	1.67/1.81 *p* = .087
Leisure: **Physical/Mental**	2.02/2.26 *p* = .062	2.40/2.16 *p* = .080	2.02/2.40 *p* = .002	2.26/2.16 *p* = .480	2.26/2.40 *p* = .020	2.16/2.02 *p* = .093	2.21/2.21 *p* = .997	2.13/2.28 *p* = .009
Leisure: **Social/Solitary**	1.84/1.94 *p* = .366	1.95/2.03 *p* = .385	1.84/1.95 *p* = .311	1.94/2.03 *p* = .409	1.94/1.95 *p* = .945	2.03/1.84 *p* = .028	1.89/1.99 *p* = .083	1.89/1.98 *p* =.074
Leisure: **Expensive/Cheap**	2.21/2.25 *p* = .704	2.33/2.33 *p* = .967	2.21/2.33 *p* = .352	2.25/2.33 *p* = .471	2.25/2.33 *p* = .265	2.33/2.21 *p* = .080	2.27/2.29 *p* = .590	2.23/2.33 *p* = .079
Personal Life: **Traditional/Non**	1.47/1.82 *p* = .004	1.69/1.64 *p* = .528	1.47/1.69 *p* = .034	1.82/1.64 *p* = .073	1.82/1.69 *p* = .073	1.64/1.47 *p* = .014	1.58/1.73 *p* = .021	1.63/1.66 *p* = .461
Personal Life: **Happy/Unhappy**	1.85/1.80 *p* = .433	1.88/1.90 *p* = .618	1.85/1.88 *p* = .666	1.80/1.90 *p* = .096	1.80/1.88 *p* = .113	1.90/1.85 *p* = .223	1.86/1.85 *p* = .747	1.83/1.89 *p* = .044

MS	Males in Stereotypical roles	S	Stereotypical roles
MCs	Males in Counter-stereotypical roles	Cs	Counter-stereotypical roles
FS	Females in Stereotypical roles	M	Males
FCs	Females in Counter-stereotypical roles	F	Females

	Significant (*p* < .05)
	Marginally significant (*p* < .1)
	Not significant

Question 1; Earnings

Booklet Question: How much do you think [insert character name] earns each year?

Response options to the above question were on a 6-point scale, ranging from (1) < £10,000 to (6) > £50,000. Response averages are stated below (from highest to lowest):

- Females in counter-stereotypical roles (i.e. male typical jobs): 3.76
- Males in stereotypical roles (i.e. male typical jobs): 3.72
- Males in counter-stereotypical roles (i.e. female typical jobs): 3.08
- Females in stereotypical roles (i.e. female typical jobs): 2.65

Interestingly, it was found that females working in male typical roles were rated as earning the highest amount, just above average ratings for males in these roles. However, on the whole, it was males that were judged as earning more than females; a somewhat unsurprising finding that echoes the frequently reported wage inequality among the sexes (e.g. Eurostat, 2013). This result was particularly driven by the fact that males working in female typical roles were judged as earning significantly more than females in the same roles.

It was also found that male typical jobs were judged as being better paid than female typical jobs. This pattern of results is likely to reflect the high status associated with some of the typically male jobs used in this study (e.g. surgeon, judge, architect) compared to the lower status associated with many of the typically female jobs used (e.g. cleaner, hairdresser, au-pair).

Question 2; Job satisfaction

Booklet Question: How satisfied do you think [he/she] is with [his/her] job?

Response options to the above question were on a 5-point scale ranging from (1) 'Extremely Satisfied' to (5) 'Extremely Dissatisfied'. Again, response averages are stated below (from highest to lowest):

- Female in stereotypical roles (i.e. female typical jobs): 2.14
- Female in counter-stereotypical roles (i.e. male typical jobs): 2.11
- Male in stereotypical roles (i.e. male typical jobs): 2.11

- Male in counter-stereotypical roles (i.e. female typical jobs): 2.09

However, it was found that ratings of job satisfactions were very similar across roles and genders with typical ratings falling close to 2 (quite satisfied). Indeed no significant differences emerged on any of the eight comparisons that were examined.

Questions 3 and 4

As mentioned above, questions 3 and 4 were open-ended and elicited a varied array of responses. In order to get an overview of themes running throughout the data, a female assistant read the responses and identified broad categories according to which the responses could subsequently be rated. The suggested categories were then discussed with the experimenter and a final set was decided upon. For question 3 (about the hypothesised leisure activities of a character), four rating categories were chosen:

(1) Male vs. Female leisure activities

(2) Physical vs. Mentally-oriented leisure activities

(3) Social vs. Solitary leisure activities

(4) High vs. Low cost leisure activities

With question 4 (about the hypothesised personal life of a character), two rating categories were selected:

(1) Traditional vs. Non-traditional personal life

(2) Happy vs. Unhappy personal life

Each of the six categories stated above had three potential rating responses (decided upon by the experimenter). For instance, with the category of Male vs. Female leisure activity, ratings could be either 1 (typically male), 2 (neutral) or 3 (typically female). The female assistant rated participants' responses according to each of the six categories and a male rater was also recruited to complete the same task (given the subjectivity of the response-ratings, it was felt that a gender balance in this assignment was important). Both raters were provided with instructions explaining their task and given examples of response ratings for each of the six categories (see Appendix 14). The raters first analysed the data independently before then meeting to compare results and to reach a consensus on conflicting ratings. All inconsistencies

were resolved after discussion so all data was kept[76]. Ratings of the responses to each of the four categories for question 3 will be outlined first below.

Question 3. Leisure activities

Booklet Question: What are [his/her] leisure activities?

Category 1: Male-typical vs. Female-typical leisure activities

The leisure activity responses were first rated according to whether they were generally more male (e.g. 'football') or female-typical (e.g. 'ballet'). Responses were rated according to the following scale: 1 = Male-typical, 2 = Neutral, 3 = Female-typical.

However, as this category of male vs. female leisure activities specifically relates to gender, the female means were transformed (i.e. reversed) so as to make them directly comparable to the male means. For example a female scoring 3 (i.e. female typical leisure activity) was transformed to a score of 1. In this way, a score of 1 or close to 1 indicates that the leisure activity is rated as typical of the sex of the character in the picture, regardless of whether they are in a male or female typical role. For instance, the mean score of 1.59 for females working in a counter-stereotypical role (in Table 4.1) means that they are thought to have more female-typical leisure activities than male-typical leisure activities. Equally, if the men working in a male counter-stereotypical role had achieved this score, it would signify that they were thought to have more male-typical leisure activities than female-typical leisure activities. These reversed female means are stated in Table 4.1 above, (as the *t*-tests were based on the reversed data), but the original (unrecoded) means are provided in the descriptive summary of results below. Note that as the subsequent rating categories do not involve gender, this transformation was not necessary in further analyses.

Response averages to this category are stated below (from lowest i.e. male typical to highest i.e. female typical):

- Males in stereotypical roles (i.e. male typical jobs): 1.43
- Males in counter-stereotypical roles (i.e. female typical jobs): 1.95
- Females in counter-stereotypical roles (i.e. male typical jobs): 1.95

[76] Unfortunately, more detailed descriptives on the ratings and how many original disagreements there were among the raters cannot be provided due to loss of data.

- Females in stereotypical roles (i.e. female typical jobs): 2.41

Unsurprisingly, it was found that males working in male typical jobs were judged as having significantly more male-typical leisure activities than those working in counter-stereotypical roles. Similarly, females working in female typical jobs were judged as having significantly more female-typical leisure activities than those working in counter-stereotypical roles. Of most interest is the fact that both males and females working in counter-stereotypical roles achieved identical ratings. Firstly, females in counter-stereotypical roles (e.g. carpenter, surgeon) were judged as having slightly more male-typical leisure activities than female-typical leisure activities. On the other hand, males working in counter-stereotypical roles (e.g. nurse, florist) were *not* judged as having more female-typical leisure activities than male-typical leisure activities. However, this final comparison was not significant.

Category 2: Physical vs. Mentally-oriented leisure activities

The leisure activity responses were next rated according to whether they were typically more physically (e.g. 'rugby') or mentally-oriented (e.g. 'reading'). The rating scale was as follows: 1 = Physical, 2 = Physical and Mental, 3 = Mental. Response averages are stated below (from lowest i.e. physical activities to highest i.e. mental activities):

- Males in stereotypical roles (i.e. male typical jobs): 2.02
- Females in counter-stereotypical roles (i.e. male typical jobs): 2.16
- Males in counter-stereotypical roles (i.e. female typical jobs): 2.26
- Females in stereotypical roles (i.e. female typical jobs): 2.40

In this category, it was found that leisure activities were rated as typically more mentally-oriented than physically-oriented in each of the above comparison groups. Overall, female-typical leisure activities were rated as being more mentally oriented than the male-typical leisure activities, and indeed males on the whole were judged as having significantly more physical leisure activities than females.

Category 3: Social vs. Solitary leisure activities

The leisure activity responses were next rated according to whether they were typically more social (e.g. 'meeting friends') or solitary-oriented (e.g. 'reading'). The rating scale was as follows: 1 = Social, 2 = Neutral, 3 = Solitary. Response averages are stated below (from lowest i.e. social activities to highest i.e. solitary activities):

- Males in stereotypical roles (i.e. male typical jobs): 1.84
- Males in counter-stereotypical roles (i.e. female typical jobs): 1.94
- Females in stereotypical roles (i.e. female typical jobs): 1.95
- Females in counter-stereotypical roles (i.e. male typical jobs): 2.03

Leisure activities were typically rated as being more social than solitary. The only significant difference found in relation to this category was that females working in male-typical roles were rated as having less socially-oriented leisure activities than males occupying these same, male-typical roles.

Category 4: High vs. Low cost leisure activities

Finally, the leisure activity responses were rated according to whether they were typically expensive (e.g. 'golf') or cheap (e.g. 'reading'). The rating scale was as follows: 1 = Expensive, 2 = Reasonable, 3 = Cheap. Response averages are stated below (from lowest i.e. expensive to highest i.e. cheap activities):

- Males in stereotypical roles (i.e. male typical jobs): 2.21
- Males in counter-stereotypical roles (i.e. female typical jobs): 2.25
- Females in stereotypical roles (i.e. female typical jobs): 2.33
- Female counter-stereotypical roles (i.e. male typical jobs): 2.33

Leisure activities were typically rated as being more cheap than expensive. While no significant differences emerged in any of the group comparisons, there was a trend suggesting that females are thought to have cheaper leisure activities than males.

Next, responses to the last of the four booklet questions were rated according to two different categories. Findings are described in further detail below.

Question 4. Personal life

Booklet Question: Briefly describe [his/her] personal life

Category 1: Traditional vs. Non-traditional

The personal life responses were first rated according to whether they were typically more traditional (e.g. 'married with children') or non-traditional (e.g. member of a 'lesbian couple').

The rating scale was as follows: 1 = Traditional, 2 = Neutral, 3 = Non-traditional. Response averages are stated below (from lowest i.e. traditional to highest i.e. non-traditional):

- Males in stereotypical roles (i.e. male typical jobs): 1.47
- Females in counter-stereotypical roles (i.e. male typical jobs): 1.64
- Females in stereotypical roles (i.e. female typical jobs): 1.69
- Males in counter-stereotypical roles (i.e. female typical jobs): 1.82

Firstly, it is apparent that personal lives were rated as being more traditional than non-traditional across all groups. A striking finding from the above category is that males working in stereotypical roles are judged as leading the most traditional personal lives while males in counter-stereotypical roles are judged as leading the least traditional personal lives. With the females it is surprising to note that those working in counter-stereotypical roles are judged as leading slightly more traditional personal lives than those in stereotypical roles (although this difference was not significant). Both of these findings suggest that it is now more frequent and less surprising for women to enter male-typical roles than for men to enter female-typical roles. Overall it is those working in stereotypical roles that are judged as leading more traditional personal lives than those working in counter-stereotypical roles.

Category 2: Happy vs. Unhappy

Finally, the personal life responses were rated according to whether they were typically happy (e.g. 'happily married with children') or unhappy (e.g. 'lonely'). The rating scale was as follows: 1 = Happy, 2 = Neutral, 3 = Unhappy. Response averages are stated below (from lowest i.e. happy to highest i.e. unhappy):

- Males in counter-stereotypical roles (i.e. female typical jobs): 1.80
- Males in stereotypical roles (i.e. male typical jobs): 1.85
- Females in stereotypical roles (i.e. female typical jobs): 1.88
- Females in counter-stereotypical roles (i.e. male typical jobs): 1.90

Only one significant difference in relation to the happiness ratings of people's lives was noted, with males judged as typically happier than females. It is surprising to note that males working in counter-stereotypical roles are judged as the happiest group while females working in counter-stereotypical roles are judged as the unhappiest.

Picture Booklets: Summary and Conclusion

Overall, the picture booklets provide interesting supplementary data on the perception of men and women working in stereotypical and counter-stereotypical occupational roles.

Males are thought to earn more than females when both sexes are depicted in gender-typical social roles, however, females are thought to earn more than males when depicted in counter-stereotypical roles. These findings reflect both the frequently-reported issue of males earning more than their female-counterparts in the work place and arguably the low-status of female-typical jobs relative to male-typical jobs. Despite this, the characters in the pictures presented to participants were generally judged as being quite satisfied with their jobs, whether working in stereotypical or counter-stereotypical occupational roles.

With the free-response questions, ratings of the characters' supposed leisure activities revealed that (1) male-typical leisure activities are judged as being more physical than mental (while the opposite is true for female-typical activities), (2) men and women working in gender-typical occupational roles were rated as having more male-typical and female-typical leisure activities (respectively) than men and women working in counter-stereotypical roles, and (3) females working in male-typical roles were thought to have less socially-oriented leisure activities than males occupying these same, male-typical roles. However, no significant differences in the cost of leisure activities were found across stereotypical and counter-stereotypical roles.

Finally, analysis of responses relating to the personal lives of those depicted working in various social roles showed that (1) those working in stereotypical roles were thought to lead significantly more traditional personal lives than those in counter-stereotypical roles (2) males occupying counter-stereotypical social roles were judged as leading less traditional personal lives than females working in counter-stereotypical jobs, (3) traditionality ratings of the females' personal lives was independent of whether they worked in stereotypical or counter-stereotypical roles while, and (4) no significant difference emerged as regards the happiness of men and women working in stereotypical or counter-stereotypical roles.

In conclusion, while the booklet data provides interesting information as regards occupational stereotypes, integrating the current findings into the social psychological literature on the perception of men and women working in different roles is not in keeping with the main theme of this thesis. Future research could further examine the themes which have emerged

in the current analysis and explore the relationship between explicitly held stereotypical views about people working in different occupational roles and implicit occupational stereotypes.

4.8 Chapter Discussion

Chapter 4 investigated the use of counter-stereotype information as a moderator of gender stereotype use. To begin, Experiment 7 employed an association learning paradigm in which participants learnt a list of gender counter-stereotypical word pairs (e.g. David, Beautician). It was hypothesised that this AL task would alert participants to the fact that males can also occupy female-biased social roles and conversely, that females can occupy male-biased roles. Contrary to expectations, only a 2.41% improvement in accuracy was found across blocks while response times tended to rise in Block 2 as opposed to decrease. This pattern of results clearly failed to provide support for the use of this AL training as a strategy for overcoming gender stereotyping. However, interpretation of these results is complicated by the fact that the design of Experiment 7 was somewhat different to the preceding studies in this thesis. The design modifications (greater number of trials across fewer blocks, questionnaires after the judgement task as opposed to before) may have contributed to the fact that Block 1 accuracy of Experiment 7 was unusually high, thus leaving less room for improvement across blocks. Overall, it was initially hypothesised that stereotype reduction would occur slowly upon encountering numerous counter-stereotype exemplars (in line with the bookkeeping theory of stereotype reduction). However, as the AL task was a relatively subtle way of presenting counter-stereotype information to participants, the results of Experiment 7 suggested that a more striking training task in which counter-stereotype information is more obviously presented may be required to induce change.

Experiment 8 built on the above findings and employed the use of overt counter-stereotype pictures as a means of stereotype reduction. This training involved presenting participants with pictures of people working in gender counter-stereotypical roles, and also answering questions about the characters in these pictures. It was hypothesised that the questions would focus participants on the characters presented (specifically their jobs), and that the pictures would be a salient reminder that people can work in gender atypical roles. It was found that accuracy did increase significantly after this picture training, and importantly, did not increase in Experiment 9; a control experiment in which participants were presented with pictures of people working in gender stereotypical roles. However, interpretation of the results is again not straightforward as Block 1 accuracy was much higher in Experiment 9 than Experiment 8;

the reason(s) for which remain unknown. Response times decreased across blocks in both Experiment 8 and 9, independently of the type of picture training received. On the whole, while the results should be interpreted with caution because of the differential Block 1 performance across studies, it appears that activating counter-stereotype gender associations did lead to a revision of participants' stereotyped beliefs and ultimately helped participants to control stereotype use in the judgement task. Furthermore, the use of more striking stimuli as part of the counter-stereotype training in Experiment 8 provides support for the conversion theory of stereotype change i.e. that stereotypes are likely to change rapidly, upon encountering few, yet striking, counter-stereotype exemplars.

Finally, the findings of Experiment 8 support the call of Macrae and Bodenhausen (2000) to move beyond the use of verbal stimuli (category labels) in research on category activation and to use more realistic stimuli. While future research would undoubtedly benefit from an investigation of the cognitive processes involved in stereotype activation upon encountering real people, the use of pictures of people at work is a promising step in the right direction towards identifying further effective means of stereotype reduction.

5. Individual differences in gender stereotyping

5.1 Introduction

Allport (1935) posited that the attitude construct is critical in social psychology research as attitudes routinely influence our actions and the way we perceive the world. For instance, if we hold positive attitudes towards a particular social group, then we will interact with and perceive this group in a more positive manner than social groups that we hold more negative attitudes towards. Given the pervasive influence that attitudes have over every day behaviour (e.g. see Ajzen & Fishbein, 2005 for an overview), it is important to understand how they develop and how individual differences in attitudes can impact on prejudice and stereotyping.

A growing body of evidence now suggests that, although prejudice can form based on particular experiences with a group, it more commonly mirrors a general predisposition to evaluate a number of outgroups negatively (Hing & Zanna, 2010). Indeed such dispositional susceptibility to prejudice was long-ago proposed by Allport (1954), and more recently affirmed, with evidence suggesting that individual differences in personality traits (e.g. authoritarianism; Altemeyer, 1996), cognitive biases (e.g. categorisation; Ashburn-Nardo, Voils, & Monteith, 2001) and socio-political ideologies (e.g. a desire for group hierarchies; Sidanius & Pratto, 1999) can all influence levels of prejudice towards social groups.

In the domain of stereotyping, research also indicates that stereotype use is influenced by certain individual differences. Carter, Hall, Carney, and Rosip (2006) investigated personal differences in willingness to rely on the use of stereotypical information when interacting with people of different social groups (as captured by a self-report measure they devised; the Acceptance of Stereotyping Questionnaire). Among other findings, Carter and colleagues report that higher acceptance of stereotyping was associated with (1) higher implicit and explicit stereotyping of specific groups, (2) with less liberal gender role values, (3) with less agreeable and more agentic personality traits, and (4) with greater use of social categories (such as gender or race) when rating the similarity of faces. Therefore, while some people may consider stereotype application to be a useful starting point from which to guide behaviour towards another person, others are more doubtful of the usefulness and validity of stereotypes and prefer to build knowledge of others from the ground up. With such inter-personal differences in attitudes towards the utility of stereotypes, Carter et al. (2006) argue that any attempt to examine stereotype use must consider the issue of individual differences.

In a similar vein, Moskowitz (1993) explains that social categorisation is a motivated process, and people differ in the extent to which their processing is driven by their motives. He investigated how individual differences in a person's need to control and impose structure on their social world (as measured by the Personal Need for Structure scale (PNS), of Thompson, Naccarato, and Parker (1992)), influences the categorisation of information.

Moskowitz found that levels of PNS do indeed dictate the extent to which people engage in the categorisation process; specifically that high levels of PNS predict increased formation of spontaneous trait inferences on the basis of single behaviours. Previous research has also found that high PNS predicts the influence of social stereotypes on judgement (Neuberg & Newsom, 1993), the formation of stereotypes about new social groups (Schaller, Boyd, Yohannes, & O'Brien, 1995) and can increase the likelihood of assimilating new information into previously existing mental representations, as opposed to creating new representations (Thompson, Roman, Moskowitz, Chaiken, & Bargh, 1994).

As stereotypes are instantly activated upon exposure to a category cue, the influence of a person's attitude on *spontaneous* behaviour is an important consideration in research. The processes involved in such behaviour are proposed by the MODE model (Motivation and Opportunity as Determinants of behaviour) of Fazio (1990). This model posits that a perceiver's behaviour is based on a conscious consideration of the current information available to them - provided that they have adequate motivation *and* the opportunity to consider this information. However, if motivation or opportunity are lacking, then the most accessible attitudes are hypothesised to predict behaviour.

Researchers have also recognised the influence of other variables on the prediction of behaviour. For instance, in their composite model of attitude-behaviour relations, Eagly and Chaiken (1993, 1998) assert that factors such as a person's habits, their attitude towards the target(s) (of the behaviour), practical outcomes (i.e. rewards/punishments that may stem from performing the behaviour), approval vs. disapproval from others and self-identity outcomes (i.e. the effect of the behaviour on an individual's self-concept) will affect a person's behaviour.

Despite the research outlined above depicting the many ways in which individual differences in attitudes can affect prejudice and stereotyping, these issues have been largely overlooked in relation to gender stereotyping and the stereotype reduction literature. Consequently, a central aim of this thesis was to investigate whether an individual's pre-existing personal beliefs or goals can moderate or even entirely inhibit gender stereotyping.

As mentioned in Section 1.7, in the domain of gender stereotyping, past research has been equivocal on how personal beliefs can influence stereotype activation and application. However, a large body of research has grown to suggest that individual differences can affect levels of stereotype *application (*e.g. Monteith, 1993), automatic stereotype *activation* (e.g. Kawakami et al., 1998; Lepore & Brown, 1997) and even *pre-conscious* activation of stereotypes (e.g. Moskowitz et al., 1999). More specifically, in relation to the online processing of gender-biased role nouns, Gabriel et al. (2010) found that levels of sexism (as measured by the Modern Sexism Scale, Swim et al., 1995) moderated the processing of gender-stereotyped role nouns in a sentence reading task (with English speaking participants). Their findings suggest that readers' cognitive representations of gender are indeed moderated by sexist beliefs.

In conclusion, the importance of including individual difference measures when investigating factors that bias and guide one's social behaviour, is glaringly apparent. In an attempt to address this issue, participants across Experiments 1-9 were each administered one or more individual difference measures. The findings from these measures, and their relationship with the behavioural judgement task, are outlined throughout this chapter so as to more comprehensively investigate how individual differences may moderate the processing of gender-biased role nouns in English.

5.2 The Individual difference measures

To begin, each of the individual difference measures employed in this research is briefly described below.

The Ambivalent Sexism Inventory (ASI; Glick & Fiske, 1996).

The ASI is a 22-item questionnaire relating to men and women and their relationships in contemporary society. It is composed of two subscales assessing both Hostile Sexism (i.e. sexist antipathy endorsing male power e.g. "Most women fail to appreciate fully all that men do for them") and Benevolent Sexism (i.e. a subtle form of prejudice that endorses stereotypic views of women in restricted roles and depicts women as creatures who should be protected, supported and adored e.g. "Every man ought to have a woman whom he adores"). Respondents rate their agreement with the questionnaire items on a 6-point Likert scale ranging from (0) disagree strongly to (5) agree strongly. The ASI score is calculated as the mean score across items on both subscales.

Across six studies, Glick and Fiske (1996) established convergent, discriminant and predictive validity of the scale. It has also been cross-culturally validated (over 15,000 respondents in 19 countries, with findings demonstrating that both hostile and benevolent sexism are widespread across cultures (Glick & Fiske, 2001; Glick et al., 2000; Glick, Sakalli-Ugurlu, Ferreira, & de Souza, 2002). A full copy of the ASI is provided in Glick and Fiske (1996).

The Bem Sex Role Inventory (BSRI; Bem, 1974).

The BSRI is a 60-item measure of gender role orientation. The scale consists of 20 masculine traits (e.g. Ambitious, Independent), 20 feminine traits (e.g. Gentle, Loyal) and 20 Neutral traits (Adaptable, Likable). Respondents are asked to indicate on a 7-point Likert scale ranging from 1 (never or almost never true) to 7 (always or almost always true) how well each of the personality characteristics describes him/herself. On the basis of their responses, each person receives three major scores: a Masculinity score, a Femininity score and, importantly, an Androgyny score. The Androgyny score is of most theoretical interest in this chapter as it captures the nature of the participant's total sex role. It is calculated as the absolute value of the Student *t*-test ratio between a participant's masculinity and femininity scores. Note that this is a 'no difference' measure that does not distinguish between individuals who score (1) high in both masculinity and femininity i.e. androgynous individuals or (2) those who score low in both masculinity and femininity i.e. sex undifferentiated individuals[77].

Bem originally administered the BSRI to 917 students (561 male and 356 female). Internal consistency scores for all three sets of traits were found to be high across samples, while there was also evidence of good test-retest reliability over a four week period[78]. A full copy of the BSRI is provided in Bem (1974).

[77] Bem later developed a scoring measure which does distinguish between participants in this way (as she also suspected that scoring low on both measures was an indicator of low self esteem as opposed to androgyny). However, Bem and Lenney (1976) previously found that only 1% of their participants scored below the midpoint on both the Masculinity and Femininity scales and thus felt justified that their past participants were appropriately classified as androgynous. Indeed, only 1 participant in the current thesis scored under the midpoint on both scales, thus the original method for calculating BSRI scores was deemed satisfactory for this research.

[78] However, although the BSRI has been widely used in psychology research, it has also received frequent criticism and questioning (e.g. Pedhazur & Tetenbaum, 1979; Wheeless & Dierks-Stewart, 1981). Nevertheless, it is still used relatively often in current research (as evidence by a Web of Knowledge search for the questionnaire e.g. 156 publication results from the years 2000 - Nov 2013).

The Inventory of Attitudes towards Sexist and Non-sexist Language – General (IASNL-G; Parks & Roberton, 2000).

The IASNL-G is a 21-item measure of attitudes towards sexist and non-sexist language composed of three distinct sections. Section one measures respondents' beliefs about sexist language (12 items e.g. "worrying about sexist language is a trivial activity"), Section two measures respondents' recognition of sexist terms (4 items e.g. "people should care about all mankind, not just themselves") while Section three investigates the willingness of respondents to use inclusive language (5 items e.g. "how willing are you to use the term *camera operator* rather than *cameraman*"). Across all sections respondents give their responses on a 5-point Likert scale (0-5, with response labels varying according to section[79]), with the IASNL calculated as the mean score across all items and sections. Low scores indicate a negative attitude towards non-sexist language while high scores indicate a supportive attitude towards non-sexist language.

In their initial study, Parks and Roberton established that internal consistency of the IASNL was high, and that the scale was found to have strong content, construct and discriminant validity across a number of different age groups and geographical regions (Parks & Roberton, 2000, 2001). A full copy of the IASNL is provided in Parks and Roberton, (2000).

Ethics Questionnaire (EthicsQ; McMinn, Lindsay, Hannum, & Troyer, 1990).

This measure was a written assessment of sexist pronoun use, presented to participants as an Ethics Questionnaire. Based on items from McMinn et al. (1990, Experiment 2), participants were presented with ethical dilemmas about a subject who was introduced using a stereotype-biased role noun. The participants' task was to write down the first thing they would do in response to the ethical dilemmas they were presented with e.g. "A business executive discovers a long-time employee has been stealing from the company. What should the executive do first?" However, due to the small number of items in the original measure of McMinn and colleagues (6, and only 4 of which were retained for use in the current set of studies[80]), a further 8 were now developed. Therefore, participants were presented with 12 items (containing 4 male-biased, 4 female-biased and 4 neutrally-rated role nouns). Two lists of six items were formed (with 2 male-biased, 2 female-biased and 2 neutrally-rated items in

[79] Section 1: the scale goes from (1) Strongly disagree to (5) Strongly agree
Section 2: the scale goes from (1) Not at all sexist to (5) Definitely sexist
Section 3: the scale goes from (1) Very unwilling to (5) Very willing
[80] The original terms *Robber* and *Professor* were omitted as they were not sufficiently stereotype-biased.

each). One list was presented at the very beginning of the experimental session (so as to assess sexist pronoun use before participants were alerted to the issue of gender stereotype biases in the judgement task) while the other was presented at the very end (so as to investigate whether sexist pronoun use diminished following the judgement/training task). Scores were simply calculated as the sum of sexist pronouns used (in some cases the sum of pronouns used before *and* after the judgement task while in other cases the sum used before the judgement task only (to examine pronoun use before training)). Finally, the order of presentation of these lists was counterbalanced across participants.

Despite the fact that the 'Ethics Questionnaire' of McMinn and colleagues has not been frequently employed in psychological research, it was reasoned that adding a measure of stereotype application would be an interesting addition to this research project. A full copy of the newly developed Ethics Questionnaire is provided in Appendix 15.

The gender-career Implicit Association Test (IAT; Greenwald, McGhee & Schwartz, 1998).

This individual difference measure uses speed of response to assess the strength with which the concepts *men* and *women* (represented by 5 male and 5 female proper names) are associated with the attributes of *career* and *family* (represented by 7 career-related terms and 7 family-related terms). Briefly, participants are instructed to press a specific key for items that form a certain concept-attribution pair (e.g. men/career) and a different key for the opposing pair (e.g. women/family). The specified concept-attribution pairings are then reversed (e.g. men /family and women/career) and participants complete the task once again. While there are further key switch elements of the task, the IAT score is derived from the difference in response latencies to these two tasks.

Note that this questionnaire was presented according to Greenwald's guidelines (http://faculty.washington.edu/agg/iat_materials.htm), and was scored using the most recently recommended scoring algorithm (Greenwald, Nosek, & Banaji, 2003). There is now a good deal of empirical evidence for the construct validity of this measure although there are also concerns that the IAT measures well-learned environmental associations as opposed to attitudes towards specific attitude objects (e.g. stigmatised group members)(Dasgupta & Greenwald, 2001).

The Modern Sexism Scale (MSS; Swim et al., 1995).

The MSS is an 8-item questionnaire aimed at measuring relatively subtle forms of sexism, such as denying the continued discrimination against women and not admitting a lack of support for policies aimed at helping women e.g. "Discrimination against women is no longer a problem in the United Kingdom[81]". Respondents are asked to indicate on a 5-point Likert scale the extent to which they agree with a series of statements ranging from 1 (strongly agree) to 5 (strongly disagree). The MSS score is calculated as the mean score across items, with higher values indicating greater levels of modern sexism relative to lower values.

Swim and colleagues (1995, Study 2) demonstrated construct validity of the MSS as this measure predicted voting preferences and explanations for job segregation, yet they did not investigate reliability of the scale. However, Parks and Roberton (2004) state that in previous research, Cronbach's alphas for the MSS have ranged from .65 to .84 (e.g. Campbell, Schellenberg, & Senn, 1997; Swim & Cohen, 1997) while they themselves report a Cronbach's alpha for the MSS of .76. A full copy of the MSS is provided in Swim et al. (1995).

The Brief Fear of Negative Evaluation scale (BFoNE; Leary, 1983).

The BFoNE (developed from the Fear of Negative Evaluation (FNE) scale of Watson & Friend, 1969) is a 12-item questionnaire investigating the extent to which people experience anxiety at the prospect of being negatively evaluated by others e.g. "I often worry that I will say or do the wrong things". Respondents are asked to indicate on a 5-point Likert scale how characteristic of themselves a series of statements are, ranging from 1 (Not at all characteristic of me) to 5 (Very characteristic of me). High scores on the BFoNE indicate higher fear of negative evaluation compared to lower scores.

The reliability of the BFoNE was unaffected by the reduced number of items compared to the FNE, with 93% of participants classified identically by both versions of the questionnaire (Leary, 1983). High internal consistency has also been reported (Cronbach's alpha of .90) along with good test-retest reliability. However, Leary also cautions that the validity of the scale is in need of more investigation. A full copy of the BFoNE is provided in Leary (1983).

[81] Note this item originally stated 'United States' but was modified to suit the British study sample. Similarly, item 6 of this scale makes reference to 'America', but this was changed to 'Britain' for the purposes of this study.

5.3 Individual difference analyses

A list of the individual difference measures used in each experiment is provided in Table 5.1 below. In Experiments 1-6, the measures were administered before the behavioural judgement task, while in Experiments 7-9 they were administered after the behavioural task[82].

Table 5.1. Experiments 1-9 and their associated individual difference measures.

	Experiment Description	**Questionnaires**
Experiment 1	Performance Feedback	EthicsQ, BSRI, ASI, IASNL,
Experiment 2	Control (no feedback)	EthicsQ, BSRI, ASI, IASNL,
Experiment 3	Long-Term Feedback	EthicsQ, BSRI, ASI, IASNL,
Experiment 4	Social Consensus Fb[83]	EthicsQ, BSRI, ASI, IASNL, & BFoNE
Experiment 5	Reverse Social Consensus Fb	EthicsQ, BSRI, ASI, IASNL, & BFoNE
Experiment 6	Accuracy and Consensus Fb	EthicsQ, BSRI, ASI, IASNL, & BFoNE
Experiment 7	Learning Association	BSRI, ASI, MSS, IAT
Experiment 8	Counter-stereotypic pictures	ASI
Experiment 9	Stereotypic pictures	ASI

In the majority of analyses that follow, individual difference data were combined across a number of experiments so as to increase statistical power[84]. Both correlational and regression analyses were conducted using the measures described above as well as two behavioural measures, outlined below:

a) Initial Performance

The first behavioural variable was a measure of performance in Block 1 only, hereafter referred to as the Initial Performance measure. This measure investigated the difference in

[82] Questionnaires were completed before the behavioural task in Experiments 1-6 so as to ensure that participants were not primed as to what the individual difference measures were investigating via the judgement task. However, a different approach was taken with Experiments 7-9; questionnaires were completed after the behavioural task so as to ensure that participants' performance on the behavioural task was not being influenced via priming from the individual difference measures.

[83] 'Fb' is used as an abbreviation of 'feedback'

[84] Note that analyses for single experiments did not show a reliable pattern so correlational analyses based on larger samples were thought to be more useful; data from 236 students were used in the majority of comparisons (those involving Experiments 1-6), although this number decreased depending on the exact measures under scrutiny.

performance to stereotype congruent and incongruent trials in Block 1 only, as Experiments 1-6 are almost identical up until this point (differing only in the instructional information given to participants that was experiment-specific). Block 1 accuracy to stereotype congruent pairings is typically higher than Block 1 accuracy to stereotype incongruent pairings, while Block 1 RTs to stereotype congruent pairings are typically faster (i.e. lower) than Block 1 RTs to stereotype incongruent pairings. High positive scores on the accuracy data reveal that higher accuracy was attained in response to congruent pairings over incongruent pairings, while high positive scores on the RT data reveal that a participant was faster on congruent pairings compared to incongruent pairings[85].

b) Performance Improvement

The second behavioural variable was a measure of change in susceptibility to stereotyping across blocks, hereafter referred to as the Performance Improvement measure. This measure investigated the difference in performance to stereotype congruent and incongruent trials in Block 1 (which is typically quite high) compared to the difference in performance to stereotype congruent and incongruent trials in Block 3 (which is typically lower than in Block 1, either due to the inclusion of a training strategy or practice effects)[86]. With both the accuracy and response time data, high positive scores on this measure indicate greater success i.e. a smaller gap between congruent and incongruent responding in Block 3 compared to pre-training scores in Block 1.

Essentially, while analyses with Initial Performance investigate whether the individual difference measures relate to initial behavioural measures of stereotyping, analyses with Performance Improvement examine whether they relate to effects of training.

Individual difference correlations; Experiments 1-6

As participants in Experiments 1-6 each completed the BSRI, ASI, IASNL, and Ethics Questionnaire, the individual difference data from these studies was combined for analysis

[85] Specifically, Initial Performance was calculated as Block 1 congruent scores minus Block 1 incongruent scores for the accuracy data, and Block 1 incongruent scores minus Block 1 congruent scores for the RT data.

[86] Specifically, this measure was calculated as the Block 1 congruent minus incongruent scores less the Block 3 congruent minus incongruent scores for the accuracy data, but Block 1 incongruent minus congruent scores less the Block 3 incongruent minus congruent scores for the response time data. In this way, high positive scores meant a big improvement (i.e. reduction of stereotyping). It is also worth noting that small values on these measures result from (a) participants showing little effect of stereotyping in Block 1 thus leaving little room for change and (b) participants showing large effects in Block 1, but who are unaffected by the training.

resulting in 235 participants (113 M, 122F)[87]. Participants in Experiments 4-6 additionally completed the BFoNE and again, their data was combined resulting in 118 participants (57M, 61F). Based on all combined data, the first analysis examined correlations between the questionnaires themselves. Findings are displayed in Table 5.2 below. Note that overall ASI scores were calculated for participants along with their Hostile and Benevolent subscale scores, and finally, the Ethics Questionnaire in this instance refers to total sexist pronoun use (i.e. sexist pronoun use at the beginning and end of the study[88]).

Table 5.2. Individual difference measures; Correlations (using Pearson's correlation coefficients).

	BSRI	**ASI**	**Hostile**	**Benev**	**IASNL**	**BFoNE**	**EthicsQ**
BSRI	1						
ASI	.09	1					
Hostile	.12	.86***	1				
Benevolent	.07	.86***	.53***	1			
IASNL	-.16*	-.45***	-.49***	-.32***	1		
BFoNE	.21*	.02	.06	-.03	-.04	1	
EthicsQ	.04	.17**	.14*	.15*	-.18**	.09	1

*$p < .05$, **$p < .01$, ***$p < .001$ (two-tailed)

A number of significant correlations were found between the various individual difference measures.

Firstly, a small, negative correlation was found between the BSRI and IASNL, with higher BSRI scores (indicating sex-typed individuals) correlating with lower IASNL scores (indicating a less supportive attitude to non-sexist language).

[87] Individual difference data from one other participant was not included due to both outlier and technical issues (use of 12 sexist pronouns in part 1 of the Ethics Questionnaire, and missing ASI data respectively).

[88] Note that cases of counter-stereotype pronoun use were extremely infrequent and so were not analysed in detail.

A significant, positive correlation between the BSRI and BFoNE was also revealed, with higher BSRI scores (indicating sex-typed individuals) correlating with higher BFoNE scores (indicating higher fear of negative evaluations).

Next, a highly significant negative correlation between the IASNL and the ASI was found, with higher IASNL scores (indicating a more supportive attitude to non-sexist language) found to correlate with lower ASI scores (indicating lower ambivalent sexism). This same pattern of results was also found with both Hostile and Benevolent subscale scores and the IASNL.

Finally, total sexist pronoun use (the EthicsQ data) was found to correlate with both the ASI (positively) and the IASNL (negatively). Scatterplots revealed that those scoring higher in sexism tended to use a greater number of sexist pronouns, while those with a more supportive attitude to non-sexist language tended to use fewer sexist pronouns.

A Wilcoxon test[89] was conducted to evaluate whether sexist pronoun use was greater before the experimental session or after. The results revealed a significant difference, $z = -3.12$, $p = .002$. The mean of the ranks in sexist pronoun use at the beginning of the experiment was 65.54 while the mean of the ranks in sexist pronoun use at the end of the experiment was 58.64[90].

Analysis of Experiments 1-6: ASI, BSRI, IASNL & Ethics Q

Hierarchical regression analyses were next conducted on the data from Experiments 1-6 so as to examine the contributions of the individual difference measures BSRI, ASI, IASNL and sexist pronoun use (the predictor variables[91]) in predicting performance on the behavioural measures of Initial Performance and Performance Improvement (the outcome variables) outlined earlier. Five dummy-coded experiment variables were also included as predictors, with Experiment 1 chosen as the reference because of its large sample size (Field, 2009).

The experiment dummies were entered in Step 1 using forced entry (due to the different effects of training found in Experiments 1-6 that are already known from the main analyses).

---

[89] A Wilcoxon test was used as the EthicsQ data was not normally distributed, nor was there independence of observations across the two groups (sexist pronoun use before and after the behavioural experiment).

[90] Note that in cases where data was missing on this measure (i.e. participants failed to type a response to a particular question on this measure), omissions were replaced by the participant's average for that section.

[91] The BFoNE is examined separately in Analysis 5 and 6.

Therefore, any significant correlations between the experiment dummies and the behavioural measures, and any experiment dummies that emerged as significant predictors of behavioural performance will be indicated, but not further discussed in this chapter.

Next, the ASI, the BSRI and the IASNL were entered in Step 2 as these individual difference measures are frequently used in social psychology research and have established validity. Finally, the EthicsQ data was entered in Step 3 as it is an infrequently used measure, with unknown predictive validity. Moreover, the original items on this measure were modified and extended by the researcher for the current research. This regression procedure was followed for Analyses 1-4 which are now described further below.

Analysis 1; Initial Performance - Accuracy

For an initial assessment of the relationships between the individual difference measures and the outcome measure of Initial Performance, Pearson's correlation coefficients are presented in Table 5.3 below.

Table 5.3. Analysis 1: Correlation coefficients of the nine predictors and Initial Performance[92].

	r
Initial Performance	
(Outcome variable)	1
Exp2vs1	.05
Exp3vs1	-.03
Exp4vs1	.07
Exp5vs1	-.06
Exp6vs1	.00
ASI	.28***
IASNL	-.17**
BSRI	.08
Pronouns (Before)	.01

***$p < .001$; **$p < .01$ (two-tailed)

[92] Although it is common practice to include the questionnaire correlations in this table, these will not be included here (or in further, equivalent correlation tables) as correlations among each of the measures are already provided in Table 5.2.

Both the ASI and the IASNL significantly correlated with the Initial Performance measure, r = .284, p < .001 and r = -.169, p = .005 respectively. Those scoring higher in sexism, and those with a less positive attitude towards non-sexist language use tended to have a greater gap between their accuracy scores on congruent and incongruent trials (indicating relatively poor stereotype incongruent performance) than those scoring lower in sexism or with a more positive attitude towards non-sexist language use (indicating relatively good stereotype incongruent performance).

Analysis of the regression output revealed that the assumptions of multiple regression analyses were largely met according to guidelines outlined by Field (2009)[93], yet the distribution of the Initial Performance data was positively skewed and did not fully approximate a Normal distribution. Consequently, the behavioural measure was log-transformed and the analysis was re-run. However, the data remained slightly skewed and the residuals deviated somewhat from a straight line on the P-P plot. Therefore, the regression analysis was re-run without the BSRI and Pronoun measures (as they failed to significantly correlate with the predictor), yet some doubt remains over the reliability of the model.

The final model was a significant fit of the Initial Performance scores, F (7, 227) = 3.58, p = .001, but explained only 9.9% of variance in pre-training performance on the judgement task. Table 5.4 below shows the regression coefficients for both steps of the hierarchical regression.

[93] Note that for this and all subsequent analyses, these criteria include having little more than 5% of cases with standardised residuals outside of ±2, 1% outside ±2.5 and 0.1% outside ±3, the distribution of the criterion variable should approximate the normal distribution, residuals should fall very close to the straight line on the P-P plot, homoscedasticity should be investigated with a scatterplot (of the residual z scores vs. predicted z scores; there should be an absence of funnelling), a Durbin-Watson statistic close to 2, and VIF values well below the upper limit of 10 so as not to breach the assumption of multicollinearity.

Table 5.4. Analysis 1: Hierarchical multiple regression analysis: Initial Performance as the dependent variable.

	b	**SE *B***	**B**
Step 1			
Constant	0.06	0.01	
Exp2v1	0.01	0.02	.06
Exp3v1	0.00	0.02	.00
Exp4v1	0.02	0.02	.06
Exp5v1	-0.01	0.02	-.03
Exp6v1	0.01	0.02	.02
Step 2			
Constant	0.02	0.06	
Exp2v1	0.02	0.02	.08
Exp3v1	0.01	0.02	.03
Exp4v1	0.02	0.02	.10
Exp5v1	0.00	0.02	.01
Exp6v1	0.01	0.02	.04
ASI	0.03	0.01	.28***
IASNL	0.00	0.00	-.04

R^2 = .008 for step 1 (p > .8): ΔR^2 = .091 for step 2 (p < .001): ***p < .001.

Table 5 above indicates that the model improved significantly with the introduction of the two individual difference measures in Step 2. However, only the ASI was found to significantly predict accuracy of initial performance on the judgement task, $t(227) = 3.98$, $p < .001$.

Two final regression analyses were then run on these data with the dummy-coded experiment variables and the individual subscales of the ASI (separate analyses were run as both scales were highly correlated, $r = .538$, $p < .001$[94]). This data revealed that both the Hostile and Benevolent subscales contributed a very similar, significant amount to the regression model when analysed separately, (6.4% and 6.3% respectively[95]).

[94] Such separate analyses with the ASI subscales were conducted throughout this chapter in cases where the ASI emerged as a significant predictor of behavioural performance.
[95] Hostile subscale**:** $t(228) = 3.96$, $p < .001$; Benevolent subscale: $t(228) = 3.94$, $p < .001$.

Analysis 2; Initial Performance - Response Times

The response time data were examined following exactly the same procedure as the accuracy data. Again, for an initial assessment of the relationships between the predictor variables and the outcome measure, correlation coefficients are presented in Table 5.5 below.

Table 5.5. Analysis 2: Correlation coefficients of the nine predictors and Initial Performance.

	r
Initial Performance	
(Outcome variable)	1
Exp2vs1	.18**
Exp3vs1	.03
Exp4vs1	.12*
Exp5vs1	.08
Exp6vs1	-.47***
ASI	.11*
IASNL	-.10
BSRI	.04
Pronouns (Before)	.02

*$p < .05$, **$p < .01$, ***$p < .001$ (two-tailed)

In this analysis the ASI was the only individual difference measure to significantly correlate with the Initial Performance measure, $r = .111$, $p = .044$. This positive correlation suggests that those who scored higher in sexism had a greater gap between their RT performance on congruent and incongruent trials. It is also worth noting that the IASNL had a marginally significant correlation with the behavioural measure, $r = -.098$, $p = .067$, again suggesting that those with a less supportive attitude towards non-sexist language use tended to have a greater gap between RT performance on congruent and incongruent pairings.

Analysis of the regression output revealed that the assumptions of multiple regression analyses were largely met aside from a slight problem with homoscedasticity; two quite distinct groups of data points were evident on the scatterplot of the residual z scores vs. predicted z scores, but with the spread of data somewhat greater in one of these groups over the other.

However, given that all other assumptions of the multiple regression were fulfilled, it is likely that the model is a reliable fit of the data. The regression was re-run but with the ASI retained as the only individual difference predictor.

The final model was a significant fit of the Initial Performance scores, F (6, 228) = 12.56, p < .001, and explained 24.8% of variance in initial RT performance on the judgement task. Table 5.6 below shows the regression coefficients for both steps of the hierarchical regression.

Table 5.6. Analysis 2: Hierarchical multiple regression analysis: Initial Performance as the dependent variable.

	b	**SE *B***	**β**
Step 1			
Constant	0.26	0.04	
Exp2v1	0.06	0.06	.07
Exp3v1	-0.05	0.06	-.06
Exp4v1	0.01	0.06	.01
Exp5v1	-0.02	0.06	-.02
Exp6v1	-0.36	0.05	-.47***
Step 2			
Constant	0.16	0.06	
Exp2v1	0.07	0.06	.08
Exp3v1	-0.04	0.06	-.05
Exp4v1	0.02	0.06	.03
Exp5v1	-0.01	0.06	-.01
Exp6v1	-0.35	0.05	-.47***
ASI	0.05	0.02	.12*

R^2 = .234 for step 1 (p < .001): ΔR^2 = .014 for step 2 (p < .05). * p < .05, ***p < .001.

However, in the table above, it can be seen that the vast majority of the variance in Initial Performance scores is explained in Step 1 (23.4%), with just 1.4% explained with the addition of the ASI in Step 2. That said, it should be noted that this increase with the ASI is also statistically significant, t(228) = 2.07, p = .04. Two final regression analyses on the Hostile and Benevolent subscales revealed that the Benevolent subscale marginally contributed to the

model (+1.2% of additional variance explained, $t(228) = 1.91$, $p = .057$) but the Hostile subscale did not (+0.6%, $t(228) = 1.33$, $p > .18$).

Analysis 3; Performance Improvement - Accuracy

A hierarchical multiple regression analysis was next carried out on the Performance Improvement data. Again, for an initial assessment of the relationships between the predictor variables and the outcome measure, correlation coefficients are presented in Table 5.7 below.

Table 5.7. Analysis 3: Correlation coefficients of the nine predictors and Performance Improvement.

	r
Performance Improvement (Outcome variable)	1
Exp2vs1	-.10
Exp3vs1	-.03
Exp4vs1	-.03
Exp5vs1	-.15*
Exp6vs1	.03
ASI	.24**
IASNL	-.11
BSRI	.02
Total Pronouns	.02

*$p < .05$; **$p < .001$ (two-tailed)

The ASI and Performance Improvement were significantly positively correlated, $r = .244$, $p < .001$, with those scoring higher on the ASI (indicating increased sexism) tending to show a greater effect of training. This is likely due to the fact that those scoring lower in sexism had less scope for improvement across blocks than those scoring higher in sexism, as evidenced by a negative correlation between the ASI and Block 1 incongruent scores, $r = -.283$, $p < .001$[96].

[96] However there is a general issue with this Performance Improvement outcome variable in that it does not take account of the fact that, among participants who *do* show evidence of stereotyping in Block 1,

There was also a marginally significant negative correlation between the IASNL and Performance Improvement, r = -.106, p = .051. A scatterplot of these data suggests that those scoring lower in the IASNL (reflecting a less supportive attitude towards non-sexist language use) showed a greater effect of training, (again, likely due to the fact that they had more scope for improvement from Block 1 to Block 3 than those with a more positive attitude to sexist language use). This negative correlation is also unsurprising given that those with a less positive attitude to non-sexist language use should have more scope for improvement across blocks, evidenced by a positive correlation between the IASNL and Block 1 incongruent scores, r = .155, p = .017.

Analysis of the regression output revealed that some of the assumptions of multiple regression analyses were not adequately met (as outlined by Field, 2009). Firstly, the distribution of the Performance Improvement data did not fully approximate a normal distribution, while the residuals also deviated somewhat from a straight line on the P-P plot. Scatterplots (depicting standardized residuals against standardized predictive values) suggested that the assumption of homoscedasticity was not met, as evidenced by slight evidence of funnelling of the data points. Finally, the standardised residuals revealed a concern about outliers as 15 cases (6.36%) had standardised residuals falling outside of ±2, 10 of which had a standardised residual falling outside ±2.5 (4.24%), while 5 of these 10 cases had a standardised residual outside ±3 (2.12%), thus data points indicated an unacceptable level of error within the model. To address these violations, the 5 cases with a standard residual outside ±3 were removed as outliers and the analysis was re-run. A further 6 cases were again omitted according to this same outlier criterion, leaving a final data set of 224 points that better met each of the assumptions of multiple regression[97].

Correlations between the predictors and outcome variable were again examined, with the ASI and IASNL once more emerging as the only individual difference measures to significantly correlate with Performance Improvement (r = .285, p < .001 and r = -.159, p = .009, respectively). One final regression was therefore run with just these two predictors (and the dummy coded experiment variables).

It was found that the final model was a significant fit of Performance Improvement, $F(7, 216)$ = 4.13, p < .001, but explained only 11.8% of variance in Performance Improvement across

---

it is the less sexist of these people that would be more likely to be influenced by subsequent trainings, due to a higher motivation to overcome stereotype biases than those scoring higher in sexism.

[97] 5.8% of standardised residuals outside ±2, 3.5% outside ±2.5 and 1.3% outside ±3.

blocks. Table 5.8 below shows the regression coefficients for both steps of the hierarchical regression.

Table 5.8. Analysis 3: Hierarchical multiple regression analysis: Performance Improvement as the dependent variable.

	b	**SE *B***	**β**
Step 1			
Constant	0.07	0.02	
Exp2v1	-0.05	0.03	-.15
Exp3v1	-0.05	0.03	-.15
Exp4v1	-0.05	0.03	-.15
Exp5v1	-0.06	0.03	-.21**
Exp6v1	-0.01	0.02	-.02
Step 2			
Constant	0.01	0.07	
Exp2v1	-0.04	0.03	-.13
Exp3v1	-0.04	0.02	-.13
Exp4v1	-0.04	0.02	-.12
Exp5v1	-0.05	0.03	-.17*
Exp6v1	0.00	0.02	-.01
ASI	0.04	0.01	.25***
IASNL	0.00	0.00	-.02

R^2 = .049 for step 1 (p = .052): ΔR^2 = .069 for step 2 (p < .001); *p < .05, **p < .01, ***p < .001.

Ultimately, the ASI was the sole significant predictor of Performance Improvement (t(217) = 3.49, p = .001), as the IASNL failed to significantly predict this behavioural measure (t(217) = .31, p = .755).

Two final regression analyses were then run on these data with the dummy-coded experiment variables and the individual subscales of the ASI. In their respective analyses, both subscales made a significant contribution to the regression model, but the Hostile subscale explained a greater amount of the variance in Performance Improvement scores (6.7% vs. 3.9% respectively)[98].

[98] Hostile subscale**:** t(217) = 4.06, p < .001; Benevolent subscale: t(217) = 3.03, p = .003.

Analysis 4; Performance Improvement – Response times

A hierarchical multiple regression analysis was again carried out using forced entry, with the dummy-coded experiments and the individual difference measures included as predictors. The correlation coefficients between the predictor variables and Initial Performance are presented in Table 5.9 below.

Table 5.9. Analysis 1: Correlation coefficients of the nine predictors and Performance Improvement.

	r
Performance Improvement (Outcome variable)	1
Exp2vs1	.29***
Exp3vs1	-.02
Exp4vs1	.03
Exp5vs1	.01
Exp6vs1	-.27***
ASI	.11*
IASNL	-.11*
BSRI	.03
Total Pronouns	.09

*p < .05, ***p < .001 (two-tailed)

As with the accuracy data, both the ASI and the IASNL significantly correlated with the Performance Improvement measure, r = .108, p = .049 and r = -.111, p = .044, respectively. Those higher in sexism, and those with a less supportive attitude towards non-sexist language use showed a greater improvement in RTs across blocks given their greater scope for improvement than those scoring lower in sexism or with a more supportive attitude towards non-sexist language use.

Analysis of the regression output revealed that the assumptions of multiple regression analyses were adequately met.

One further regression was run with just the ASI and IASNL as predictors (along with the dummy coded experiment variables). The final model was a significant fit of the Performance

Improvement scores, F (7, 227) = 5.57, p < .001, and explained 14.7% of variance in Performance Improvement across blocks. However, upon further investigation, no significant improvement to the model from Step 1 was found (1.5% of additional variance was explained), with neither of the individual difference measures found to significantly predict response times on the behavioural measure[99], ASI: t(227) = 1.48, p = .140; IASNL: t(227) = .51, p = .61.

Nevertheless, two final regression analyses were run on these data with the dummy-coded experiment variables entered in Step 1 and either the Hostile or Benevolent subscale entered in Step 2. The Hostile subscale was found to marginally contribute to the regression model (1.3%, t(228) = 1.83, p = .069), however the Benevolent subscale did not (t(228) = 1.24, p = .217).

Overall, Analyses 1-4 have repeatedly demonstrated that the ASI is a consistent predictor of performance on the behavioural task. Also, while the IASNL repeatedly correlated with the behavioural measures, these results were weaker than those found with the ASI. Finally, the BSRI and Ethics Questionnaire proved ineffective at predicting accuracy or response time performance on the behavioural task. Reasons for this will be considered in Section 5.4.

Analysis of Experiments 4-6: BFoNE

In Experiments 4-6 (which all involved a form of social consensus feedback), participants were additionally administered the BFoNE so as to investigate whether scores on this measure correlate with or predict performance on the behavioural task. As a reminder, the BFoNE investigates the extent to which people experience anxiety at the prospect of being negatively evaluated by others, with high scores indicating a higher fear of negative evaluation relative to lower scores. In each of the subsequent analyses involving the BFoNE, hierarchical multiple regression analyses were carried out with the BFoNE as the sole individual difference predictor, combined with two dummy-coded experiment variables. Experiment 5 was chosen as the reference as this experiment was a control condition (Field, 2009), whereas experiments 4 and 6 involved stereotype reduction trainings. The experiment dummies were entered in Step 1 using forced entry, followed by the BFoNE in Step 2. Also, in the BFoNE analysis the regressions were initially conducted with all participants, and secondly with the data file split by gender. This latter analysis was conducted based on results from Chapter 3 indicating that

[99] The regression table for Step 1 is not presented here, as it is effects involving the individual difference measures that are of most theoretical interest in the chapter.

male participants were more sensitive to the social consensus feedback than females, as they displayed consistently better performance than their female counter-parts (in stark contrast to the findings of Chapter 2 involving performance feedback).

Analysis 5: BFoNE - Accuracy

To begin with, the correlation coefficients between the predictors and outcome variables were examined and are presented in Table 5.10 below.

Table 5.10. Analysis 5: Correlation coefficients of the BFoNE and the outcome variables of Initial Performance and Performance Improvement.

	r			***r***
Initial Performance			**Performance Improvement**	
(Outcome variable)	1		**(Outcome variable)**	1
Exp4vs5	.09		Exp4vs5	.02
Exp6vs5	.00		Exp6vs5	.16*
BFoNE	.02		BFoNE	.05

*$p < .05$ (two-tailed)

The correlation analysis revealed that the BFoNE failed to significantly correlate with either of the outcome variables. Despite this, the multiple regression analysis was run with the Initial Performance measure. However, the BFoNE failed to predict accuracy scores whether examined for all participants together (t (115) = .38, p = .708) or for males (t(115) = .16, p = .875) and females (t(115) = .47, p = .643) separately.

Next, with the Performance Improvement measure, the regression output revealed that the assumptions of multiple regression analyses were not adequately met, as 4 cases had standardised residuals falling outside ±3 (3.36%), again indicating an unacceptable level of error within the model (Field, 2009). These were removed and the regression analysis was re-run. However, the final model was not a significant fit of the Performance Improvement scores, F (3, 111) = 1.59, p = .196, nor did the BFoNE significantly predict Performance Improvement on the behavioural task, t(111) = 0.32, p = .752.

However, when the data file was split by gender, it was found that the BFoNE had a significant positive correlation with Performance Improvement for male participants (r = .302, p = .012), indicating that those scoring higher in fear of negative evaluation improved in accuracy to a greater extent across blocks than those scoring lower in fear of negative evaluation. There was no significant correlation with Performance Improvement for female participants (r = -.162, p = .111), yet it is worth noting that the data tended towards a negative correlation. Scatterplots of the accuracy data revealed that there was also some tendency for female participants who scored high on the BFoNE to improve to a greater extent across blocks than those who scored lower on the BFoNE. However, it appears a significant correlation failed to emerge because of a number of participants who had high BFoNE scores but whose accuracy scores deteriorated across blocks.

Overall, this pattern of results emerged irrespective of the fact that, on average, male participants achieved lower BFoNE scores than females (indicating less fear of negative evaluation, 35.9 vs. 42.4). The final regression model was a significant fit of Performance Improvement for male participants only, F (3, 52) = 3.08, p = .035, with the BFoNE explaining 11.1% of the variance in the males accuracy on this measure. This regression output is displayed in Table 5.11 below.

Table 5.11. Analysis 5: Hierarchical multiple regression analysis of male participants' data: Performance Improvement as the dependent variable.

	b	**SE *B***	**β**
Step 1			
Constant	0.03	0.03	
Exp4vs5	0.01	0.04	.02
Exp6vs5	0.04	0.03	.19
Step 2			
Constant	-0.10	0.05	
Exp4vs5	0.01	0.03	.05
Exp6vs5	0.04	0.03	.22
BFoNE	0.00	0.00	.34*

R^2 = .040 for step 1 (p = .342): ΔR^2 = .151 for step 2 (p = .012). * p < .05.

Analysis 6: BFoNE - Response times

As with the accuracy data, hierarchical multiple regression analyses were conducted to examine the response time data. The correlation coefficients between the predictors and outcome variables are presented in Table 5.12 below.

Table 5.12. Analysis 6: Correlation coefficients of the BFoNE and the outcome variables of Initial Performance and Performance Improvement.

	r			*r*
Initial Performance			**Performance Improvement**	
(Outcome variable)	1		**(Outcome variable)**	1
Exp4vs5	.39**		Exp4vs5	.19*
Exp6vs5	-.55***		Exp6vs5	-.31***
BFoNE	.00		BFoNE	-.05

*p < .05, **p < .01, ***p < .001 (two-tailed)

The correlation analysis revealed that the BFoNE once again failed to significantly correlate with either of the outcome variables.

Next, with Initial Performance and the BFoNE, the assumptions of multiple regression analyses were largely met (aside from slight problems of homoscedasticity as the scatterplot of residual z scores vs. predicted z scores revealed distinct groups of data points with varying levels of spread). However, while the final model was a significant fit of the Initial Performance scores, F (3, 115) = 16.52, p < .001, the BFoNE did not contribute to this as it failed to significantly predict Initial Performance on the behavioural task, t(115) = 0.08, p = .938. When the data file was split by gender, a very similar pattern of results was found, with the BFoNE failing to explain a greater amount of the variance in Initial Performance scores for males (t(54) = 1.50, p = .138) or females (t(57) = 1.64, p = .106) respectively).

Next, with Performance Improvement and the BFoNE, the assumptions of multiple regression were again largely met (aside from the same issue of homoscedasticity mentioned above). The final model was a significant fit of the Performance Improvement scores, F (3, 115) = 4.34, p = .006, yet the BFoNE did not contribute to this as it failed to significantly predict Performance Improvement on the behavioural task, t(115) = 0.57, p = .572.

However, when the data file was split by gender, it was found that the BFoNE had a significantly negative correlation with Performance Improvement for male participants (r = -.256, p = .026), yet a marginally positive correlation with Performance Improvement for female participants (r = .183, p = .079).

For the male participants, it appears that those scoring higher in fear of negative evaluation did not show much improvement in RT responding to stereotype incongruent pairings from Block 1 to Block 3. While this pattern could indicate a difficulty in processing the stereotyped role nouns, the correspondingly high accuracy scores of male participants in the judgement task suggest that they were taking longer to respond so as to reply correctly.

Conversely, for female participants, those scoring higher in fear of negative evaluation showed improvement in RT responding to stereotype incongruent pairings from Block 1 to Block 3. Although these faster response times could indicate greater ease with processing of the stereotyped role nouns, the relatively low accuracy scores of female participants relative to male participants in the judgement task suggest that, although they were faster to respond, they were making more mistakes.

The final regression model was a significant fit of Performance Improvement for male and female participants, F (3, 54) = 3.92, p = .013; F (3, 57) = 4.56, p = .006, respectively. The BFoNE was found to explain 9.7% of the variance in the male participants' response times on this behavioural measure and 6.0% of the variance in the female participants' response times. This regression output for both sexes is displayed in Table 5.13 below.

Table 5.13. Analysis 6: Hierarchical multiple regression analysis of male and female participant's data separately: RT Performance Improvement as the dependent variable.

		B	**SE *B***	**β**
Male participants	***Step 1***			
	Constant	0.08	0.10	
	Exp4vs5	.10	0.14	.11
	Exp6vs5	-0.17	0.13	-.21
	Step 2			
	Constant	0.53	0.20	
	Exp4vs5	.13	0.13	.14
	Exp6vs5	-0.19	0.12	-.24
	BFoNE	-0.01	0.01	-.32*
Female participants	***Step 1***			
	Constant	0.17	0.08	
	Exp4vs5	-0.05	0.11	-.07
	Exp6vs5	-.28	0.11	-.40**
	Step 2			
	Constant	-0.28	0.23	
	Exp4vs5	0.03	0.12	.04
	Exp6vs5	-0.28	0.10	-.39**
	BFoNE	0.01	0.01	.26*

Male Participants*:* R^2 = .082 for step 1 (p = .094): ΔR^2 = .097 for step 2 (p = .015). * $p < .05$

Female Participants*:* R^2 = .134 for step 1 (p = .015): ΔR^2 = .060 for step 2 (p = .045). * $p < .05$, ** $p < .01$

Ultimately, the BFoNE was a significant predictor of Performance Improvement scores for both sexes, albeit to a slightly greater extent for male participants, $t(54) = 2.52$, p = .015, than for females, $t(57) = 2.05$, p = .045.

Overall, the BFoNE scores of male participants shed some light on the reasons for which males consistently outperformed females in Chapter 3; males scoring higher on BFoNE tended to improve accuracy to a greater extent across blocks than those scoring lower on this scale and, although an equivalent improvement in RTs across blocks was not found, this suggests that they were taking longer to respond so as to ensure a correct answer. However, it is also worth

noting that males generally indicated lower levels of fear of negative evaluation than female participants. Female participants with high BFoNE scores tended to improve RT responding across blocks, yet showed poorer levels of accuracy.

5.3.4 Analysis of Experiments 7, 8 and 9: ASI

In Experiments 7, 8 and 9 (which involved association learning, counter-stereotypical pictures and stereotypical pictures respectively), participants were each administered the ASI so as to investigate whether scores on this sexism measure may correlate with or predict performance on the behavioural task. As a reminder, high scores on the ASI indicate a higher level of ambivalent sexism relative to lower scores. Data from 102 participants (45m, 57 f) were included in this analysis[100]. Hierarchical multiple regression analyses were again carried out, using forced entry. As before, the experiments were dummy-coded but now with Experiment 7 (association learning) as the reference as this experiment had the largest sample size (Field, 2009).

Analysis 7: ASI - Accuracy

To begin with, correlation coefficients between the predictor and outcome variables are presented in Table 5.14 below.

Table 5.14. Analysis 7: Correlation coefficients of the ASI and the outcome variables of Initial Performance and Performance Improvement.

	r			*r*
Initial Performance			**Performance Improvement**	
(Outcome variable)	1		**(Outcome variable)**	1
Exp8vs7	.23*		Exp8vs7	.31***
Exp9vs7	-.02		Exp9vs7	-.22*
ASI	.25**		ASI	-.08

*$p < .05$, ** $p < .01$, *** $p < .001$, (two-tailed)

[100] Data from one participant (from Experiment 7) was omitted as the questionnaire data was not recorded correctly.

The correlation analysis revealed a significant positive correlation between the ASI and the Initial Performance measure r = .251, p = .006 (with higher scores on the ASI correlating with lower accuracy on stereotype incongruent pairings relative to congruent pairings in Block 1). However, the ASI failed to significantly correlate with Performance Improvement, r = .08, p = .212).

With Initial Performance and the ASI, the assumptions of multiple regression analyses were largely met yet there were 9 cases with standardised residuals falling outside of ±2 (8.9%). Three of these were above ±3 and were removed as outliers. The analysis was then re-run and all regression assumptions were adequately met.

The final model was found to be a significant fit of the Initial Performance scores, F (3, 95) = 6.62, p < .001 explaining 17.3% of the variance therein. Table 5.15 below shows the regression coefficients for both steps of the hierarchical regression.

Table 5.15. Analysis 7: Hierarchical multiple regression analysis: Initial Performance as the dependent variable.

	b	**SE *B***	**β**
Step 1			
Constant	0.07	0.03	
Exp8v7	0.14	0.05	.30**
Exp9v7	0.07	0.05	.17
Step 2			
Constant	-0.11	0.06	
Exp8v7	0.14	0.05	.31**
Exp9v7	0.08	0.05	.18
ASI	0.09	0.03	.32***

R^2 = .073 for step 1 (p = .027): ΔR^2 = .100 for step 2 (p = .001). ** p < .01, ***p < .001.

Table 5.15– reveals that a much greater amount of variance in Initial Performance scores is explained in Step 2 (10%), with the ASI significantly contributing to the model, t(95) = 3.40, p = .001.

Two further regression analyses on this data revealed that both the Hostile and Benevolent subscales contributed a significant amount to the regression model, but with the Hostile

subscale explaining a much larger amount of the variance in Initial performance scores than the Benevolent subscale (10.3% vs. 4.8% respectively[101]).

Next, with the Performance Improvement and ASI analysis, the regression output revealed that the assumptions of multiple regression analyses were again largely met. However, two standardised residuals above 3 were removed and the analysis re-run. As with the Initial Performance measure described above, the final model was a significant fit of the Performance Improvement scores, F (3, 95) = 5.21, p = .002, yet the ASI did not contribute to this as it failed to significantly predict Performance Improvement on the judgement task, t(95) = 0.53, p = .601. Similarly, when two further regression analyses were run on this data with the Hostile and Benevolent subscales examined separately, neither were found to predict Performance Improvement scores, t(95) = .05, p = .96 and t(95) = .94, p = .35 respectively.

Analysis 8: ASI - Response Times

Next, the response time data of Experiments 7-9 was examined, with the correlation coefficients between the predictors and outcome variables presented in Table 5.16 below.

Table 5.16. Analysis 8: Correlation coefficients of the ASI and the outcome variables of Initial Performance and Performance Improvement.

	r			***r***
Initial Performance			**Performance Improvement**	
(Outcome variable)	1		**(Outcome variable)**	1
Exp8vs7	-.06		Exp8vs7	.01
Exp9vs7	-.27		Exp9vs7	-.28**
ASI	.02		ASI	.01

**p < .01 (two-tailed)

The correlation analysis showed that the ASI again failed to significantly correlate with either of the outcome variables.

Firstly, with the Initial Performance and ASI analysis, the regression output revealed that the assumptions of multiple regression analyses were largely met, yet 2 cases with particularly

[101] Hostile subscale**:** t(95) = 3.44, p = .001; Benevolent subscale: t(95) = 2.27, p = .026.

large standardised residuals (falling outside of ±3.6) were removed and the analysis re-run. However, while the final model was a significant fit of the Initial Performance scores, F (3, 95) = 5.60, p = .001, the ASI did not contribute to this fit, t(95) = 0.69, p = .495. Similarly, when two further regression analyses were run on this data with the Hostile and Benevolent subscales examined separately, neither were found to predict Initial Performance RT scores, t(95) = .28, p = .781 and t(95) = .89, p = .378 respectively.

Secondly, with Performance Improvement and the ASI, the regression output revealed that 2 cases had large standardised residuals (falling outside of ±3.3) and so these were removed from analysis and the regression re-run. The final model was again found to be a significant fit of the Performance Improvement scores, F (3, 95) = 3.67, p = .015, yet the ASI did not contribute to this as it failed to significantly predict Initial Performance on the behavioural task, t(95) = 0.51, p = .614. Finally, when two further regression analyses were run on this data with the Hostile and Benevolent subscales examined separately, neither were found to predict Performance Improvement RT scores, t(95) = .43, p = .661 and t(95) = .43, p = .671 respectively.

Overall, unlike analyses with the ASI in Experiments 1-6, this questionnaire was now *not* found to significantly predict performance on the behavioural measures of Initial Performance or Performance Improvement. This pattern of results may have resulted from a lack of statistical power, with less than half the number of participants in the current analysis. Also, as performance failed to improve significantly across blocks in Experiment 7 (association learning) and Experiment 9 (stereotypical imagery) there was generally less improvement across blocks in this sample than the earlier sample. This is likely to have affected the potential for ASI correlations with the Performance Improvement measure, due to a much narrower range of behavioural results.

Analysis of Experiment 7: MSS, BSRI, IAT

Unlike Experiments 8 and 9, participants in Experiment 7 were also administered three other individual difference measures: the BSRI, the MSS and the Gender-Career IAT. These will be examined in more detail below.

Analysis 9: MSS, BSRI, IAT - Accuracy

A multiple regression analysis was again carried out, with all individual difference measures entered together using forced entry. As these questionnaires were examined in Experiment 7 only, it was not necessary to include any dummy-coded experiment variables. The correlation coefficients between the three predictors and the two outcome variables are presented in Table 5.17 below.

Table 5.17. Analysis 9: Correlation coefficients of the MSS, BSRI, IAT and the outcome variables of Initial Performance and Performance Improvement.

	r			*r*
Initial Performance			**Performance Improvement**	
(Outcome variable)	1		**(Outcome variable)**	1
MSS	-0.22		MSS	-0.15
BSRI	0.08		BSRI	-0.03
IAT	-0.09		IAT	-0.03

The correlation analysis revealed that the three predictors failed to significantly correlate with either of the outcome variables.

Firstly, with Initial Performance, the regression output revealed that the assumptions of multiple regression analyses were largely met. However, one case with a standardised residual of 5 was removed (as it was well above the recommended limit of 3.2) and the analysis was re-run. One further case was found to have a standardised residual over 4 so this was again removed. With these changes, the MSS was now found to significantly correlate with Initial Performance (r = -.402, p = .008) while the BSRI was found to marginally correlate with it (r = .246, p = .077). The regression analysis was thus re-run with these two questionnaires only.

The model was now found to be a significant fit of the Initial Performance scores, F (2, 32) = 4.63, p = .017, with the two questionnaires explaining 22.4% of the variance in the outcome variable. However, it was found that only the MSS provided a significant contribution to the fit of the model, t(32) = 2.60, p = .014.

Next, with the Performance Improvement measure, one case with a standardised residual above 5 was found with the regression analysis. This was again removed and the analysis re-

run. Results revealed that the MSS marginally correlated with Performance Improvement, r = -.257, p = .065, while the other predictors did not. A final regression analysis was run with the MSS only, however, the final model was not significantly better than the mean model, F (1, 34) = 2.41, p = .130, thus suggesting that the MSS was not a significant predictor of Performance Improvement.

Analysis 10: MSS, BSRI, IAT - Response Times

Next, a multiple regression analysis was conducted as above to examine the response time data. The correlation coefficients between the three predictors and the two outcome variables are presented in Table 5.18 below.

Table 5.18. Analysis 10: Correlation coefficients of the MSS, BSRI, IAT and the outcome variables of Initial Performance and Performance Improvement.

	r			***r***
Initial Performance			**Performance Improvement**	
(Outcome variable)	1		**(Outcome variable)**	1
MSS	0.02		MSS	-0.22
BSRI	0.11		BSRI	0.03
IAT	-0.11		IAT	0.13

In both sets of correlational analyses, it was revealed that there was no significant correlation between the individual difference predictors and the outcome variables. However, regression analysis were run with all 3 predictors entered together.

Initial performance was first examined. Unsurprisingly, it was found that the final model was not significantly better than the mean model, F (3, 33) = 0.22, p = .880, and all three predictors failed to significantly contribute to the model. Similarly with the Performance Improvement measure, the model once again proved no better than the mean model F (3, 33) = 1.103, p = .399, and all three predictors again failed to significantly contribute to the model.

Overall, results of this section of analyses with Experiment 7 indicate that the MSS measure is worthy of further investigation as it was found to correlate with, and predict the behavioural measures, albeit inconsistently. In contrast, the BSRI and IAT results were disappointing yet this may be due to the fact that there was no significant improvement across blocks in

Experiment 7. As there was a narrow range of behavioural results, there was less chance of finding significant correlations with the individual difference measures. Similarly, as these analyses were conducted on data from just one experiment, the sample size was small and is likely to have reduced the chances of obtaining significant effects.

5.4 Chapter Discussion

Chapter 5 sought to comprehensively investigate a variety of individual difference measures, and their relationship with the judgement task used throughout this thesis.

Behavioural measures were selected that capture two interesting aspects of the judgement task data, namely Initial Performance (measuring stereotype congruent vs. stereotype incongruent performance in Block 1 before any training takes place) and Performance Improvement (measuring stereotype congruent and incongruent performance in Block 1 vs. stereotype congruent and incongruent performance in Block 3). Correlational and regression analysis were conducted between these behavioural measures and the individual difference measures, with some dominant patterns found to emerge.

Beginning with Initial Performance, the ASI and IASNL were consistently found to correlate with this behavioural measure. However, only the ASI was found to predict Initial Performance, and achieved this across both accuracy and response times. Also, in Experiment 7 (the only study in which the MSS was administered), the MSS was found to significantly correlate with and predict initial accuracy performance. Both the ASI and MSS are measures of sexism, with patterns of responding revealing that higher levels of sexism on these scales correlated with, and predicted, poorer performance on stereotype incongruent pairings relative to congruent pairings. In contrast, no significant effects surfaced between Initial Performance and the BFoNE (when examined with all participants together, or when split by gender), or with the remaining individual difference measures (IAT, BSRI, and the Ethics Questionnaire).

With Performance Improvement, the ASI and IASNL were again found to consistently correlate with the behavioural scores, and the ASI alone emerged as a significant predictor (this time in the accuracy data only). When Performance Improvement was first examined with the BFoNE, no significant correlations were revealed. However, when the data of male and female students was analysed separately, the BFoNE was found to significantly correlate with and predict Performance Improvement across accuracy and response times for male participants;

those scoring higher on the BFoNE took longer to respond to incongruent pairings in Block 3, yet were more accurate in their responding relative to Block 1. With the female participants, the BFoNE was found to correlate marginally with, and significantly predict Performance Improvement scores in the response time data only. Female participants with high BFoNE scores tended to respond faster across blocks, yet did not show correspondingly high levels of accuracy indicating that this fast performance may simply have been due to practice effects. Overall, this pattern of results is helpful in deciphering why male participants consistently showed superior performance over female participants in Chapter 3.

Next, while the MSS correlated with Performance Improvement in the accuracy data, this individual difference measure did not emerge as a significant predictor. Overall, when this MSS data is combined with that of the Initial Performance data, it is evident that this questionnaire could prove a useful individual difference measure to incorporate into future research, as it was not comprehensively investigated in this thesis. Indeed, as mentioned earlier, the MSS was previously found to moderate processing of gender-stereotyped role nouns in a sentence reading task (Gabriel et al., 2010).

Finally, no further significant relationships were found between Performance Improvement and the remaining individual difference measures (IAT, BSRI, and the Ethics Questionnaire).

In sum, there is little doubt that the ASI was the most consistent questionnaire to correlate with and predict performance on the behavioural measures. With the accuracy data, both the Hostile and Benevolent subscales were repeatedly found to emerge as significant predictors of behavioural performance, but the Hostile subscale the stronger predictor of the two. Only infrequent, marginal effects were found with the response time data. Overall it appears that people's opinions (particularly those concerning sexist antipathy) of men and women and their relationship in modern society are a window into predicting levels of stereotyping towards gender-biased role nouns in English.

Of interest with these results is the fact that the IASNL was consistently found to correlate with both behavioural measures but repeatedly failed to emerge as a significant predictor of task performance. However, correlations between the IASNL and behavioural measures were consistently lower than between the ASI and behavioural measures, thus effects with the IASNL were simply not strong enough to predict behavioural performance. Similarly, there is likely to be some overlapping variance between the ASI and IASNL; some of that variance is taken up by the ASI in a multiple regression making it harder to see what is already a much weaker effect of IASNL.

Overall, the results outlined above contribute to the growing literature on the effects of individual differences in stereotyping, with findings in line with those of past researchers who have noted the influence of individual differences on levels of stereotype use (e.g. Carter et al., 2006; Matheson & Kristiansen, 1987; Monteith, 1993). However, one point worth discussing is the size of the effects that were found in the current chapter. In cases where a questionnaire was found to significantly contribute to a regression model, it typically accounted for about 10% of variance in the behavioural measure data. Although some variance in responding could also be attributed to the experiment from which data was sourced, this clearly leaves a large amount of variance unaccounted for. One potential reason for this trend in results is that the chosen measures of prejudice and stereotyping may have had relatively little relevance to the behaviour under scrutiny in the judgement task. For instance, some of the questionnaires may have been tapping attitudes or behaviours that are much more specific (or indeed much broader) than the gender stereotyping behaviour tapped by the judgement task (Fishbein & Ajzen, 1975).

In this vein, Attitude Representation Theory (ART; Lord & Lepper, 1999) posits that when revealing attitudes towards a social category, a perceiver may have a particular exemplar in mind. Consequently, if this exemplar is not assessed in a subsequent behavioural measure, it is likely that there will be low attitude-behaviour correspondence; a process that is of particular importance when investigating behaviour to moderately typical and atypical group members (Lord, Lepper, & Mackie, 1984).

In the current research, a wide variety of individual difference measures were employed. However, as the behavioural task involved making judgements about gender-biased role nouns, it is arguably the Ethics Questionnaire that was most closely related to this task. Contrary to expectations, this individual difference measure consistently failed to emerge as a predictor of behavioural performance. The Ethics Questionnaire involved typing a response to an ethical dilemma, with levels of sexist pronoun use as the dependent variable. However many participants chose not to respond with a pronoun i.e. either began their response by fully repeating the role noun presented in the dilemma, or wrote short replies without a subject in the sentence. The Ethics Questionnaire is also an infrequently used individual difference measure and many items were added to the scale for the purposes of the current research, thus potentially changing the nature of the scale. It is possible that the above considerations may have contributed to the lack of significant findings between the Ethics Questionnaire and either of the behavioural measures.

The remaining individual difference measures purported to tap levels of sexism, attitudes to non-sexist language use, sex-role perception, implicit gender-career associations and fear of negative evaluation. Although each of these relate to the issue of gender (aside from the BFoNE), it is clear that they each seek to measure much broader concepts than the judgement task which examines immediate gender inferences in response to stereotyped role terms. While the IAT may be similar to the judgement task in some respects (i.e. they both involve fast responding to gender-related stimuli), this test was used with Experiment 7 only (due to a lack of significant findings). In hindsight, it may have proven beneficial to also incorporate the IAT with some of the other experiments, specifically those that revealed a definite change in behavioural performance across blocks (so as to increase the chances of finding effects). This may have been particularly worthwhile as implicit measures are thought to better predict spontaneous, automatic behaviour and non-conscious responding while explicit measures better predict more controlled, deliberative behaviour and conscious responding (see Maio, Haddock, Manstead, & Spears, 2010 for a discussion). Indeed, this distinction between automatic and controlled processing is somewhat problematic for the behavioural task used throughout this thesis.

Although the judgement task does not assess purely automatic, non-conscious responding, it *does* measure relatively spontaneous reactions, with participants typically making a response about one second after presentation of a target. It therefore appears that the judgement task measures effects that are at the boundary of what both implicit and explicit individual difference measures purport to optimally assess i.e. spontaneous reactions as opposed to purely automatic vs. controlled processing. Thus, it seems likely that neither category of implicit or explicit individual difference measures could adequately capture individual variance in responding to the behavioural judgement task.

With this in mind, it may prove more beneficial for future research to employ implicit individual difference measures in conjunction with behavioural tasks that measure implicit responding to stereotyped role nouns, and explicit individual difference measures in conjunction with behavioural tasks that measure explicit responding to stereotyped role nouns. In this way perhaps a more accurate picture of the role of individual differences in such gender processing could be drawn. That said, an interesting line of enquiry would also be an investigation into the crossed combinations of implicit and explicit tasks and attitudinal measures so as to gain a more comprehensive understanding of how such processes are related.

More generally, it is apparent that the vast majority of individual difference measures used in this chapter rely on self-report scales in which participants indicate their agreement with a particular attitude statement. While such measures are certainly useful as regards gaining insight into the surface attitudes of participants about a variety of social groups, results are not always interpretable in a straightforward manner as many factors (e.g. social norm biases, participants being unaware of their true attitudes) can influence responding (Maio et al., 2010). Such factors may be another reason for which a lack of significant effects were observed between some of the individual difference and behavioural measures, and is further reason to employ more implicit individual difference measures in future research.

In conclusion, the current research has established that individual differences (particularly in levels of ambivalent sexism) can influence response to stereotype incongruent gender information in English. It is hoped that results and observations from this chapter may inform the selection of individual difference measures for future studies on the topic of gender stereotype use. Moreover, as with past findings in which individual differences have been found to influence human perception and behaviour, the above results promote movement of research away from a deterministic outlook that places emphasis on the role of contextual influences in judgement, to a view with greater emphasis on the values, needs, goals and motives that a perceiver carries into the social perception process (Moskowitz, 1993).

6. General Discussion

6.1 Introduction

This thesis set out to devise and explore training strategies aimed at overcoming the activation of gender stereotypes in English so as to result in lower levels of stereotype application. Past research has shown relative success in this endeavour, for instance through explicit mention of the sex of a referent in a text (e.g. Duffy & Keir, 2004; Kreiner et al., 2008; Lassonde, 2013), mental imagery (Blair et al., 2001), counter-stereotype affirmation training (e.g. Gawronski et al., 2008; Kawakami et al., 2007), and explicit reminders about gender-biased occupations (Oakhill et al., 2005). However, while stereotyping has previously been reduced, these effects are typically short-lived and stereotyping is rarely found to disappear completely. Recent reviews of the field of stereotype reduction have highlighted a number of concerns over past training regimes and suggested directions for further research. The training methods developed within this thesis sought to build on these guidelines and advance current research on gender stereotype reduction.

One of the fields in which stereotype reduction has been studied is psycholinguistics. In the field, the focus has been on reduction of stereotyping of people described by occupational role nouns. This has typically been investigated in text comprehension using sentence reading tasks (e.g. Duffy & Keir, 2004; Kreiner et al., 2008; Lassonde, 2013). An exception to this was observed in the work of Oakhill et al. (2005) who successfully lowered levels of stereotype use after explicitly reminding participants that certain occupational roles can now be fulfilled by either sex. The judgement task of Oakhill et al. (2005) was chosen as a means of stereotype assessment in the current research so as to continue in this line of enquiry and try to identify further means of stereotype reduction in response to single, gender-biased occupational terms.

This final chapter will begin with an overview of the main findings from the experimental work of this thesis. The theoretical implications of this research will then be considered, followed by a brief discussion of some study limitations, and recommendations for future research in the domain of stereotype reduction.

6.2 Main findings: A summary

Across a series of nine experiments, a number of stereotype reduction strategies was investigated using the judgement task of Oakhill and colleagues (2005). Using this judgement

task, an initial assessment of stereotype application was taken before the introduction of a stereotype reduction training method. In this way it was possible to (a) investigate whether the training method led to reduced levels of stereotype use and (b) evaluate the efficacy of one training relative to another. For ease of reference, the accuracy and RT performance to critical stereotype incongruent trials across blocks in each experiment is provided in Table 6.1 and 6.2 below respectively.

Table 6.1 Accuracy (%) performance across blocks in Experiments 1-9

Exp	Experiment Title	Block 1	Block 2	Block 3	Block 4	Improvement
1	Performance Fb	81.62	91.34	96.08	--	14.46%
2	Control (no Fb)	77.50	80.27	78.33	--	0.83%
3	Performance Fb No. 2	81.49	89.47	87.97	90.86	9.37%
4	Social Consensus Fb	76.50	79.40	82.64	--	6.14%
5	Reverse consensus Fb	79.79	81.43	81.68	--	1.89%
6	Social & Accuracy Fb	79.09	84.92	88.42	--	9.33%
7	Association Learning	88.16	90.57	--	--	2.41%
8	Counter-Ster. Pictures	74.86	84.73	--	--	9.87%
9	Stereotype Pictures	83.21	83.33	--	--	0.12%

Table 6.2 Response time (ms) performance across blocks in Experiments 1-9

Exp	Experiment Title	Block 1	Block 2	Block 3	Block 4	Improvement
1	Performance Fb	1120	843	713	--	407ms[102]
2	Control (no Fb)	1114	870	864	--	250ms
3	Performance Fb No. 2	1140	921	813	804	336ms
4	Social Consensus Fb	1057	1051	794	--	263ms
5	Reverse consensus Fb	999	879	749	--	250ms
6	Social & Accuracy Fb	1018	945	712	--	306ms
7	Association Learning	848	864	--	--	-16ms
8	Counter-Ster. Pictures	1032	807	--	--	225ms
9	Stereotype Pictures	979	861	--	--	118ms

[102] Note that positive numbers in this column indicate performance improvement, negative numbers indicate performance deterioration.

Experiments 1-3.

Chapter 2 investigated the use of performance-related feedback as a stereotype reduction strategy. This feedback involved telling participants whether they were correct or incorrect (about whether, for example, *sister* and *electrician* can refer to the same person) after each of their Block 2 responses, and providing them with their cumulative percentage score. It was found that providing participants with such direct, repetitive and accurate feedback on their performance helped to reduce levels of stereotyping immediately following the training (Experiment 1), but also to a novel set of stimuli (although with somewhat limited success, Experiment 3), and with effects still evident one week later (Experiment 3). In contrast, no improvement in accuracy of responding was found in Experiment 2; a control study in which feedback was not given to participants. Performance-related feedback was thus proven to be a highly successful means of stereotype reduction. Moreover, while the majority of stereotype reduction studies investigate only the immediate effects of trainings (Lenton et al., 2009; Paluck & Green, 2009), these studies demonstrated the value of further verifying that training successfully extends to newly introduced stimuli and also assessing the durability of results. It is clear that only through such stringent testing of strategies will a truly useful means of stereotype reduction be identified and a shift away from research on 'quick fix' methods can commence.

Experiments 4-6.

Chapter 3 sought to explore the use of social consensus information as a stereotype reduction strategy. Although this feedback has previously proved successful in reducing stereotyping against those suffering from obesity (Puhl et al., 2005) and racial stereotyping (Stangor et al., 2001), its effects had remained untested in the field of gender stereotyping. Specifically, participants were provided with fictitious social consensus feedback (ostensibly from their peers) which sought to convey that stereotype endorsement was extremely infrequent among this group (Experiment 4). In this way, it was hypothesised that participants would adapt their behaviour to bring their responding in line with the perceived responding of their peer group. It was revealed that participants did indeed reduce stereotyping following presentation of the social feedback. Experiment 5 was designed as a control study aimed at establishing the mechanism(s) through which the social consensus feedback operated. The results suggested that this strategy relies on social compliance mechanisms (as opposed to simply alerting participants to the issue of stereotype biases) in order to successfully reduce stereotyping. Finally, Experiment 6 combined both accuracy and social consensus feedback in an effort to

determine whether both sources of information would lead to higher levels of stereotype reduction than one source of feedback. Although trends in the data supported the use of both sources of information (as opposed to social consensus feedback alone), no consistent significant differences were found between the two experiments in response to stereotype incongruent word pairs.

The findings from Chapter 3 suggest that social consensus information is useful means of stereotype reduction, yet when the results of Experiment 4 were compared with those of Experiment 1, it was established that this consensus feedback does not achieve the same level of success as providing participants with more straight-forward, performance-related feedback. It is possible that the former strategy may prove most useful in contexts in which participants are likely to be particularly concerned with the opinions or attitudes of others, while the latter strategy may have potential as a more general stereotype reduction training method in a broader array of contexts.

Experiments 7-9.

A different approach to stereotype reduction was examined in Chapter 4; the strengthening of counter-stereotype associations in the cognitive network. As mentioned earlier, the use of counter-stereotype information as a stereotype reduction strategy has previously met with some success (e.g. Blair et al., 2001; Kawakami et al., 2001). However, debate persists over whether presenting participants with such information does indeed lead to stereotype reduction (through increased variability and fragmentation in the perceivers' stereotype representations) or to stereotype maintenance (through subtyping processes in which counter-stereotype information is grouped together and explained away as a type of 'exception to the rule').

Experiment 7 employed an association learning paradigm in which it was hypothesised that getting participants to learn gender incongruent pairings would alert them to the fact that men and women can occupy gender atypical occupational roles (without explicitly pointing this out). This hypothesis was not supported as stereotyping was still evident following the learning task. However, due to some design modifications with this study (greater number of trials across fewer blocks, questionnaires after the judgement task as opposed to before) relative to the previous studies, it is not fully clear whether Block 1 performance (which was unusually high) masked any improvement that association learning paradigm might bring about, or whether the training was simply too subtle to induce change.

The stereotype reduction strategy of Experiment 8 made more direct use of counter-stereotypes than Experiment 7, as participants were presented with (and answered questions about) striking pictures of men and women working in counter-stereotypical roles. Responses to stereotype incongruent pairings were found to significantly improve after this picture task while results of Experiment 9 (a control study in which pictures of people working in gender stereotypic roles were shown to participants) confirmed this improvement was indeed due to the picture manipulation. Findings were again complicated by the fact that Block 1 performance was lower in Experiment 8 than in Experiment 9. Thus, despite a significant improvement across blocks in the former study only, final accuracy scores were quite similar in both experiments. Although these results should be interpreted with some caution, on the whole the findings of Chapter 4 suggest that counter-stereotype information leads to stereotype reduction as opposed to stereotype maintenance. Finally, the use of striking pictures as part of the counter-stereotype training in Experiment 8 provides support for the conversion theory of stereotype change, i.e. that stereotypes are likely to change quickly upon encountering few, yet striking, counter-stereotype exemplars.

In Experiments 8 and 9, data was also collected on the broader gender stereotypes that people hold in relation to occupations and lifestyle. This was achieved by asking participants a series of questions about each of the pictures they were presented with as part of the behavioural training. These data highlight some interesting trends regarding the perception of men and women working in stereotypical and counter-stereotypical occupational roles (see Chapter 4 for further details).

Chapter 5: Individual differences

An investigation of Individual differences in stereotype reduction has typically been something of an afterthought in psychology research, with conflicting evidence on the influence of this variable on task performance. In an attempt to address this issue, individual difference measures were consistently administered across all studies in this thesis in conjunction with the behavioural tasks (with the results of accompanying correlational and regression analyses outlined in Chapter 5).

Contrary to expectations, the findings of the current research proved somewhat underwhelming as relatively small amounts of variance on the behavioural task were accounted for by the individual difference measures used. That said, some important observations were made: (1) the Ambivalent Sexism Inventory stood out as a measure worthy of inclusion in future gender-related research, as it consistently captured a significant amount

of variance in task responding, (2) the Modern Sexism Scale and gender-career Implicit Association Test were both found to warrant further investigation (as they were only administered in Experiment 7 and it is likely that significant effects went undetected due to a lack of improvement in responding across blocks), and (3) scores on the Brief Fear of Negative Evaluation scale (BFoNE) provided some insight into the unusually good performance of male participants in Chapter 3; a number of male participants who scored particularly high on fear of negative evaluation also greatly improved their accuracy performance across blocks on the judgement task, thus suggesting they were particularly influenced by the social feedback (although it is worth noting that males still scored lower on fear of negative evaluation than females did overall).

Experiments 1-9: Participant Gender

Finally, although sex differences in performance were not anticipated in this thesis (based on the previous findings of Oakhill et al., 2005), differential performance between the sexes *was* consistently noted across the chapters. That said, the detailed effects of Participant Gender were different in different studies. For instance, females consistently outperformed males in Chapter 2 (performance feedback and control). This trend was then reversed in Chapter 3 (social consensus feedback, reverse social consensus feedback and combined social and accuracy feedback), while Chapter 4 provided more mixed results (females outperformed males in Experiment 7 and 9 yet males responded with much greater accuracy in Experiment 8). Thus, Participant Gender interacted differentially with levels of stereotyping overall, depending on the chapter in question.

A recurrent pattern in participants' responses was the tendency for female participants to respond faster and more accurately to female kinship/stereotyped terms over male terms while male participants tended to respond faster and more accurately to male kinship/stereotyped terms over female terms (as evidenced by significant interactions of Participant Gender by Kinship term gender or Participant Gender by Stereotype bias). These interactions suggest a facilitation when responding to stimuli congruent with a participant's own sex.

Overall the findings in relation to Participant Gender are highly ambiguous, with no clear reasons identified for the variable effects observed with this factor across chapters. Future research could further explore the reasons underlying this pattern of results in an attempt to fully understand the occurrence of sex differences in processing of gender stereotypes.

6.3 Theoretical implications

The general discussion sections of Chapters 2-5 explored some theoretical implications of the experimental data presented in the respective chapters. The broader implications of these results in relation to the processing of gender stereotype biases and the field of stereotype reduction will now be discussed.

As mentioned in Chapter 1**,** the question of when gender stereotypes are activated in online processing has been of interest to researchers of late (Duffy & Keir, 2004; Irmen, 2007; Pyykkönen et al., 2010; Reynolds, et al., 2006). The findings from this thesis contribute to this discussion, with results providing support for the elaborative activation of gender stereotype biases i.e. that gender stereotypes are activated when they are not needed to understand a text or to increase text coherence. In the current studies it was found that such gender biases were activated in response to single words even though they were detrimental to performance on the judgement task (findings that replicate those previously found by Oakhill et al., 2005). Correspondingly, these results fail to provide support for the minimalist account of text comprehension (McKoon & Ratcliff, 1986; 1992), which predicts that elaborative gender inferences will not be made online if they are not required for local coherence or based on easily available (i.e. explicitly stated) information. Although stereotype activation in the judgement task used in these studies does not directly reflect what happens in normal text processing, the current findings lend some support to past claims that stereotype biases are used in a forward, elaborative manner.

Overall the studies in this thesis provide support for the use of explicit stereotype reduction strategies (as opposed to implicit) and the claim of Bargh (1994) that awareness of the automatic influence of a stimulus is a crucial factor in the subsequent purposeful control of thought and behaviour. Once they are made aware of the biases influencing their behaviour, participants can form intentional strategies to overcome and control them if they have sufficient motivation and attentional resources. Each of the successful stereotype reduction manipulations in this research (performance feedback, social consensus feedback, combined social and accuracy feedback and the use of counter-stereotypical pictures) relied on striking, explicit training strategies in order to reduce stereotyping, while Experiment 7 (in which a relatively subtle association learning paradigm was used) failed to effect stereotype change. It is possible that participants were simply not aware of succumbing to gender biases in that latter study and thus did not expend any additional cognitive effort in trying to control or overcome them. In contrast, participants who received an explicit training strategy seem to

have been reminded that stereotypes are maladaptive forms of categories (in that their content is not always accurate) and became motivated to respond without the influence of these stereotypes. Indeed, at their simplest, it seems that the explicit training strategies used in this thesis simply reminded participants of things they already knew e.g. that a woman can be a surgeon and a man can be a nurse. The fact that the findings from these (and other) explicit training methods can be distilled down to such a simple statement has broad implications for stereotype reduction at a societal level - it appears that simply making counter-stereotypes much more visible in society (i.e. increasing exposure to counter-stereotype material or exemplars) could instigate real change in the cognitive representations of gender-biased terms (either through bookkeeping or conversion processes, or indeed a combination of both - these processes have been described in Chapter 1).

In this vein, the use of pictures as a stereotype reduction manipulation is an example of a training strategy that could be easily applied at a broad, societal level so as to increase exposure to counter-stereotypes. For instance, it seems likely that frequent depiction of men and women working in gender atypical roles in educational material would effect change in students' cognitive representations of gender to accommodate this information[103]. Such gender-fair pictures could also be used in other gender-related contexts when occupational stereotypes may be in use (e.g. with certain job adverts). Ultimately, with such exposure, counter-stereotypic associations should become more accessible and the issue of gender 'atypical' roles may become obsolete. If true, this approach shows much promise for inducing long-term stereotype change and could, with time, result in people delaying the assignment of gender to a referent when gender-biased occupational terms are encountered (as they hold back until more definitive gender information is supplied).

The results of Experiments 1-9 are also in line with the meta-analytic findings of Lenton et al. (2009) who posit that a significant moderator of the reduction of automatic gender stereotypes is the intervention method used. As a reminder, they established that interventions based on either (A) distracting or redirecting participant's attention before category activation, or (B) facilitating the holding of multiple, different representations within the activated stereotype were typically more successful than those based on (C) stereotype prevention or inhibiting expression of stereotypes. As the first type of intervention did not fit with the main focus of this thesis (i.e. overcoming stereotype application *after* stereotype

[103] Although some studies report that school books have become more gender fair across recent years (e.g. Diekman & Murnen, 2004; Moser & Hannover, 2013), effects of these changes on longer term cognitive representations of gender remain unknown.

activation had already occurred through presentation of a gender-biased role noun) only the latter two types of intervention were examined herein.

Indeed, category B and C interventions were interlinked in the current work. For instance strategies were devised that aimed at creating awareness of category heterogeneity e.g. through feedback or pictures (category B above), yet the judgement task itself involved inhibition or suppression of the stereotype bias so as to result in lower levels of stereotype use (category C above). Therefore, while the findings of this research do not provide conclusive evidence on the most successful type of intervention method for stereotype reduction, they at least provide some support for the use of suppression-based interventions in research. Such interventions have previously been identified as relatively unsuccessful because of rebound effects in which the stereotype becomes hyper-accessible and returns more insistently after intentional suppression (e.g. Macrae et al., 1994). Although longer term effects of the training methods used in this thesis were only examined in Experiment 3, no evidence of rebound effects was found at this stage.

Of interest is the fact that, in the work of Macrae and colleagues (1994), participants were explicitly told to avoid stereotypes (about skinheads), whereas suppression occurred without such instruction in the current studies. Instead, participants were motivated to suppress stereotypes spontaneously, so as to perform better on the judgement task. This difference in task designs may be a reason for the contrasting conclusions on the use of suppression in these two pieces of research. Indeed, past research by Monteith and colleagues (1998) revealed that participants with a motivation or desire to avoid stereotypic thinking were able to achieve stereotype reduction without the occurrence of subsequent rebound effects or heightened stereotype accessibility.

Although the results of this thesis provide strong support for the malleability of gender stereotype biases (as a number of successful stereotype reduction strategies have been identified), the results are also consistent with past research documenting the persistency of stereotyping effects (e.g. Dunning & Sherman, 1997; Oakhill et al., 2005; Reynolds et al., 2006). There remains much scope for future research on this topic, as the processing of stereotype incongruent pairings never quite achieved the same level of effortlessly fast and accurate responding as that of stereotype congruent and neutral pairings. This same level of success (or lack of complete success) was achieved by Oakhill et al. (2005) when they originally used this judgement paradigm (i.e. responding to stereotype incongruent pairings was consistently slower and less accurate than responding to trials in the other conditions). But how does

performance on the incongruent trials of Oakhill and colleagues relate to those found in the current thesis?

As described in Chapter 1, Oakhill et al. (2005) conducted six studies in which aspects of stimuli presentation were modified (Experiment 1, 2, 3, 5, 6)[104] or participants were explicitly reminded that most professions can now be occupied by both men and women (Experiment 4). Two of their studies stood out as particularly successful in terms of improving responding to stereotype incongruent trials; Experiment 3 (93% accuracy/670ms RTs) and Experiment 4 (92.3% accuracy/635ms RTs).

These results are not particularly surprising as, in Experiment 3, the role term was presented for a relatively long time before the kinship term was presented (1,800ms), thus allowing for conscious reflection on responding and, in Experiment 4, participants were provided with an explicit performance strategy to help them in the judgement task. In comparison, Experiment 1 in this thesis (in which performance-related feedback was offered to the participants) was the only study to find a higher level of accuracy (96.08%) than that achieved by Oakhill et al., while response times were not improved upon. While such cross-study comparison is useful in terms of investigating whether past results were broadly replicated, it should be noted that subtle differences in design and procedure between the original and current experiments limit the conclusions that can be drawn from such comparisons.

Finally, to return briefly to the issue of stereotype persistence, one reason for the endurance of stereotype effects may be that certain design limitations adversely influenced the reduction of stereotyping in the current research. Although various limitations have been mentioned in earlier discussion sections, some of the broader methodological concerns pertaining to the present work, and related suggestions for future research, will now be outlined.

6.4 Methodological limitations and future research

In this thesis the judgement task of Oakhill et al. (2005) was consistently used as the behavioural measure of stereotyping so that the success of the different training strategies

[104] Modifications included: (Experiment 1) kinship and role terms were presented on-screen together, (Experiment 2) kinship and role terms were presented separately with the role term shown first (for 500ms), (Experiment 3) stimuli were presented as in Experiment 2 but with the role term now shown for 1800ms and participants were asked about their response strategies after the study, (Experiment 5) role terms were presented first for 500ms then kinship terms remained for just 600ms before disappearing and (Experiment 6) the opposite pattern i.e. kinship terms were presented first for 500ms then the role nouns remained for 600ms.

could be readily compared. If different stereotyping measures had been used, the factors underlying particularly good or poor performance would have been difficult to establish i.e. reasons could have been task-related or training-related. However, one cause for concern with this judgement task was the variable Block 1 performance found across experiments. As studies were extremely similar up until the completion of Block 1 (at which point a training phase would typically commence), the reasons for such variable performance remain unknown. It is possible that a different behavioural measure or simple design changes in the judgement task may have resulted in more consistent results across initial blocks. For instance, forcing participants to respond within a short time window would have constrained response times, and possibly also helped to standardise accuracy scores (due to less scope for explicit reflection on responding). However, such hypotheses could only be verified with further experimentation – it is of course also possible that individual differences or minor alterations in the experimental instructions of studies 1-9 differentially influenced participants' expectations and perception of the behavioural task across experiments, thus resulting in more variable performance than was anticipated by the experimenter.

A concern highlighted in past research is the over-reliance on single session, laboratory-based studies in the stereotype reduction literature. This research trend is problematic for two reasons (1) from a single session, it is not possible to distinguish between a temporary change in behaviour and a longer-lasting change in underlying representations and (2) interventions are often detached from a 'real-life' context thus casting doubt on their usefulness beyond a laboratory setting (Lenton et al., 2009; Paluck & Green, 2009). An attempt was made to address the first of these two issues in Experiment 3 (in which stereotype reduction was observed one week after the initial training). While the reduced levels of stereotyping suggest that there was a real change in the underlying stereotype, in fact, it cannot be confirmed that this was the case. Participants may simply have learned how to respond correctly in this highly specific experimental context, yet may still fail to overcome gender stereotype biases in different environments. Indeed, the same may be said for each of the stereotype reduction strategies used in this thesis. Future research aimed at investigating transfer of successful training to new contexts would help to establish the true durability and stability of stereotype change and could simultaneously address issue 2 mentioned above by trying to confirm the value of the training in a broader, ecologically valid setting.

Potential ways in which interventions from the current work could be applied to new contexts include testing whether these training methods can assist in (1) overcoming gender biases in text processing in more natural discourse processing tasks or (2) reducing stereotyping when

people are faced with other gender-related decisions such as evaluating C.V.'s or hiring people for jobs. The strategies also could be extended to related fields of research and their effects on the reduction of stereotyping against other social groups (e.g. black and ethnic minorities, the elderly, those suffering from obesity etc.) could be investigated. Indeed, past training strategies that have reduced stereotyping against one particular social group have also shown success with other groups. For example, social consensus feedback has been used to reduce weight-related and race-related stereotyping, while the stereotype negation training of Kawakami and colleagues (2000; 2007) has been found to reduce, gender-, race- and skin head-related stereotypes. Such transfer of interventions across different stereotyped groups is a logical and promising route for future research with the effective interventions developed in this thesis. Ideally, a training method would be identified that induces stereotype reduction towards a variety of social groups and can thus be widely applied in research.

Finally, given the vast increase in application of cognitive neuroscience methodologies to psychological research over the past decade or so, it seems highly likely that applying such tools (e.g. eye-tracking, EEG or fMRI) to the study of stereotype reduction could prove an extremely promising line of research. While these methodologies have previously been used to examine activation of gender stereotypes (e.g. Irmen, 2007; Knutson, Mah, Manly, & Grafman, 2007; Osterhout et al., 1997) little research has investigated what happens 'online' when stereotype reduction occurs. Future studies using such methodologies would permit more fine-grained analyses of the time course and biological underpinnings of stereotype activation, use *and* stereotype reduction.

6.5 Final conclusion

The primary aim of this thesis was to identify strategies aimed at overcoming the activation of the gender stereotype biases that are associated with many occupational role nouns in English. A secondary goal was to fill gaps in the established research literature e.g. by (1) examining the durability of the effects and their extension to novel stimuli, (2) taking advantage of existing knowledge on the power of conformity and social norms, (3) tackling the over-reliance on verbal stimuli in past research, and (4) conducting a thorough investigation into individual differences that may influence levels of stereotyping. Overall a number of effective interventions were identified, thus providing further evidence for the malleability of stereotypical gender biases, given appropriate strategies and conditions.

Future research should continue in its endeavor to isolate the individual differences in stereotyping, and the interventions and contexts that best facilitate the goal of stereotype reduction. Once the individual influence of each of these variables has been established, then a theory-driven examination of the inter-relations between them can begin. In this way, a much clearer depiction of how to reduce gender biases in response to occupational role terms, and gender stereotyping more generally, could be achieved.

Reference List

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracín, B. T., Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173–222). Mahwah, NJ: Erlbaum.

Allport, G. W. (1935). Attitudes. In C. Murchinson (Ed.), *A handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Altemeyer, B. (1996). *The authoritarian specter*. Cambridge, MA: Harvard University Press.

Anderson, J. R., & Bower, G.H. (1973). *Human Associative Memory*. Washington D.C.: V.H.Winston.

Anderson, N. H. (1981). *Foundations of information integration theory: Concepts and applications*. Chichester, UK: Wiley.

Anglin, J. M. (1977). *Word, Object, and Conceptual Development*, New York: Norton.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70.

Ashburn-Nardo, L., Voils, C. I., & Monteith, M. J. (2001). Implicit associations as the seeds of intergroup bias: How easily do they take root? *Journal of Personality and Social Psychology*, *81*(5), 789-799.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*(3), 136–141.

Banaji, M. R., Hardin, C. D., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, *65*(2), 272-281.

Bargh, J. A. (1990). Goal and intent: Goal-directed thought and behavior are often unintentional. *Psychological Inquiry*, *1*(3), 248–251.

Bargh, J. A. (1992). Does subliminality matter to social psychology? Being aware of the stimulus versus aware of its influence. In R. F. Bornstein & T. Pittman (Eds.), *Perception without awareness* (pp. 236-255). New York: Guilford Press.

Bargh J. A. (1994). The four horsemen of automaticity: awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1–40). Hillsdale, NJ: Erlbaum.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). New York: Guilford Press.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230-244.

Bargh, J. A., & Gollwitzer, P. M. (1994). Environmental control of goal-directed actions and behavior. In W. Spaulding (Ed.), *Integrations of motivation and cognition: The Nebraska Symposium on Motivation* (Vol. 41, pp 71-124). Lincoln: University of Nebraska press.

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

Bem, S. L., & Lenney, E. (1976). Sex typing and the avoidance of cross-sex behavior. *Journal of Personality and Social Psychology*, *33*(1), 48-54.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*(3), 242–261.

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology*, *70*(6), 1142-1163.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828-841.

Blanton, H., & Christie, C. (2003). Deviance regulation: A theory of action and identity. *Review of General Psychology*, *7*(2), 115-149.

Bodenhausen, G.V., & Macrae, C.N. (1998). Stereotype activation and inhibition. In R. S. Wyer Jr. (Ed.), *Advances in social cognition* (Vol 11, pp. 1-52). Mahwah, NJ: Erlbaum.

Brehm, S. S., & Brehm, J. W. (1981). *Psychological reactance: A theory of freedom and control*. New York: Academic Press.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R.S. Wyer (Eds.), *Advances in social cognition* (Vol. 1, pp. 1-36). Hillsdale, NJ: Erlbaum.

Brewer, M. B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology*, *41*(4), 656-670.

Brown, R. (2010). *Prejudice: Its social psychology* (2nd ed.). Oxford: Wiley-Blackwell.

Brown, R., & Turner, J. C. (2002). The role of theories in the formation of stereotype content. In C. McGarty, V.Y. Yzerbyt, & R. Spears (Eds.), *Stereotypes as explanations: The formation of meaningful beliefs about social groups* (pp. 67–89). Cambridge: Cambridge University Press.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, *64*(2), 123-152.

Campbell, B., Schellenberg, E. G., & Senn, C. Y. (1997). Evaluating measures of contemporary sexism. *Psychology of Women Quarterly*, *21*(1), 89–102.

Carlston, D. E. (1994). Associated systems theory: A systematic approach to the cognitive representation of persons and events. In R. S. Wyer (Ed.), *Advances in social cognition: Vol. 7. Associated systems theory* (pp. 1–78). Hillsdale, NY: Lawrence Erlbaum.

Carlston, D. E. (2010). Models of implicit and explicit mental representation. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 38–61). New York, NY: Guildford Press.

Carlston, D. E., & Smith E. R. (1996). Principles of mental representation. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principle*s (pp. 184–210). New York: Guilford Press.

Carreiras, M., Garnham, A., Oakhill, J. V., & 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*, 639–663.

Carter, J. D., Hall, J. A., Carney, D. R., & Rosip, J. C. (2006). Individual differences in the acceptance of stereotyping. *Journal of Research in Personality*, *40*(6), 1103–1118.

Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, *71*(3), 464-478.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893-910.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed.), *Discourse Production and Comprehension* (pp. 1–40). Norwood, NJ: Ablex.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407-428.

Correll, S. J. (2004). Constraints into preferences: Gender, status, and emerging career aspirations. *American Sociological Review*, *69*(1), 93–113.

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: the struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359-378.

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*(5), 642–658.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*(5), 800-814.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5-18.

Devine, P. G. (2005). Breaking the prejudice habit: Allport's "inner conflict" revisited. In J. Dovidio, P. Glick, & L. Rudman (Eds.), O*n the nature of prejudice: Fifty years after Allport* (pp. 327–342). Malden, MA: Blackwell.

Devine, P. G., & Monteith, M. J. (1993). The role of discrepancy-associated affect in prejudice reduction. In D. M. Mackie & D. L. Hamilton (Eds.), *Affect, cognition and stereotyping: Interactive processes in intergroup perception* (pp. 317-344). San Diego: Academic Press.

Devine, P. G., & Monteith, M. J. (1999). Automaticity and control in stereotyping. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psycholo*gy (pp. 339-360). New York: Guildford Press.

Devine, P. G., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology*, *60*(6), 817.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*(5), 835-848.

Diamond, M. (2000). Sex and gender: Same or different? *Feminism & Psychology*, 10, 46-54.

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin*, *26*(10), 1171–1188.

Diekman, A. B., & Murnen, S. K. (2004). Learning to be little women and little men: The inequitable gender equality of nonsexist children's literature. *Sex Roles*, *50*(5-6), 373–385.

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, *22*(1), 22–37.

Dovidio, J.F., & Gaertner, S.L. (1998). On the nature of contemporary prejudice: The causes, consequences, and challenges of aversive racism. In J. Eberhardt & S.T. Fiske (Eds.), *Confronting racism: The problem and the response* (pp. 3–32). Newbury Park, CA: Sage.

Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory & Cognition*, *32*(4), 551–559.

Dunning, D., & Sherman, D. A. (1997). Stereotypes and tacit inference. *Journal of Personality and Social Psychology*, *73*(3), 459-471.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*(3), 316–326.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes.* Fort Worth, TX: Harcourt Brace.

Eagly, A. H., & Chaiken, S. (1998). Attitude structure and function. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 269–322). New York: McGraw-Hill.

Eagly, A. H., & Kite, M. E. (1987). Are stereotypes of nationalities applied to both women and men? *Journal of Personality and Social Psychology*, *53*(3), 451–462.

Eagly, A. H., Mladinic, A., & Otto, S. (1991). Are women evaluated more favorably than men?: An analysis of attitudes, beliefs, and emotions. *Psychology of Women Quarterly*, *15*(2), 203–216.

Erber, R., & Fiske, S. T. (1984). Outcome dependency and attention to inconsistent information. *Journal of Personality and Social Psychology*, *47*(4), 709-726.

Eurostat. (2013, March). Gender pay gap statistics. *Statistics Explained.* Retrieved 12/10/2013, http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Gender_pay_gap_statistics

Fagan, J. F. (1976). Infants' recognition of invariant features effaces. *Child Development*, 47, 627-638.

Fagan, J. F., & Shepherd, P. A. (1982). Theoretical issues in the early development of visual perception. In M. Lewis & L. Taft (Eds.), *Developmental disabilities in preschool children* (pp. 9-34). New York: Spectrum.

Fagan, J. F., & Singer, L. T. (1979). The role of simple feature differences in infants' recognition of faces. *Infant Behavior and Development*, 2, 39-45.

Fazio, R. H. (1990). The MODE model as an integrative framework. *Advances in Experimental Social Psychology*, *23*, 75–109.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013-1027.

Field, A. P. (2009). *Discovering statistics using SPSS: and sex and drugs and rock 'n' roll* (3 ed.). London: Sage.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley Publishing Co.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed. Vol. 2, pp.357-411). New York: McGraw-Hill.

Fiske, S. T. (2000). Stereotyping, prejudice, and discrimination at the seam between the centuries: Evolution, culture, mind, and brain. *European Journal of Social Psychology*, *30*(3), 299–322.

Fiske, S. T., & Neuberg, S. L. (1989). Category-based and individuating processes as a function of information and motivation: Evidence from our laboratory. In D. Bar-Tal, C. F. Graumann, A. W. Kruglanski, & W. Stroebe (Eds.), *Stereotyping and prejudice: Changing conceptions* (pp. 83–103). New York: Springer.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum model of impression formation from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 3, pp. 1–74). San Diego, CA: Academic Press.

Ford, T. E., & Ferguson, M. A. (2004). Social consequences of disparagement humor: A prejudiced norm theory. *Personality and Social Psychology Review*, *8*(1), 79–94.

Gabriel, U., Garnham, A., Sarrasin, O., Gygax, P., & Oakhill, J. (2010). Gender representation in different languages and grammatical marking on pronouns: When beauticians, musicians, and mechanics remain men. *Unpublished Manuscript.*

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. *Behavior Research Methods*, *40*(1), 206–212.

Gaertner S. L., & Dovidio J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination and racism* (pp. 61–89). Orlando, FL: Academic.

Gardiner, G. S. (1972). Complexity training and prejudice reduction. *Journal of Applied Social Psychology*, *2*(4), 326–342.

Garnham, A. (1992). Minimalism versus constructionism: A false dichotomy in theories of inference during reading. PSYCOLOQUY 3(63) reading-inference-1.1.

Garnham, A. (2001). *Mental models and the interpretation of anaphora*. Hove, England: Psychology Press.

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*(6), 481–500.

Garnham, A., & Oakhill, J. V. (1996). The mental models theory of language comprehension. In B.K. Britton & A.C. Graesser (Eds.), *Models of understanding text* (pp. 313-339). Hillside, NJ: Erlbaum.

Garnham, A., Oakhill, J., & Reynolds, D. (2002). Are inferences from stereotyped role names to characters' gender made elaboratively? *Memory & Cognition*, *30*(3), 439–446.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology*, *44*(2), 370–377.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, *60*(4), 509-517.

Girden, E. R. (1992). *ANOVA: Repeated measures*. Newbury Park: Sage Publications.

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*(3), 491-512.

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist*, *56*(2), 109-118.

Glick, P., Fiske, S. T., Mladinic, A., Saiz, J. L., Abrams, D., Masser, B., … Alao, A. (2000). Beyond prejudice as simple antipathy: hostile and benevolent sexism across cultures. *Journal of Personality and Social Psychology*, *79*(5), 763-775.

Glick, P., Sakalli-Ugurlu, N., Ferreira, M. C., & de Souza, M. A. (2002). Ambivalent sexism and attitudes toward wife abuse in Turkey and Brazil. *Psychology of Women Quarterly*, *26*(4), 292–297.

Gollwitzer, P. M. (1993). Goal achievement: The role of intentions. *European Review of Social Psychology*, *4*(1), 141–185.

Gottfredson, L. S. (1981). Circumscription and compromise: A developmental theory of occupational aspirations. *Journal of Counseling Psychology*, *28*(6), 545-579.

Gottfredson, L. S. (2005). Using Gottfredson's theory of circumscription and compromise in career guidance and counseling. . In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 71–100). New York: Wiley.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371–395.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197-216.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1-20.

Gygax, P., & Gabriel, U. (2011). Gender representation in language: More than meets the eye. In R. Mishra, & N. Srinivasan (Eds.), *Language and cognition: State of the art* (pp. 72–92). München: Lincom AP.

Hagoort, P., & Brown, C. M. (1999). Gender electrified: ERP evidence on the syntactic nature of gender processing. *Journal of Psycholinguistic Research*, *28*(6), 715–728.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483.

Hagoort, P., Brown, C. M., & Osterhout, L. (1999). The neurocognition of syntactic processing. *The Neurocognition of Language*, 273–316.

Hamilton, S. (2008). *Automatic gender stereotyping, an ERP Investigation*. Unpublished masters thesis, University of Sussex, UK.

Hamilton, D. L., & Sherman, J. W. (1994). Stereotypes. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition,* Vol. 2 (2nd. ed., pp. 1-68). Mahwah, NJ: Erlbaum.

Hamilton, D. L., Sherman, S. J., & Ruvolo, C. M. (1990). Stereotype-based expectancies: Effects on information processing and social behavior. *Journal of Social Issues*, *46*(2), 35–60.

Hantzi, A. (1995). Change in stereotypic perceptions of familiar and unfamiliar groups: The pervasiveness of the subtyping model. *British Journal of Social Psychology*, *34*(4), 463–477.

Hardin, C. D., & Conley, T. D. (2000). A relational approach to cognition: Shared experience and relationship affirmation in social cognition. In G. Moskowitz (Ed.), *Cognitive social psychology: The Princeton symposium on the legacy and future of social cognition* (pp. 3-17), Hillsdale, NJ: Erlbaum.

Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology*, *37*(1), 25-38.

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 512–521.

Hegarty, P., & Buechel, C. (2006). Androcentric reporting of gender differences in APA journals: 1965-2004. *Review of General Psychology*, *10*(4), 377-389.

Hellinger. M., & Bußmann, H. (Eds.). (2001–2003). *Gender across languages* (Vols. 1–3). Philadelphia: John Benjamins Company.

Higgins, E. T., Bargh, J. A., & Lombardi, W. J. (1985). Nature of priming effects on categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(1), 59-69.

Hill, S. E., & Flom, R. (2007). 18-and 24-month-olds' discrimination of gender-consistent and inconsistent activities. *Infant Behavior and Development*, *30*(1), 168–173.

Hilton, J. L., & von Hippel, W. (1990). The role of consistency in the judgment of stereotype-relevant behaviors. *Personality and Social Psychology Bulletin*, *16*(3), 430–448.

Hilton J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology. 47*, 237–71.

Hing, L. S. S., & Zanna, M. P. (2010). Individual Differences. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), T*he SAGE handbook of prejudice, stereotyping, and discrimination* (pp. 163-178). Thousand Oaks, CA: SAGE Publications.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological Bulletin*, *107*(2), 139-155.

Irmen, L. (2007). What's in a (role) name? Formal and conceptual aspects of comprehending personal nouns. *Journal of Psycholinguistic Research*, *36*(6), 431–456.

Irmen, L., & Roßberg, N. (2004). Gender markedness of language: The impact of grammatical and nonlinguistic information on the mental representation of person information. *Journal of Language and Social Psychology*, *23*(3), 272–307.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541.

Jacoby, L. L., Toth, J. P., Lindsay, D. S., & Debner, J. A. (1992). Lectures for a layperson: Methods for revealing unconscious processes. In R. F. Bornstein & T. S. Pittman (Eds.), *Perception without awareness* (pp. 81-120). New York: Guilford Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Johnston, L., & Hewstone, M. (1992). Cognitive models of stereotype change: 3. Subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology*, *28*(4), 360–386.

Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and Social Psychology Bulletin*, *26*(8), 1002–1012.

Kawakami, K., Dion, K. L., & Dovidio, J. F. (1998). Racial prejudice and stereotype activation. *Personality and Social Psychology Bulletin*, *24*(4), 407–416.

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, *78*(5), 871-888.

Kawakami, K., Dovidio, J. F., & van Kamp, S. (2005). Kicking the habit: Effects of nonstereotypic association training and correction processes on hiring decisions. *Journal of Experimental Social Psychology*, *41*(1), 68–75.

Kawakami, K., Dovidio, J. F., & van Kamp, S. (2007). The impact of counterstereotypic training and related correction processes on the application of stereotypes. *Group Processes & Intergroup Relations*, *10*(2), 139–156.

Kelman, H. C. (1958). Compliance, identification, and internalization: Three processes of attitude change. *The Journal of Conflict Resolution*, *2*(1), 51–60.

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.

Knutson, K. M., Mah, L., Manly, C. F., & Grafman, J. (2007). Neural correlates of automatic beliefs about gender and race. *Human Brain Mapping*, *28*(10), 915–930.

Kreiner, H., Sturt, P., & Garrod, S. (2008). Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. *Journal of Memory and Language*, *58*(2), 239–261.

Krueger, J., & Rothbart, M. (1990). Contrast and accentuation effects in category learning. *Journal of Personality and Social Psychology*, *59*(4), 651-663.

Kuklinski, J. H., & Hurley, N. L. (1996). It's a Matter of Interpretation. In D. M. Mutz, P. M. Sniderman, and R. Brody, (Eds.), *Political persuasion and attitude change* (pp. 125–144). Ann Arbor: University of Michigan Press.

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology*, *68*(4), 565-579.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*(2), 284-308.

Kutas, M., & Van Petten, C. (1994). Psycholinguistics electrified: Event related potential investigations. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 83–143). San Diego, CA: Academic Press.

LaFrance, M., & Woodzicka, J. A. (1998). No laughing matter:Women's verbal and nonverbal reactions to sexist humor. In J.Swim & C. Stangor (Eds.), *Targets of prejudice*. San Diego: Academic Press.

Lambdin, J. R., Greer, K. M., Jibotian, K. S., Wood, K. R., & Hamilton, M. C. (2003). The animal= male hypothesis: Children's and adults' beliefs about the sex of non–sex-specific stuffed animals. *Sex Roles*, *48*(11-12), 471–482.

Lassonde, K. A., & O'Brien, E. J. (2013). Occupational stereotypes: activation of male bias in a gender-neutral world. *Journal of Applied Social Psychology*. *43*(2), 387-396.

Leary, M. R. (1983). A brief version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin*, *9*(3), 371–375.

Lenton, A. P., Bruder, M., & Sedikides, C. (2009). A meta-analysis on the malleability of automatic gender stereotypes. *Psychology of Women Quarterly*, *33*(2), 183–196.

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology*, *72*(2), 275-287.

Levine, J. M., Resnick, L. B., & Higgins, E. T. (1993). Social foundations of cognition. *Annual Review of Psychology*, *44*(1), 585–612.

Liben, L. S., Bigler, R. S., & Krogh, H. R. (2002). Language at work: Children's gendered interpretations of occupational titles. *Child Development*, *73*(3), 810–828.

Linville, P. W., Fisher, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology,* 57, 165–188.

Lippmann, W. (1922). The world outside and the pictures in our heads. *Public Opinion*, *4*, 1–22.

Locke, V., MacLeod, C., & Walker, I. (1994). Automatic and controlled activation of stereotypes: Individual differences associated with prejudice. *British Journal of Social Psychology*, *33*(1), 29–46.

Lord, C. G., & Lepper, M. R. (1999). Attitude representation theory. *Advances in Experimental Social Psychology*, *31*, 265–343.

Lord, C. G., Lepper, M. R., & Mackie, D. (1984). Attitude prototypes as determinants of attitude–behavior consistency. *Journal of Personality and Social Psychology*, *46*(6), 1254-1266.

Lun, J., Sinclair, S., Whitchurch, E. R., & Glenn, C. (2007). (Why) do I think what you think? Epistemic social tuning and implicit prejudice. *Journal of Personality and Social Psychology*, *93*(6), 957-972.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, *51*, 93-120.

Macrae, C. N., Bodenhausen, G. V., & Milne, A. B. (1998). Saying no to unwanted thoughts: self-focus and the regulation of mental life. *Journal of Personality and Social Psychology*, *74*(3), 578-589.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Ford, R. L. (1997). On regulation of recollection: The intentional forgetting of stereotypical memories. *Journal of Personality and Social Psychology*, *72*(4), 709-719.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology*, *67*(5), 808-817.

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology*, *23*(1), 77–87.

Maio, G. R., Haddock, G., Manstead, A. S., & Spears, R. (2010). Attitudes and Intergroup Relations. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), *The SAGE handbook of prejudice, stereotyping, and discrimination* (pp. 261-275). Thousand Oaks, CA: SAGE Publications.

Matheson, K., & Kristiansen, C. M. (1987). The effect of sexist attitudes and social structure on the use of sex-biased pronouns. *The Journal of Social Psychology*, *127*(4), 395–398.

Maurer, K. L., Park, B., & Rothbart, M. (1995). Subtyping versus subgrouping processes in stereotype representation. *Journal of Personality and Social Psychology*, *69*(5), 812-824.

McCann, C. D., Ostrom, T. M., Tyner, L. K., & Mitchell, M. L. (1985). Person perception in heterogeneous groups. *Journal of Personality and Social Psychology*, *49*(6), 1449-1459.

McConahay, J. B., Hardee, B. B., & Batts, V. (1981). Has racism declined in America? It depends on who is asking and what is asked. *Journal of Conflict Resolution*, *25*(4), 563–579.

McKoon, G., & Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(1), 82-91.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*(3), 440-466.

McMinn, M. R., Lindsay, S. F., Hannum, L. E., & Troyer, P. K. (1990). Does sexist language reflect personal characteristics? *Sex Roles*, *23*(7-8), 389–396.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207-238.

Merry, E. (1995). *First names: the definitive guide to popular names in England and Wales 1944-1994 and in the regions 1994*. London: Her Majestry's Stationery Office.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371-378.

Miller, C. L. (1983). Developmental changes in male/female voice classification by infants. *Infant Behavior and Development*, *6*(2), 313–330.

Miller, C. L., & McFarland, C. (1991). When social comparison goes awry: The case of pluralistic ignorance. In J. Suls & T.A. Wills (Eds.), *Social comparison: Contemporary theory and research* (pp. 287-313). Hillsdale, NJ: Erlbaum.

Miller, C. L., & Read. S. J. (1991). On the coherence of mental models of persons and relationships: A knowledge structure approach. In G. J. O. Fletcher & F. D. Fincham (Eds.), *Cognition in close relationships* (pp. 69–99). Hillsdale, NJ: Erlbaum.

Miller, C. L., Younger, B. A., & Morse, P. A. (1982). The categorization of male and female voices in infancy. *Infant Behavior and Development*, *5*(2), 143–159.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology*, *65*(3), 469-485.

Monteith, M. J., Devine, P. G., & Zuwerink, J. R. (1993). Self-directed versus other-directed affect as a consequence of prejudice-related discrepancies. *Journal of Personality and Social Psychology*, *64*(2), 198-210.

Monteith, M. J., Sherman, J. W., & Devine, P. G. (1998). Suppression as a stereotype control strategy. *Personality and Social Psychology Review*, *2*(1), 63–82.

Monteith, M. J., Zuwerink, J. R., & Devine, P. G. (1994). Prejudice and prejudice reduction: Classic challenges, contemporary approaches. In P. G. Devine, D. L. Hamilton, & T. M. Ostrom (Eds.), *Social cognition: Impact on social psychology* (pp. 323–46). San Diego, CA: Academic Press.

Moser, F., & Hannover, B. (2013). How gender fair are German schoolbooks in the twenty-first century? An analysis of language and illustrations in schoolbooks for mathematics and German. *European Journal of Psychology of Education*, 1–21.

Moskowitz, G. B. (1993). Individual differences in social categorization: The influence of personal need for structure on spontaneous trait inferences. *Journal of Personality and Social Psychology*, *65*(1), 132-142.

Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, *77*(1), 167-184.

Mussweiler, T., & Neumann, R. (2000). Sources of mental contamination: Comparing the effects of self-generated versus externally provided primes. *Journal of Experimental Social Psychology*, *36*(2), 194–206.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226-254.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264-336). Hillsdale, NJ: Erlbaum.

Nelson, T. E., Biernat, M. R., & Manis, M. (1990). Everyday base rates (sex stereotypes): Potent and resilient. *Journal of Personality and Social Psychology*, *59*(4), 664-675.

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, *53*(3), 431-444.

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of Personality and Social Psychology*, *65*(1), 113-131.

O'Brien, M., & Huston, A. C. (1985). Development of sex-typed play behavior in toddlers. *Developmental Psychology*, *21*(5), 866-871.

Oakhill, J., Garnham, A., & Reynolds, D. (2005). Immediate activation of stereotypical gender information. *Memory & Cognition*, *33*(6), 972–983.

Olson, J. M., Roese, N. J., & Zanna, M. P. (1996). Expectancies. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 211–238). New York, NY, US: Guilford Press.

Operario, D., & Fiske, S. T. (2004). Stereotypes: Content, Structures, Processes, and Context. In M. B. Brewer & M. Hewstone (Eds.), *Social cognition* (pp. 120–141). Malden: Blackwell Publishing.

Osterhout, L., Bersick, M., & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, *25*(3), 273–285.

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, *34*(6), 739–773.

Paluck, E. L. (2006). Diversity training and intergroup contact: A call to action research. *Journal of Social Issues*, *62*(3), 577–595.

Paluck, E. L. (2011). Peer pressure against prejudice: A high school field experiment examining social network change. *Journal of Experimental Social Psychology*, *47*(2), 350–358.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*, 339–367.

Parks, J. B., & Roberton, M. A. (2000). Development and validation of an instrument to measure attitudes toward sexist/nonsexist language. *Sex Roles*, *42*(5-6), 415–438.

Parks, J. B., & Roberton, M. A. (2004). Attitudes toward women mediate the gender effect on attitudes toward sexist language. *Psychology of Women Quarterly*, *28*(3), 233–239.

Pedhazur, E. J., & Tetenbaum, T. J. (1979). Bem Sex Role Inventory: A theoretical and methodological critique. *Journal of Personality and Social Psychology*, *37*(6), 996-1016.

Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology*, *59*(3), 475-486.

Perdue, C. W., & Gurtman, M. B. (1990). Evidence for the automaticity of ageism. *Journal of Experimental Social Psychology*, *26*(3), 199–216.

Pettigrew, T. F. (1981). Extending the stereotype concept. In D. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 303–331). Hillsdale, NJ: Erlbaum.

Petty, R. E., & Jarvis, W. B. G. (1998, October). *What happens to the "old" attitude when attitudes change?* Paper presented at the annual meeting of the Society of Experimental Social Psychology, Lexington, KY.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811-832.

Poulin-Dubois, D., Serbin, L. A., Eichstedt, J. A., Sen, M. G., & Beissel, C. F. (2002). Men Don't Put on Make-up: Toddlers' Knowledge of the Gender Stereotyping of Household Activities. *Social Development*, *11*(2), 166–181.

Powell, M. C., & Fazio, R. H. (1984). Attitude accessibility as a function of repeated attitudinal expression. *Personality and Social Psychology Bulletin*, *10*(1), 139–148.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, *64*(2), 243-256.

Pryzgoda, J., & Chrisler, J. C. 2000. Definitions of gender and sex: The subtleties of meaning. *Sex Roles, 43*(7/8), 553-569.

Puhl, R. M., Schwartz, M. B., & Brownell, K. D. (2005). Impact of perceived consensus on stereotypes about obese people: a new approach for reducing bias. *Health Psychology*, *24*(5), 517-525.

Pyykkönen, P. (2009). *The Importance of Semantics: Visual World Studies on Drawing Inferences and Resolving Anaphors*. Unpublished doctoral dissertation, University of Turku.

Pyykkönen, P., Hyönä, J., & van Gompel, R. P. (2010). Activating gender stereotypes during online spoken language processing: evidence from Visual World Eye Tracking. *Experimental Psychology*, *57*(2), 126-133.

Reskin, B. F., & Roos, P. A. (1990). *Job queues, gender queues: Explaining women's inroads into male occupations.* Philadelphia: Temple University Press.

Reynolds, D. J., Garnham, A., & Oakhill, J. (2006). Evidence of immediate activation of gender information from a social role name. *The Quarterly Journal of Experimental Psychology*, *59*(05), 886–903.

Roediger, H. L., Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of Memory and Consciousness* (pp. 3-42). Hillsdale, NJ: Erlbaum.

Roopnarine, J. L. (1986). Mothers' and fathers' behaviors toward the toy play of their infant sons and daughters. *Sex Roles*, *14*(1-2), 59–68.

Rothbart, M. (1981). Memory processes and social beliefs. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behaviour* (pp. 145–181). Hillsdale, NJ: Erlbaum.

Rothbart, M. (1996). Category-exemplar dynamics and stereotype change. *International Journal of Intercultural Relations*, *20*(3), 305–321.

Rothbart, M., & John, O. P. (1985). Social categorization and behavioral episodes: A cognitive analysis of the effects of intergroup contact. *Journal of Social Issues*, *41*(3), 81–104.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). " Unlearning" automatic biases: the malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology*, *81*(5), 856-868.

Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men? *Journal of Personality and Social Psychology*, *87*(4), 494-509.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. (1986). Schemata and sequential though processes in PDP models. In J. L. McClelland & D. E. Rumelhart

(Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (vol. II, pp. 7-57). Cambridge, MA: MIT Press.

Sanford, A. J. (1985). *Cognition and cognitive psychology.* London: Weidenfeld & Nicolson.

Sato, S., Gygax, P. M., & Gabriel, U. (2013). Gender inferences: Grammatical features and their impact on the representation of gender in bilinguals. *Bilingualism: Language and Cognition*, *16*(4), 792-807.

Schaller, M., Asp, C. H., Roseil, M. C., & Heim, S. J. (1996). Training in statistical reasoning. inhibits the formation of erroneous group stereotypes. *Personality and Social Psychology Bulletin*, *22*(8), 829–844.

Schaller, M., Boyd, C., Yohannes, J., & O'Brien, M. (1995). The prejudiced personality revisited: Personal need for structure and formation of erroneous group stereotypes. *Journal of Personality and Social Psychology*, *68*(3), 544-555.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1-66.

Sechrist, G. B., & Stangor, C. (2001). Perceived consensus influences intergroup behavior and stereotype accessibility. *Journal of Personality and Social Psychology*, *80*(4), 645-654.

Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. *Advances in Experimental Social Psychology*, *29*, 209–269.

Serbin, L. A., Poulin-Dubois, D., Colburne, K. A., Sen, M. G., & Eichstedt, J. A. (2001). Gender stereotyping in infancy: Visual preferences for and knowledge of gender-stereotyped toys in the second year. *International Journal of Behavioral Development*, *25*(1), 7–15.

Serbin, L. A., Poulin-Dubois, D., & Eichstedt, J. A. (2002). Infants' Responses to Gender-Inconsistent Events. *Infancy*, *3*(4), 531–542.

Sherif, M., & Sherif, C. W. (1953). *Groups in harmony and tension; an integration of studies of intergroup relations*. New York: Harper.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127-190.

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. New York: Cambridge University Press.

Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: the role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*(4), 583-592.

Smith, E. R. (1989). Procedural efficiency: General and specific components and effects on social judgment. *Journal of Experimental Social Psychology*, *25*(6), 500–523.

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology*, *70*(5), 893-912.

Smith, E. R., Branscombe, N. R., & Bormann, C. (1988). Generality of the effects of practice on social judgment tasks. *Journal of Personality and Social Psychology*, *54*(3), 385-395.

Smith, E. R., & Conrey, F. R. (2007). Mental representations are states, not things: Implications for implicit and explicit measurement. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 247-264). New York: Guilford.

Smith, E. R., & DeCoster, J. (1999). Associative and rule based processing: a connectionist interpretation of dual-process models. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 323-336). New York: Guilford.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–74.

Son Hing, L. S., Li, W., & Zanna, M. P. (2002). Inducing hypocrisy to reduce prejudicial responses among aversive racists. *Journal of Experimental Social Psychology*, *38*(1), 71–78.

Spence, J. T., & Helmreich, R. L. (1972). The Attitudes Toward Women Scale: An objective instrument to measure attitudes toward the rights and roles of women in contemporary society. *JSAS Catalog of Selected Documents in Psychology, 2*, 1-48.

Stahlberg, D., Braun, F., Irmen, L., & Sczesny, S. (2007). Representation of the sexes in language. *Social Communication*, 163–187.

Stangor, C. (2009). The study of stereotyping, prejudice, and discrimination within social psychology: A quick history of theory and research. In T. D. Nelson, (Ed.), *Handbook of prejudice, stereotyping and discrimination* (pp. 1-22). New York: Psychology Press, Taylor & Francis Group.

Stangor, C., Lynch, L., Duan, C., & Glas, B. (1992). Categorization of individuals on the basis of multiple social features. *Journal of Personality and Social Psychology*, *62*(2), 207-218.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin*, *111*(1), 42-61.

Stangor, C., Sechrist, G. B., & Jost, J. T. (2001). Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin*, *27*(4), 486–496.

Stangor, Thompson, & Ford, T. E. (1998). An inhibited model of stereotype inhibition. In R. Wyer (Ed.), *Advances in social cognition* (pp. 193-210). Mahwah, NJ: Erlbaum.

Stapel, D. A., & Koomen, W. (1998). When stereotype activation results in (counter) stereotypical judgments: Priming stereotype-relevant traits and exemplars. *Journal of Experimental Social Psychology*, *34*(2), 136–163.

Stapel, D., Martin, L., & Schwarz, N. (1998). The smell of bias: What instigates correction processes in social judgments? *Personality and Social Psychology Bulletin,* 24, 797–806.

Stephan, W. G., & Stephan, C. W. (1984). The role of ignorance in intergroup relations. In N. Miller & M. B. Brewer (Eds.), *Groups in contact: The psychology of desegregation* (pp. 229–257). Orlando, FL: Academic Press.

Strack, F., & Hannover, B. (1996). Awareness of influence as a precondition for implementing correctional goals. In P. Gollwitzer & J. A. Bargh (Eds.),*The psychology of action: Linking cognition and motivation to behaviour* (pp. 579-595). New York: Guildford.

Stricker, L. J., Messick, S., & Jackson, D. N. (1967). Suspicion of deception: Implications for conformity research. *Journal of Personality and Social Psychology*, *5*(4), 379-389.

Stroessner, S. J., Hamilton, D. L., & Mackie, D. M. (1992). Affect and stereotyping: the effect of induced mood on distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology*, *62*(4), 564-576.

Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: old-fashioned and modern prejudices. *Journal of Personality and Social Psychology*, *68*(2), 199-214.

Swim, J. K., & Cohen, L. L. (1997). Overt, Covert, And Subtle Sexism A Comparison Between the Attitudes Toward Women and Modern Sexism Scales. *Psychology of Women Quarterly*, *21*(1), 103–118.

Taylor, S. E. (1981). A categorization approach to stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 88–114). Hillsdale, NJ: Erlbaum.

Tetlock, P. E. (1992). The impact of accountability on judgment and choice: Toward a social contingency model. *Advances in Experimental Social Psychology*, *25*(3), 331–76.

Thompson, M. M., Naccarato, M. E., & Parker, K. E. (1992). *Measuring cognitive needs: The development and validation of the Personal Need for Structure (PNS) and Personal Fear of Invalidity (PFI) measures*. Unpublished Manuscript.

Thompson, Roman, R. J., Moskowitz, G. B., Chaiken, S., & Bargh, J. A. (1994). Accuracy motivation attenuates covert priming: The systematic reprocessing of social information. *Journal of Personality and Social Psychology*, *66*(3), 474-489.

UK Nursing and Midwifery Council. (2011). *Analysis of diversity data*. Retrieved 12/10/2013, http://www.nmc-uk.org/About-us/Equality-and-diversity/Analysis-of-diversity-data-2011/

Van Berkum, J. J., Brown, C. M., & Hagoort, P. (1999). Early referential context effects in sentence processing: Evidence from event-related brain potentials. *Journal of Memory and Language*, *41*(2), 147–182.

Watson, D., & Friend, R. (1969). Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, *33*(4), 448-457.

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology*, *45*(5), 961-977.

Wegner, D. M. (1994). *White bears and other unwanted thoughts*. New York, NY: Guilford.

Wegener, D. T., & Petty, R. E. (1997). The flexible correction model: The role of naive theories of bias in bias correction. *Advances in Experimental Social Psychology*, *29*, 141–208.

Wenzlaff, R. M., & Wegner, D. M. (2000). Thought suppression. *Annual review of psychology*, *51*(1), 59–91.

Wheeless, V. E., & Dierks-Stewart, K. (1981). The psychometric properties of the Bem sex-role inventory: Questions concerning reliability and validity. *Communication Quarterly*, *29*(3), 173–186.

Wilbourn, M. P., & Kee, D. W. (2010). Henry the nurse is a doctor too: Implicitly examining children's gender stereotypes for male and female occupational roles. *Sex Roles*, *62*(9-10), 670–683.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117-142.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological review*, *107*(1), 101.

Wittenbrink, B., & Henly, J. R. (1996). Creating social reality: Informational social influence and the content of stereotypic beliefs. *Personality and Social Psychology Bulletin*, *22*(6), 598–610.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262-274.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: variability in automatically activated attitudes. *Journal of Personality and Social Psychology*, *81*(5), 815-827.

Wyer, R. S., & Carlston, D. E. (1979). *Social cognition, inference and attribution*. Hillsdale, NJ: Erlbaum.

Wyer, N. A., & Hamilton, D. L. (1998). The balance between excitation and inhibition in stereotype use. In R. Wyer (Ed.), *Advances in social cognition* (pp. 1-52). Mahwah, NJ: Erlbaum.

Zarate, M. A., & Smith, E. R. (1990). Person categorization and stereotyping. *Social Cognition*, *8*(2), 161–185.

Zimbardo, P. G. (1972). Comment: Pathology of imprisonment. *Society*, *9*(6), 4–8.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162-185.

Appendix 1 **Ethical Approval Certificate**

Life Sciences & Psychology Cluster based Research Ethics Committee University of Sussex CERTIFICATE OF APPROVAL	
Reference Number:	JOEF0311
Title of Project:	Cognitive representations of gender stereotypes
Principal Investigator:	Jane Oakhill
Student:	Eimear Finnegan
Collaborators:	
Duration of Approval (not greater than 4 years)	24 months
Expected Start Date:*	March 2011

This project has been given ethical approval by the Life Sciences and Psychology Cluster based Research Ethics Committee (C-REC).

*NB. If the actual project start date is delayed beyond 12 months of the expected start date, this Certificate of Approval will lapse and the project will need to be reviewed again to take account of changed circumstances such as legislation, sponsor requirements and University procedures.

Please note and follow the requirements for approved submissions:

Amendments to protocol.

- Any changes or amendments to approved protocols must be submitted to the C-REC for authorisation prior to implementation.

Feedback regarding the status and conduct of approved projects

- Any incidents with ethical implications that occur during the implementation of the project must be reported immediately to the Chair of the C-REC.

The principal investigator is required to provide a brief annual written statement to the committee, indicating the status and conduct of the approved project. These reports will be reviewed at the annual meeting of the committee. A statement by the Principal Investigator to the C-REC indicating the status and conduct of the approved project will be required on the following date(s):

December 2011, 2012……

Authorised Signature	Jennifer Rusted
Name of Authorised Signatory (C-REC Chair or nominated deputy)	**Jennifer Rusted**
Date	27 March 2011

Appendix 2 **Role Nouns: Experiments 1-6**

High ratings indicate masculinity, low ratings indicate femininity, and ratings of 50% indicate neutrality.

Male stereotyped role nouns:

		Bias	*SD*
1	Bricklayer	88.24	-11.26
2	President	86.80	-16.96
3	Boxer	86.27	-15.62
4	Mechanic	85.20	-10.35
5	Football coach	84.71	-15.15
6	Lorry driver	82.75	-15.50
7	Hunter	82.16	-13.16
8	Factory manager	79.41	-14.06
9	Electrician	79.22	-17.98
10	Pilot	77.84	-13.76
11	Golfer	77.45	-12.62
12	Politician	77.14	-14.43

Neutral-rated role nouns:

		Bias	*SD*
1	Pedestrian	49.80	-03.74
2	Proof reader	50.39	-15.74
3	Author	49.60	-10.29
4	Trainee	49.41	-08.81
5	Neighbour	50.78	-15.47
6	Gynaecologist	49.22	-19.68
7	Jogger	48.82	-11.77
8	Concert go-er	48.43	-09.67
9	Relative	52.04	-08.41
10	Office worker	47.65	-13.94
11	Artist	52.55	-13.54
12	Adolescent	52.94	-10.64

Female stereotyped role nouns:

		Bias	*SD*
1	Beautician	13.27	-10.88
2	Fortune teller	18.82	-12.91
3	Au pair	19.39	-16.38
4	Secretary	21.76	-13.81
5	Dressmaker	22.94	-13.90
6	Cleaner	25.29	-13.32
7	Flight attendant	27.20	-13.56
8	Social worker	27.84	-12.05
9	Model	28.43	-14.33
10	Nurse	29.02	-21.28
11	Chocolate lover	29.22	-15.47
12	Birth attendant	30.82	-18.01

Appendix 3 **Filler Role Nouns: Experiments 1-6**

Eight of the male and female definitional terms used in the matching condition were also repeated in the mismatching condition (those in bold font were not repeated).

Condition		Male Definitional Role nouns	Condition		Female Definitional Role nouns
Matching	1	Policeman	**Matching**	1	Landlady
	2	Groom		2	Heroine
	3	Postman		3	Mistress
	4	Salesman		4	Spinster
	5	Bachelor		5	Hostess
	6	Steward		6	Bride
	7	Waiter		7	Waitress
	8	King		8	Princess
	9	**Craftsman**		**9**	**Mermaid**
	10	**Prince**		**10**	**Ballerina**
Mismatching	1	Policeman	**Mismatching**	1	Landlady
	2	Groom		2	Heroine
	3	Postman		3	Mistress
	4	Salesman		4	Spinster
	5	Bachelor		5	Hostess
	6	Steward		6	Bride
	7	Waiter		7	Waitress
	8	King		8	Princess
	9	Son		9	Stewardess
	10	Sir		10	Milkmaid
	11	Knight		11	Salesgirl
	12	Master		12	Duchess
	13	Pope		13	Countess
	14	Hero		14	Dame
	15	Husband		15	God mother
	16	Landlord		16	Policewoman
	17	God father		17	Grandmother
	18	Count		18	Seamstress
	19	Gigolo		19	Geisha
	20	Baron		20	Lesbian
	21	Fireman		21	Matron
	22	Grandfather		22	Baroness
	23	Milkman		23	Nun
	24	Host		24	Step mother
	25	Duke		25	Maid of honour
	26	Best man		26	Barmaid
	27	Barman		27	Wife
	28	Step brother		28	Queen
	29	Step father		29	Madam
	30	Priest		30	Daughter

Appendix 4 **Information Sheet**

Information sheet

Purpose of study:

This study investigates attitudes and decision making about different role names in English.

The task:

To begin with you will complete 4 different questionnaires related to how people should react in a variety of different scenarios, your personality and attitude/belief measures. These should take about 20 minutes to complete.

They will be followed by the main experiment made up of three blocks of decision-making trials in which you must decide whether or not two terms presented on screen can be used to refer to the same person. The second block differs from the other two in that you will receive feedback after each of your responses indicating whether you were correct or incorrect.

Try to concentrate throughout the experiment, as there are many different sections and instructions change frequently. Also, if tired, please use the time after each section to rest before beginning another.

The entire study should take approximately 1hr to complete and you will receive £6 on completion. If you are a psychology undergraduate you can opt to receive 4 course credits instead.

Please address any immediate concerns or questions to the researcher. Similarly, if you have any further queries or concerns after the experiment please contact the researcher at e.finnegan@sussex.ac.uk . If you are happy to proceed, please read through the consent form and sign it before beginning the experiment.

Many thanks,

Eimear Finnegan.

Remember, you may withdraw from the study at any time, without giving a reason

Appendix 5 **Consent Form**

University of Sussex

Volunteer Consent

I consent to take part in this study, of which the nature and purpose have been explained to me. I have read and understand the attached information sheet.

I understand that any information I provide is confidential, and that no information that I disclose will lead to the identification of any individual in the reports on the project, either by the researcher or by any other party. In addition, all data will be used for the sole purpose of this experiment.

I understand that my participation is voluntary, that I can choose not to participate in part or all of the project, and that I can withdraw at any stage of the project without being penalised or disadvantaged in any way.

Name: _____

Signature _____

Date: _____

Appendix 6 **Role Nouns: Experiment 3**

The items previously used in Experiments 1 and 2 are numbered 1-6, while new items are numbered 7-12. Items taken from Kennison and Trofe (2003) are presented in bold font; all others were sourced from Gabriel et al. (2008). In the former case, average ratings of both male and female participants are provided along with the standard deviations. The ratings of female participants are shown on the left hand side while those of the male participants are shown on the right hand side.

Male stereotyped role nouns:

		Block 1 and 2				**Block 3 and 4**	
		Bias	***SD***			**Bias**	***SD***
1	Bricklayer	88.24	-11.26	1	President	86.80	-16.96
2	Mechanic	85.20	-10.35	2	Boxer	86.27	-15.62
3	Football coach	84.71	-15.15	3	Lorry driver	82.75	-15.50
4	Factory manager	79.41	-14.06	4	Hunter	82.16	-13.16
5	Electrician	79.22	-17.98	5	Pilot	77.84	-13.76
6	Politician	77.14	-14.43	6	Golfer	77.45	-12.62
7	Engineer	77.00	-13.13	7	Farmer	76.80	-14.49
8	Mathematician	74.90	-12.06	8	Carpenter	75.49	-18.15
9	Murderer	74.40	-13.12	9	Physicist	73.53	-12.62
10	Judge	72.75	-14.43	10	Butcher	73.33	-22.15
11	Technician	72.75	-15.63	11	Inventor	71.63	-12.31
12	Prisoner	71.57	-12.06	12	Statistician	71.60	-13.76

Neutral-rated role nouns:

		Block 1 and 2				**Block 3 and 4**	
		Bias	***SD***			**Bias**	***SD***
1	Pedestrian	49.80	-03.74	1	Proof reader	50.39	-15.74
2	Trainee	49.41	-08.81	2	Author	49.60	-10.29
3	Neighbour	50.78	-15.47	3	Gynaecologist	49.22	-19.68
4	Concert go-er	48.43	-09.67	4	Jogger	48.82	-11.77
5	Relative	52.04	-08.41	5	Office worker	47.65	-13.94
6	Cinema go-er	46.60	-08.23	6	Artist	52.55	-13.54
7	Swimmer	53.14	-09.48	7	Musician	53.92	-12.34
8	Skier	55.88	-09.42	8	Bank Clerk	54.60	-14.60
9	Spectator	54.71	-12.86	9	**Tour guide**	**4.00 / 3.95**	**0.79 / 1.00**
10	**Entertainer**	**4.05 / 4.15**	**0.22 / 0.49**	10	**Customer**	**3.75 /4.05**	**0.72 / 0.51**
11	**Patient**	**4.00 / 4.15**	**0.56 / 0.88**	11	**Photographer**	**4.00 / 3.80**	**0.79 / 0.52**
12	**Journalist**	**4.15 / 3.90**	**1.27 / 1.21**	12	**Acrobat**	**3.85 / 4.15**	**1.42 / 1.42**

Female stereotyped role nouns:

		Block 1 and 2				Block 3 and 4	
		Bias	***SD***			**Bias**	***SD***
1	Beautician	13.27	-10.88	1	Fortune teller	18.82	-12.91
2	Secretary	21.76	-13.81	2	Au pair	19.39	-16.38
3	Dressmaker	22.94	-13.90	3	Cleaner	25.29	-13.32
4	Social worker	27.84	-12.05	4	Flight attendant	27.20	-13.56
5	Model	28.43	-14.33	5	Nurse	29.02	-21.28
6	Birth attendant	30.82	-18.01	6	Chocolate lover	29.22	-15.47
7	Sales Assistant	32.55	-13.54	7	Dancer	33.53	-14.54
8	Hairdresser	34.60	-21.59	8	Cashier	33.67	-15.23
9	Dietician	34.71	-15.54	9	Ice Skater	34.80	-12.66
10	Violinist	42.35	-13.80	10	Librarian	35.00	-16.07
11	Singer	44.80	-11.47	11	Cook	45.10	-18.91
12	Secretary	21.76	-13.81	12	**Florist**	**1.95 / 2.15**	**0.76 / 1.14**

Appendix 7 **Fictitious feedback range: Social Consensus Feedback**

Critical Items	**'Yes' judgement**	**'No' judgement**
Neutral terms	97-100%	0-3%
Male/Female stereotype congruent terms	97-100%	0-3%
Male/Female stereotype incongruent terms	95-98%	5-2%

Fillers	**'Yes' judgement**	**'No' judgement**
Definitional gender Match	98-100%	0-2%
Definitional gender Mismatch	0-2%	98-100%

Appendix 8 **Fictitious feedback range: Reverse Social Consensus Feedback**

Critical Items	**'Yes' judgement**	**'No' judgement**
Neutral terms	97-100%	0-3%
Male/Female stereotype match terms	97-100%	0-3%
Male/Female stereotype mismatch terms	35-65%	35-65%

Fillers	**'Yes' judgement**	**'No' judgement**
Definitional gender Match	98-100%	0-2%
Definitional gender Mismatch	0-2%	98-100%

Appendix 9 **Pilot study 1: Plausibility of the Social Consensus Feedback**

Questions:

Q1. What do you think the experiment is about?

Q2. Do you think you were influenced by the feedback? If yes, in what way?

Q3. How believable did you find the feedback?

1	**2**	**3**	**4**	**5**
Believable	Quite believable	Neither believable nor unbelievable	Quite unbelievable	Unbelievable

Q4. Did you find the feedback surprising? If so, in what cases? Can you think of any examples?

Q5. Any other comments?

Appendix 10 **Materials: Experiment 7**

1. Judgement task role nouns

Male stereotyped role nouns:

	Bias	*SD*
Mechanic	85.2	-10.35
Carpenter	75.49	-18.15
Electrician	79.22	-17.98
President	86.8	-16.96
Hunter	82.16	-13.16
Pilot	77.84	-13.76
Politician	77.14	-14.43
Mathematician	74.9	-12.06
Member of Parliament	74.08	-09.56
Butcher	73.33	-22.15
Physics student	72.94	-12.21
Technician	72.75	-15.63

Neutral-rated role nouns:

	Bias	*SD*
Pedestrian	49.8	-03.74
Proof reader	50.39	-15.74
Author	49.6	-10.29
Trainee	49.41	-08.81
Neighbour	50.78	-15.47
Gynaecologist	49.22	-19.68
Jogger	48.82	-11.77
Concert go-er	48.43	-09.67
Relative	52.04	-08.41
Office worker	47.65	-13.94
Artist	52.55	-13.54
Adolescent	52.94	-10.64

Female stereotyped role nouns:

	Bias	*SD*
Secretary	21.76	-13.81
Flight attendant	27.20	-13.56
Nurse	29.02	-21.28
Fortune teller	18.82	-12.91
Cleaner	25.29	-13.32
Chocolate lover	29.22	-15.47
Psychology student	31.37	-18.11
Dancer	33.53	-14.54
Dietician	34.71	-15.54
Librarian	35	-16.07
Babysitter	**1.85 / 1.95**	**0.93 / 0.89**
Florist	**1.95/ 2.15**	**0.76/ 1.14**

Note that the two female-biased role nouns in bold font (above in the judgement task and below in the learning association items) were taken from Kennison and Trofe, 2003. In these cases, average ratings of both male and female respondents are provided; the ratings of female participants are shown on the left hand side of the column while those of the male participants are shown on the right hand side of the column.

2. **Learning Association items:** Proper nouns and role nouns (including mean bias ratings and standard deviations of the latter).

	Proper Nouns	**Role Nouns**	**Bias**	***SD***
1	Thomas	**Knitter**	**1.85 / 1.80**	**0.67 / 0.83**
2	Claire	Football coach	84.71	-15.15
3	David	Beautician	13.27	-10.88
4	Rebecca	Judge	72.75	-14.43
5	James	**Manicurist**	**1.80 / 1.30**	**0.89 / 1.30**
6	Sarah	Farmer	76.80	-14.49
7	Christopher	Dressmaker	22.94	-13.90
8	Lauren	Dean	72.94	-14.04
9	Jack	Ice skater	34.80	-12.66
10	Emma	Boxer	86.27	-15.62
11	Daniel	Model	28.43	-14.33
12	Jessica	Physicist	73.53	-12.62
13	Michael	Cashier	33.67	-15.23
14	Laura	Bricklayer	88.24	-11.26
15	Samuel	Hairdresser	34.60	-21.59
16	Charlotte	Murderer	74.40	-13.12
17	Matthew	Sales assistant	32.55	-13.54
18	Emily	Lorry driver	82.75	-15.50
19	Luke	Au pair	19.39	-16.38
20	Hannah	Engineer	77.00	-13.13
21	Ryan	Birth attendant	30.82	-18.01
22	Amy	Factory manager	79.41	-14.06
23	Joshua	Social worker	27.84	-12.05
24	Sophie	Golfer	77.45	-12.62

Appendix 11 **Filler Role Nouns: Experiment 7**

Note items that were new (i.e. added to the filler list used in Experiments 1-6 (in Appendix 2)) are provided in bold font.

	Condition:	Role nouns:			Condition:	Role nouns:	
		Block 1	**Block 2**			**Block 1**	**Block 2**
1	**Match**	Tailor	Skipper	1	**Match**	Baroness	Princess
2		Grandpa	Gigolo	2		Countess	Wife
3		Policeman	Fireman	3		God mother	Queen
4		Best man	Grandfather	4		Landlady	Madam
5		Craftsman	Step brother	5		Heroine	Aunt
6		Groom	Duke	6		Mistress	Daughter
7		Step father	Milkman	7		Sister	Waitress
8		Count	Barman	8		Nun	Barmaid
9		Merman	Son	9		Spinster	Stewardess
10		Postman	Uncle	10		Hostess	Salesgirl
11		Salesman	Master	11		Milkmaid	Duchess
12		Host	Pope	12		Mermaid	Step mother
13		Bachelor	King	13		Ballerina	Bridesmaid
14		Steward	Hero	14		Step sister	Washerwoman
15		Waiter	Husband	15		Bride	Grandma
16		Priest	Landlord	16		Busgirl	Seamstress
17		Brother	Chairman	17		Policewoman	Lesbian
18		Sir	Prince	18		Grandmother	Matron
19		Knight	Baron	19		Geisha	Maid of honour
20		God father	Ball boy	20		Ball girl	Dame

1	**Mismatch**	Policeman	1	**Mismatch**	Landlady
2		Groom	2		Heroine
3		Postman	3		Mistress
4		Salesman	4		Spinster
5		Bachelor	5		Hostess
6		Steward	6		Bride
7		Waiter	7		Waitress
8		King	8		Princess
9		Son	9		Stewardess
10		Sir	10		Milkmaid
11		Knight	11		Salesgirl
12		Master	12		Duchess
13		Pope	13		Countess
14		Hero	14		Dame
15		Husband	15		God mother
16		Landlord	16		Policewoman
17		God father	17		Grandmother
18		Count	18		Seamstress
19		Gigolo	19		Geisha
20		Baron	20		Lesbian
21		Fireman	21		Matron
22		Grandfather	22		Baroness
23		Milkman	23		Nun

24	Host	24	Step mother
25	Duke	25	Maid of honour
26	Best man	26	Barmaid
27	Barman	27	Wife
28	Step brother	28	Queen
29	Step father	29	Madam
30	Priest	30	Daughter
31	**Skipper**	31	**Sister**
32	**Tailor**	32	**Aunt**
33	**Grandpa**	33	**Step sister**
34	**Busboy**	34	**Busgirl**
35	**Merman**	35	**Washerwoman**
36	**Ballboy**	36	**Grandma**
37	**Uncle**	37	**Mermaid**
38	**Brother**	38	**Countess**
39	----------	39	**Ball girl**

Appendix 12 **Stereotypical & Counter-stereotypical pictures**

1. Architect

2. Boxer

3. Bricklayer

4. Carpenter

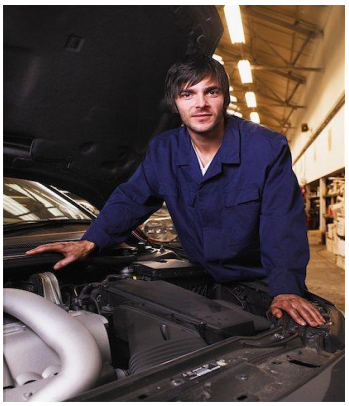
5. Electrician

6. Farmer

7. Golfer

8. Judge

9. Mechanic



10. Soldier

11. Dentist

12. Truck Driver

13. Au pair

14. Ballet dancer

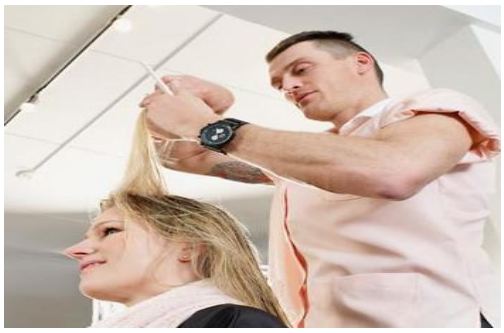
15. Cleaner

16. Flight attendant

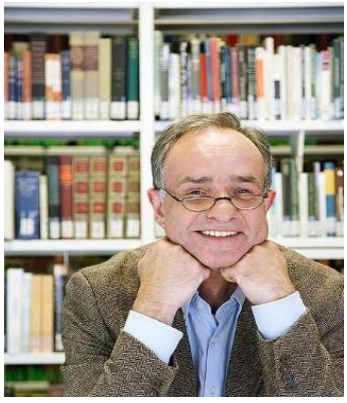
17. Florist

18. Fortune teller

19. Hairdresser



20. Librarian



21. Make-up artist

22. Model

23. Nurse

24. Primary school teacher

Appendix 13 Example Booklet for Experiment 8 (Counter-Stereotype Pictures)

Please answer 4 questions in relation to each of the pictures presented onscreen. When you have answered each of the questions for one picture, press Enter on the keyboard to proceed to the next image and repeat for all 24 pictures.

Please attempt to answer all questions

1. **This is Rebecca. She is a bricklayer**

- How much do you think Rebecca earns each year?

☐ <£10,000 ☐ £10,000-£20,000 ☐ £21,000-£30,000 ☐ £31,000-£40,000 ☐ £41,000-£50,000 ☐ > £50,000

- What are her leisure activities?

- Briefly describe her personal life:

- How satisfied do you think she is with her job?

1	**2**	**3**	**4**	**5**
Extremely Satisfied	Quite Satisfied	Neither Satisfied nor Dissatisfied	Quite Dissatisfied	Extremely Dissatisfied

2. **This is Daniel. He is a librarian**

- How much do you think Daniel earns each year?

☐ <£10,000 ☐ £10,000-£20,000 ☐ £21,000-£30,000 ☐ £31,000-£40,000 ☐ £41,000-£50,000 ☐ > £50,000

- What are his leisure activities?

- Briefly describe his personal life:

- How satisfied do you think he is with his job?

1	**2**	**3**	**4**	**5**
Extremely Satisfied	Quite Satisfied	Neither Satisfied nor Dissatisfied	Quite Dissatisfied	Extremely Dissatisfied

Appendix 14 **Rater's Instructions for the booklet analysis**

Rating categories

To Remember!

- When rating, ask yourself – Are these leisure activities ***typically*** [male/neutral/female]? etc.
- Disregard any gender related information given in the responses (e.g. his, her, he, she) unless directly relevant to the target you are rating e.g. with the category 'traditional vs. non-traditional personal life' terms such as boyfriend, wife etc. may be relevant.
- Rate according to the numbers marked next to the categories on the next page e.g. male vs. female: (1) male (2) neutral and (3) female.

Leisure activities

1. **Male vs. female leisure activities**
   male/neutral/female

2. **High vs. low cost leisure activities**
   expensive / reasonable costs / cheap

3. **Body vs. mind oriented leisure activities**
   physically active/ equally physically and mentally active / mentally active

4. **Social vs. solitary leisure activities**
   social / neutral / solitary

Personal life

1. **Traditional vs. non-traditional personal life**
   traditional /neutral/non-traditional

2. **Happy vs. unhappy personal life**
   happy / neutral / unhappy

Leisure activities

1. **Male vs. female leisure activities**
    1. Male e.g. football, rugby
    2. Neutral e.g. tennis, sight-seeing, reading
    3. Female e.g. dancing, ballet

2 Costs: Expensive vs. Cheap
    1. Expensive e.g. golf
    2. Reasonable e.g. yoga
    3. Cheap e.g. reading

3 Body vs. mind oriented
    1. Physically active e.g. rugby
    2. Physically and mentally active e.g. cooking
    3. Mentally active e.g. reading

4 Social (i.e. group) vs. solitary
    1. Social e.g. football, meeting friends
    2. Neutral e.g. going to museums
    3. Solitary e.g. reading

Personal life

1. **Traditional vs. non-traditional lifestyle**
    1. Traditional e.g. married with children
    2. Neutral e.g. single
    3. Non-traditional e.g. lesbian couple, transgender boyfriend

2. **Happy vs. Unhappy**
    1. Happy e.g. happily married with children
    2. Neutral e.g. lives with housemates
    3. Unhappy e.g. Extremely stressed all the time

Appendix 15 **Ethics Questionnaire (Chapter 5)**

Participant instructions: You must type responses to questions indicating what you think people should do *first* in a number of different situations.

Items: Note that items written in bold font were taken from the original questionnaire of McMinn et al., 1990 while all others were created for the purpose of this thesis.

Neutral

1. A musician gets increasingly nervous before every performance. What should the musician do first?
2. A pedestrian finds £20 in the street. What should the pedestrian do first?
3. A child notices the neighbour's cat is stuck up their tree. What should the child do first?
4. A swimmer is offered performance-enhancing drugs at a competition. What should the swimmer do first?

Female-biased

1. A sales assistant discovers some money is missing from the till. What should the sales assistant do first?
2. A dancer twists an ankle while rehearsing but the show opens tonight. What should the dancer do first?
3. **A nurse discovers a hospital patient has been given blood contaminated with the AIDS virus. What should the nurse do first?**
4. **A librarian notices that many students are being too loud in the library. What should the librarian do first?**

Male-biased

1. **A business executive discovers a long-time employee has been stealing from the company. What should the executive do first?**
2. **A truck driver has just witnessed a pedestrian being hit by a car. What should the truck driver do first?**
3. An engineer discovers the building is unsafe for inhabitants. What should the engineer do first?
4. A mathematician notices a colleague is having difficulty settling into the department. What should the mathematician do first?