

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/96719>

Copyright and reuse:

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

**Advancing the applicability of absolute implicit measures:
Using the Simple Implicit Procedure (SIP) to measure
responses to pathogen threats.**

by

Brian O'Shea

B.A., National University of Ireland, Galway, 2010

M.Sc., London School of Economics and Political Science, 2012

A thesis submitted in partial fulfilment of the requirements for the

degree of

Doctor of Philosophy in Psychology

University of Warwick, Department of Psychology

April, 2017

Table of Contents

Acknowledgements.....	iv
Declaration.....	vi
Inclusion of Published Work	vii
Summary.....	ix
Abbreviations.....	x
Tables.....	xi
Figures.....	xiii
Chapter 1: Introduction.....	1
Abstract.....	2
Origins of research into implicit social cognition.....	3
Unobtrusive and indirect research methods.....	5
Selective attention and implicit memory	6
The development of implicit measures	8
What is an implicit measure?.....	9
Limitations of existing implicit measures.....	11
Implicit and explicit correlations and predicting behaviour	13
Theoretical basis of implicit measures.....	16
Overview of the thesis	18
Chapter 2: Disease rates are associated with racial prejudice across the US and the world....	21
Abstract.....	22
Introduction.....	23
Parasite stress theory (PST)	23
Present study	28
Study 1 (US) and Study 2 (World)	29
Method.....	29
Results.....	33
Study 3	38
Method.....	38
Results.....	41
Discussion.....	43
Chapter 3: Measuring implicit attitudes: A positive framing bias flaw in the Implicit	
Relational Assessment Procedure (IRAP)	46
Abstract.....	47
Introduction.....	48
Absolute vs. relative measures of implicit cognition.....	48

Framing effects and language biases	53
Present study	55
Method	56
Results	62
Discussion	70
Chapter 4: The Simple Implicit Procedure (SIP): A new method for measuring implicit cognition	78
Abstract	79
Associative versus propositional measures of implicit cognition	81
Limitations of the IRAP	84
The Simple Implicit Procedure: SIP	87
Overview of studies	91
Study 1	91
Method	94
Results	100
Discussion	114
Study 2	116
Method	117
Results	120
Discussion	128
Study 3	129
Method	131
Results	133
Discussion	137
General discussion	137
Chapter 5: Some words are more like pictures: Word type response biases in reaction time tasks	144
Abstract	145
Introduction	146
Overview	147
Study 1	148
Method	149
Results	150
Discussion	152
Study 2	152
Method	153

Results.....	156
Discussion.....	158
Study 3	158
Method	158
Results.....	159
Discussion.....	161
General discussion	161
Chapter 6: Disgust versus fear: Using the SIP to measure racial prejudice.....	164
Abstract	165
Introduction.....	166
Present study	169
Method	170
Results.....	177
Discussion.....	185
Chapter 7: Concluding remarks and future directions	191
References.....	202
Appendix 1: Chapter 2	241
Appendix 2: Chapter 3	257
Appendix 3: Chapter 4	267
Appendix 4: Chapter 5	274
Appendix 5: Chapter 6.....	281

Acknowledgements

My ideas, experimental rigour, coding abilities and cogency have been nurtured and constantly pushed to new heights I never thought I could attain, by my two wonderful supervisors, Gordon and Derrick. I am forever indebted to you for imparting so much of your wisdom. Thank you.

This PhD thesis would not have been possible without the generous support from the Psychology Department at the University of Warwick. I must also acknowledge the Experimental Psychological Society, the European Association of Social Psychologists, the Society for the Psychological Study of Social Issues, the Society for Personality and Social Psychology, the British Psychological Society and the Psychology Postgraduate Affairs Group for funding to support data collection, study visits or disseminating my research at international conferences.

I would like to extend my gratitude to the Learning and Implicit Processes Lab at the University of Ghent, especially to Jan De Houwer for giving me a place to work and the excellent advice given. Maarten De Schryver also greatly assisted me with scraping data from online repositories. Kate Ratliff, Colin Smith and all their lab members at the University of Florida provided me with three months of great memories and valuable intellectual contributions to this thesis. GO GATORS.

Blanca, I know dealing with me towards the end of this thesis submission was not easy but I am eternally grateful for everything you did for me. Many thanks to those in H122/124 for making my time at Warwick so enjoyable and offering a safe space to rant. Thanks to all the Undergraduates, Masters, PhD Students, Post-doctorates and Professors at Warwick who gave me a medium where I could express my ideas and receive feedback. Anna and Steve, I really appreciate your comments on one of the chapters. I would also like to thank all the

participants who volunteered to take part in my research, especially those who had to repeat an implicit measure a number of times.

The new knowledge found in this thesis would not have been made available (at least this soon), were it not for my loving parents, Anne and Michael. You did everything you could to help me overcome my weakness, nourished my strengths and supported my decisions. I never felt pressure to aim for this level of educational attainment but I knew it would make you proud. Go raibh míle maith agaibh.

Declaration

This thesis is submitted to the University of Warwick, in support of my application for the degree of Doctor of Philosophy. This thesis has not been submitted for a degree at another university. I hereby confirm that I completed this thesis independently, including all the experimental data collection and analysis.

Inclusion of Published Work

All the following publications and manuscripts under review were submitted during the period of my PhD registration. The copyright of the published papers resides with the publishers but the reproduction of the papers in this thesis is permitted under the terms of the Copyright agreement.

Sections of Chapters 2 and 6 were published in:

O'Shea, B. (2016). Parasites and their implications for social and cultural psychology.

PSYPAG Quarterly, 99, 30-34.

Chapter 2 is expected to be submitted for publication by mid-May:

O'Shea, B., Watson, D.G., Brown, G. D.A., & Fincher, C. L. (in preparation). *Disease*

rates are associated with racial prejudice across the U.S. and the world.

Chapter 3 was published as:

O'Shea, B., Watson, D. G., & Brown, G. D. A. (2016). Measuring implicit attitudes: A positive framing bias flaw in the Implicit Relational Assessment Procedure (IRAP). *Psychological Assessment*, 28, 158-170.

Chapter 4 is expected to be submitted for publication before the end of the year.

O'Shea, B., Brown, G. D. A., De Houwer, J., & Watson D. G. (in preparation). *The Simple*

Implicit Procedure (SIP): A new method of measuring implicit cognition.

Chapter 5 is expected to be submitted for publication before the end of August:

O'Shea, B., Watson D. G., & Brown, G. D. A. (in preparation). *Some words are more like*

pictures: Word type response biases in reaction time tasks

Ideas from the following manuscripts also appear throughout the thesis:

O'Shea, B. (2015). Capitalism versus a new economic model: Implicit and explicit attitudes of protesters and bankers. *Social Movement Studies*, 14, 311-330.

O'Shea, B. (revise and resubmit). *Attitudes towards atheism and religious belief: Limitations*

with the Implicit Relational Assessment Procedure (IRAP).

Conway, J. G., Pogge, G., Redford, L., **O'Shea, B.**, Klein, R. A., Ratliff, K. A. (under review).

Can carelessness be captured? Assessing careless responding in attitudes toward novel stimuli.

Summary

In all areas of psychological research, particularly in the area of implicit cognition, investigations depend on the tools of measurement. A central aim of the work presented in this thesis was to improve the accuracy and usability of tools that measure implicit cognition. This development was situated in the context of parasite-stress theory (PST) and racial prejudice.

Chapter 1 explains the origins of research on implicit cognition and what makes a measure implicit. Chapter 2 introduces PST, which predicts that increased exposure to infectious diseases will lead to avoidance of, or disdain towards, out-groups because such avoidance will reduce the likelihood of contracting an illness. This prediction was confirmed for both implicit and explicit attitudes, using complementary correlational and experimental methodologies.

The findings reported in Chapter 3 indicate that participants remember positive statements better than they remember negative statements and hence display a Positive Framing Bias (PFB) in the Implicit Relational Assessment Procedure (IRAP). Chapter 4 introduces the Simple Implicit Procedure (SIP) which is user-friendly and is not subject to the PFB. Estimates of implicit attitudes obtained from the SIP correlate with explicit measures, provide increased specificity of where an individual's implicit biases lie, have acceptable reliability and are not limited by practice/experience effects. The SIP can assist researchers in devising strategies aimed at understanding and ameliorating the precise mechanisms driving prejudice.

Chapter 5 describes how using nouns, verbs or adjectives in implicit measures can affect outcomes. In Chapter 6, the SIP was used to show that females' racial biases remain stable over time, even when primed with diseases. Males appear to be particularly susceptible to expressing increased prejudice when primed with disease images.

In summary, this thesis identifies various response biases that can greatly influence the results obtained from measures of implicit attitudes. Recommendations for overcoming these biases are described.

Abbreviations

BF	Bayes Factor
BIAT	Brief Implicit Association Test
BIS	Behavioural Immune System
GNAT	Go/No-Go Association Task
IAT	Implicit Association Test
IRAP	Implicit Relational Assessment Procedure
NHST	Null Hypothesis Significance Testing
PFB	Positive Framing Bias
PST	Parasite Stress Theory
SIP	Simple Implicit Procedure
SC-IAT	Single Category Implicit Association Test
SPFT	Sorting Paired Features Task
ST-IAT	Single Target Implicit Association Test
RRT	Relational Responding Task
RT	Reaction Time
VJT	Valence Judgement Task

Tables

Table 2.1	26
Table 2.2	34
Table 2.3	35
Table 2.4	37
Table 4.1	95
Table 4.1	132
Table 4.3	134
Table 4.4	136
Table 5.1	150
Table 5.2	154
Table 5.3	155
Table 6.1	172
Table 6.2	184
Table S2.1	244
Table S2.2	245
Table S2.3	246
Table S2.4	247
Table S2.5	248
Table S2.6	249
Table S2.7	250
Table S2.8	251
Table S2.9	252
Table S2.10	254
Table S3.1	258

Table S3.2	259
Table S5.1	279
Table S5.2	280

Figures

Figure 2.1	42
Figure 3.1	51
Figure 3.2	66
Figure 3.3	68
Figure 3.4	69
Figure 4.1	83
Figure 4.2	89
Figure 4.3	90
Figure 4.4	105
Figure 4.6	109
Figure 4.7	112
Figure 4.8	122
Figure 4.9	123
Figure 4.10	125
Figure 4.11	127
Figure 4.12	135
Figure 5.1	151
Figure 5.2	157
Figure 5.3	160
Figure 6.1	174
Figure 6.2	177
Figure 6.3	180
Figure 6.4	181
Figure 6.5	183

Figure S2.1	255
Figure S2.2	256
Figure S3.1	266
Figure S5.1	277

Chapter 1: Introduction

Abstract

The origins of one of the most studied topics in social psychology - Implicit Social Cognition - are discussed, as are the unobtrusive/indirect research methods used to overcome the problems (e.g., a tendency to make socially desirable responses) that are apparent when using explicit self-reports. The Implicit Association Test is introduced along with a brief overview of other popular implicit measures. It is precisely defined what an implicit measure is, emphasising that automaticity is a core feature. A section is devoted to explaining under what conditions implicit and explicit measures are or are not related. The value of using both measures to predict behaviour is also explained. I conclude by discussing some theoretical issues relating to implicit measures and end with an overview of the thesis.

Origins of research into implicit social cognition

Most psychologists agree that to gain a comprehensive understanding of an individual's behaviour one needs knowledge not only of the external contexts in which an individual is situated but also of their internal psychological attributes (i.e., attitudes, stereotypes and personality traits). The cognitive revolution of the 1960s and 70s challenged behaviourism's grip on psychology and restored the scientific respectability of the study of internal psychological processes (Miller, 2003). Without this impetus, the study of implicit¹ cognition and investigations into the importance of unconscious/unexpressed biases influencing behaviour could have been further delayed.

The prominence in social psychology of the study of individuals' attitudes was apparent as far back as the 1930s, when George Allport (1935), described attitudes as the "most distinctive and indispensable concept" in the discipline. In the 1940s the behaviourist Leonard Doob (1947, p.136) defined an attitude as "an implicit, drive-producing response considered socially significant in the individual's society". It was generally accepted, albeit implicitly, that attitudes were influenced by unconscious mechanisms (see Greenwald & Banaji, 1995). This (implicit) acceptance was likely due to psychoanalytic theory according to which attitudes could be influenced by unconscious processes. Although psychologists often want to dissociate

¹ The term "implicit" is often used as a synonym for other labels such as "unconscious", "unaware", "intuitive", "indirect" and "automatic". Likewise, "explicit" often overlaps with "conscious", "aware", "analytic", "direct" and "controlled". The implicit-explicit distinction was used throughout this thesis. Additionally, the terms "implicit cognition", "implicit attitudes", "implicit social cognition" and "implicit bias" are synonyms of each other in this thesis.

themselves from Sigmund Freud's controversial and often unfalsifiable ideas, he was one of the first individuals to bring the idea of the unconscious to the mainstream (Freud, 1915, 2005).

Researchers aimed to measure attitudes using more objective methods such as questionnaires. Nevertheless, the near-universal use of self-report questionnaires throughout social psychology's past has led to numerous problems. One of the major problems was construct validity weaknesses, as seen by the lack of correspondence between attitudes and behaviour (Greenwald, 1990, see also Nisbett & Wilson, 1977). For example, LaPiere (1934) reported that when he visited 251 accommodation venues with a Chinese couple, only one venue refused them admission. Following these visits, letters were sent to all 251 venues asking for a response to the question "Will you accept members of the Chinese race as guests in your establishment?" and over 90% responded that they would not. Regardless of the limitations of the study, (e.g., the Chinese couple was accompanied by an individual in a high-status profession; the person rejecting the letter request may have been different from the person who accepted their face to face request), LaPiere's study emphasises the values of using unobtrusive/indirect² research methods instead of confining oneself to explicit self-reports (direct measures). Sheeran (2002) provides a more recent review of the relationship between attitudes (intentions) and behaviour.

² The terms "indirect" and "direct" are often used to describe the procedural characteristics of a measurement procedure, while the terms "implicit" and "explicit" are often used to describe the psychological features or attributes assessed by measurement procedures. These classifications follow recommendations by De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009; see also De Houwer & Moors, 2010).

Unobtrusive and indirect research methods

“Unobtrusive” refers to methods in which participants are generally unaware that their behavioural responses are under investigation, while the term “indirect” refers to when participants are aware their responses are being monitored but it is unclear exactly which aspect of their behaviour is being assessed (Banaji & Greenwald, 2016). In the 1970’s, two social forces made unobtrusive research methods both more appealing and more popular: (1) The rise in the scientific study of prejudice across the US due to racial tensions, and (2) the growing realisations that participants often respond in a socially desirable manner when making explicit self-reports (Banaji & Greenwald, 2016; Jones & Sigall, 1971). The first two successful uses of unobtrusive methods relating to racial prejudice showed that when a black or white individual using a telephone called looking for help, white callers were more likely to receive assistance, and when an open unsent student application with a portrait photograph attached was left at an airport telephone booth, participants were more likely to voluntarily submit applications for the white students (Benson, Karabenick, & Lerner, 1976; Gaertner & Bickman, 1971; for a review of unobtrusive methods see Crosby, Bromley, & Saxe, 1980 and for a recent example of the unobtrusive technique showing that white people offer less help to black relative to white individuals in an emergency situation, see Kunstman & Plant, 2008).

Crosby and colleagues (1980) indicated that discrimination is more prevalent than explicit self-reports would lead us to believe, and that remote (not face to face) interactions were more likely to give rise to stronger racial prejudice when unobtrusive methods were used. Unobtrusive methods reduce the Hawthorn effect (i.e., the fact that participants’ knowledge of being in an experiment modifies their behaviour from how they would have responded without that knowledge: Adair, 1984). However, unobtrusive methods can be ethically problematic, require more resources to run than simple questionnaires, and the context is more difficult to control (Blackstone, 2017). Of most relevance, they do not provide an adequate opportunity to

measure individual differences because asking participants to fill out questionnaires or demographic information would rouse suspicions. Indirect measures, such as implicit measures, were developed to overcome the limitation of unobtrusive methods not assessing individual differences.

Selective attention and implicit memory

Research relating to both selective attention and implicit memory greatly influenced the development of research into implicit social cognition (Payne & Gawronski, 2010). With respect to selective attention, a key distinction was one between controlled and automatic/involuntary modes of information processing (e.g., Shiffrin & Schneider, 1977). For example, our attention will be instantly drawn if we detect words of importance originating from unattended sources, such as hearing one's name (Moray, 1959). The development of sequential priming techniques (e.g., Devine, 1989; Dovidio, Evans, & Tyler, 1986; Fazio, Jackson, Dunton, & Williams, 1995; Gaertner & McLaughlin, 1983) was strongly influenced by the selective attention literature (i.e., weakly learned associations need cognitive effort to retrieve but strong associations will be activated automatically)³. These sequential priming measures did however suffer from low reliability and produced small effect sizes (Payne & Gawronski, 2010), and the existence of these weaknesses could explain the popularity of other implicit measures (see below).

Research on implicit memory pioneered by Jacoby and colleagues (e.g. Jacoby, Toth, Lindsay, & Debnor, 1992) (for review see Schacter, Chiu, & Ochsner, 1993) strongly

³ Since the focus of this thesis is not related to priming measures to estimate an individual's implicit social cognition, these measures will not be discussed further. For more recent information on these types of measures see Cameron, Brown-Iannuzzi, & Payne, 2012; Payne, Cheng, Govorun, & Stewart, 2005 and Wentura & Degner, 2010)

influenced Greenwald and Banaji's (1995) seminal paper on implicit cognition (see also Banaji, 2001). Essentially, studies of implicit memory showed that participants found it easier to perceive stimuli that they had previously seen (perceptual fluency) but attributed this ease to characteristics of the stimulus, rather than to the recent past encounter. These ideas led Greenwald and Banaji (1995) to coin the term “implicit cognition” and define it as “An implicit **C** is the introspectively unidentified (or inaccurately identified) trace of past-experience that mediates **R**. In this template, **C** is the label for a construct (such as attitudes), and **R** names the category of responses (such as object evaluative judgements) assumed to be influenced by that construct” (p. 5). Their review focused on how implicit cognition was specifically related to attitudes, stereotypes and the self, and this relationship is also a major focus of this thesis.

In implicit memory research, experimenters normally have perfect control over the stimuli previously presented, while in implicit cognition, experimenters generally have little control over previously presented stimuli (e.g., an individual's life history) and, therefore, require more mentalistic explanations for behavioural responses (i.e., mental associations, particularly evaluative and semantic associations, see Greenwald et al., 2002; Hahn & Gawronski, 2015). An exception is when participants are exposed to completely new stimuli. Any attitude (both implicit and explicit) that subsequently develops following a positive or negative induction (i.e., through evaluative conditioning: see Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010) to novel stimuli (e.g., an unknown group or tribe), is most likely due to the controlled exposure rather than prior experiences (e.g., Gregg, Seibt, & Banaji, 2006; Olson & Fazio, 2001). This experimental set up has been said to involve attitude formation and refers to the initial change from having no attitude towards an object to having some attitude, either positive or negative, towards it (Oskamp & Schultz, 2005).

The development of implicit measures

Greenwald and Banaji (1995) asserted that the measurement of individual differences in implicit cognition is likely to be possible and described judgement latency (i.e., reaction time (RT)) measures as a potentially fruitful avenue to pursue this goal. They predicted that “when such measures become available, there should follow the rapid development of a new industry of research on implicit cognitive aspects of personality and social behaviour” (p. 20). Within three years, the same authors developed and published the first and still most popular RT task that aims to measure implicit attitudes, stereotypes and self-concept at the individual level. The measure was called the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) and precisely as they predicted, this task led to an acceleration in research into implicit cognition, with the area being described as “one of the liveliest and most active research areas in social psychology” (Payne & Gawronski, 2010, p. 9). The initial IAT publication has to date (July 2017) been cited almost 9,000 times. Development of the IAT to run online through Project Implicit (<https://implicit.harvard.edu/implicit/>), the world’s largest online virtual laboratory greatly accelerated the speed at which the IAT could be validated (e.g., Greenwald, Nosek, & Banaji, 2003; Greenwald & Nosek, 2001; Nosek et al., 2007; Nosek, Banaji, & Greenwald, 2002; Nosek, Greenwald, & Banaji, 2007) .

Millions of people from all over the world have completed various versions of the IAT through Project Implicit (e.g., Old-Young IAT, Fat-Thin IAT, European American-African American IAT). Participants are incentivised to complete IATs because after completing the task they are given their implicit bias score towards the categories to which they were responding. In a typical IAT, such as the Young–Old IAT, participants are presented successively with various pictures of young and old individuals as well as positively and

negatively valenced words⁴, and must sort the items using the correct keypress. In one of the two critical blocks, participants should press the “E” key on a computer keyboard if a positive word or a picture of a young person appears and press the “I” key if a negative word or a picture of an old person appears (congruent block). In the other critical block, participants should press “E” if a positive word or a picture of an old person is shown and they must press the “I” key for a negative word or a picture of a young person (incongruent block).

This task aims to measure biases participants have in associating concepts (old and young) with valenced words. The stronger the association, the more natural the sorting task will feel and hence, result in faster responses (congruent block), while weaker associations will result in a slowing down of processing, due to the need to make use of unaccustomed pairings in memory (incongruent block; Greenwald et al., 2002). Rather than using positive and negative valenced words to measure attitudes, stereotypical words could be used instead (e.g., healthy, lively, frail, slow, etc.) to measure stereotypes that individuals have.

What is an implicit measure?

Currently, the term ‘implicit social cognition’ is generally used to refer to research in social psychology that uses computerised RT measurement instruments to infer an individual’s psychological attributes (i.e., attitudes, stereotypes, self-esteem, etc.) without asking an individual to report their psychological attributes directly (Hahn & Gawronski, 2015). Implicit measures have been defined as “the outcome of a measurement procedure that results from automatic processes by which the to-be-measured attribute causally determines the outcome”

⁴ Throughout the thesis, “positively valenced words” will be used interchangeably with “positive words”. Examples of positive words include, happy, joy, love. Likewise, “negatively valenced words” will be used interchangeably with “negative words”. Examples of negative words include, evil, hurt, sick.

(De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009, p. 347). Automaticity has been argued to be one of the core features of an implicit measure and occurs when the impact of the “to be measured” attribute on an individual’s responses is uninfluenced by certain goals, substantial cognitive resources, awareness or substantial time (Bargh, 1994; De Houwer, 2006; De Houwer & Moors, 2012). Therefore, quick and accurate reactions to stimuli are necessary to limit an individual’s ability to exercise strategic control over their responses. This automaticity aspect or feature distinguishes implicit measures from traditional instruments that rely on explicit self-reports (Gawronski & De Houwer, 2014). Consequently, the speed or accuracy with which an individual responds to or associates stimuli in implicit measures is used to infer their psychological attributes (De Houwer, 2009; De Houwer, 2003; Nosek & Banaji, 2001). With implicit measures (and also explicit measures), an inference must inevitably be used because it is not possible to directly observe psychological attributes.

An ideal measure of implicit attitudes would provide an accurate index of the extent to which an individual possesses the psychological attributes that the measure was designed to capture (De Houwer et al., 2009). To validate an implicit measure, there must be evidence that variation in the “to be measured” attribute (e.g. racial bias), causes variations in the measurement outcome (i.e., the measure’s score) but an understanding of how the measurement outcome is detecting variations in the psychological attribute is also necessary (Borsboom, Mellenbergh, & van Heerden, 2004; Wentura & Rothermund, 2007). Both correlations and experimental approaches are useful when validating implicit measures. But collecting correlational data is the most time and cost effective way to validate an implicit measure (De Houwer et al., 2009). The more evidence that is accumulated showing that the implicit measure correlates in the expected manner with other measures of psychological attitudes (e.g., explicit self-reports), the more the likelihood that the correlations are due to a third factor is reduced

(e.g., Nosek & Smyth, 2007). Correlational analysis will be mainly used to validate the new implicit measure introduced in this thesis.

Limitations of existing implicit measures.

The period following the IAT's first publication has been described as the "age of measurement" in social psychology, due to the development of various implicit measures each of which aimed to measure psychological attributes accurately (Nosek, Hawkins, & Frazier, 2011). New implicit measures were developed mainly because of limitations in the IAT. But these new measures were further challenged by the need to preserve the strong psychometric properties that the IAT achieves. The brief IAT (BIAT) was developed to give researchers a tool that could be used to measure implicit bias in a shorter amount of time. However, like the IAT, the BIAT only measures implicit biases relatively (e.g., attitudes to fat people relative to attitudes to thin people). It is therefore impossible to determine using these IATs whether this bias is the result of a strong/weak pro-thin bias, a strong/weak anti-fat bias, or some combination of the two (see Blanton & Jaccard, 2006; Roddy, Stewart, & Barnes-Holmes, 2010, 2011). Likewise, the IAT cannot be used to determine how interventions that aim to increase or reduce implicit biases have their effect (e.g., a difference in implicit attitudes could be reduced by acting on the "Thin Person" category, the "Fat Person" category, or both; see Lai et al., 2014).

Another limitation of the IAT arises because some categories do not have an obvious comparison group. For example, when assessing implicit self-esteem, researchers can use the IAT to measure the positive and negative associations a person has with the self in comparison to a specified/unspecified other (or with "me" in comparison to "not me"). The type of comparison category (i.e., specified vs. unspecified other) used affects implicit self-esteem results (Karpinski, 2004). Therefore, a more appropriate approach would measure only evaluative associations with the self, without the need to use a complementary category.

To address these problems, RT tools that attempt to measure absolute attitudes and do not require a relative comparison to another group have been developed. For example, the go/no-go (Nosek & Banaji, 2001) and the extrinsic affective Simon task (De Houwer, 2003b) both claim to measure implicit attitudes non-relatively/absolutely. However, both suffer from problems ranging from a high level of task difficulty to low reliability (Bar-Anan & Nosek, 2014; De Houwer & De Bruycker, 2007). Variations of the IAT that could be described as single concept IATs, such as the Single Target/Category IAT (e.g., Bluemke & Friese, 2008; Karpinski & Steinman, 2006) have, however, shown promise for measuring absolute implicit attitudes. The Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes et al., (2006)) is another absolute implicit measures, but see Chapter 3 for limitations of the IRAP's assessment of absolute implicit attitudes.

The Single-Block IAT (Sarah Teige-Mocigemba, Klauer, & Rothermund, 2008), the Recoding Free IAT (Klaus, Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009) and the Sorting Paired Features Task (Bar-Anan, Nosek, & Vianello, 2009) were developed to overcome the problem of block structure influencing the accurate measurement of implicit biases. For example, the order in which a participant completes the compatible and incompatible blocks can influence IAT score (see Nosek, et al., 2007; Teige-Mocigemba, Klauer, & Sherman, 2010). Efforts have been made to reduce the order effect in the IAT (Nosek, Greenwald, & Banaji, 2005) but it is nevertheless difficult to determine the magnitude of order effects at the individual level and hence it is not possible to fully remove effects of this confound (De Houwer et al., 2009). Another reason for removing the block structure in the IAT was that the salience of the items within a block, not the associations between the items, can lead to an IAT effect (see Rothermund & Wentura, 2001, 2004).

Several authors have identified “unwanted factors” (Friese & Fiedler, 2009, p. 230) or “nonassociative influences” (Rothermund & Wentura, 2010, p. 234) that influence the

magnitude of IAT effects. If IAT effects can be caused by processes other than just mental associations in memory, and if it is not clear which processes influence the IAT effect, then the meaning of the effect becomes ambiguous (see Fiedler, Messner, & Bluemke, 2006). There is increasing evidence that findings relating to implicit measures can reflect several processes, such as salience of the stimuli (e.g., Houben & Wiers, 2006; Klaus Rothermund, Wentura, & De Houwer, 2005), similarity between the stimuli (De Houwer, Geldof, & De Bruycker, 2005) and cognitive ability of participants (e.g., Back, Schmukle, & Egloff, 2005; McFarland & Crouch, 2002). Greenwald et al.'s (2003) D-algorithm has greatly reduced the problem relating to cognitive ability because it accounts for an individual's variation in responding across the IAT. More research is needed to explain the relative impact that salience and similarity can have on IAT scores (De Houwer et al., 2009, see Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005, and Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007, for methods that can account for the influence of these various processes). IAT researchers have acknowledged these problems, emphasising that no measure is perfect but yet the measure can still be useful (Greenwald & Sriram, 2010): see Fazio and Olson (2003), Gawronski (2009), Gawronski and De Houwer (2014), and Nosek et al. (2011), for reviews and lists of other implicit measures that have received less attention.

Implicit and explicit correlations and predicting behaviour

Correlations between implicit and explicit measures of attitudes vary widely, from weakly positive ($r = .250$; e.g. thin people-fat people) to strongly positive ($r = .780$; democrats-republicans) with a median correlation of .48 found for 56 domains across 6000 participants (Nosek, 2007). There are a number of possible reasons for this disparity, with the most obvious being self-presentation/socially desirable responding. This is especially true for socially sensitive topics, where demand characteristics and/or impression management may distort self-report responses (e.g. Fazio, 2007; Holtgraves, 2004). Implicit and explicit correlations have

been shown to be higher for affective responses (emotions and feelings about the attitude object) on explicit measures compared to more cognitive responses (thoughts and beliefs about the attitude object) on explicit measures (Smith & Nosek, 2011, for a review see Spence & Townsend, 2008).

Reducing the time allocated for a participant to think about their response on the explicit measures produces higher correlations between implicit and explicit measures (Ranganath, Smith, & Nosek, 2008). Another crucial aspect to consider is that implicit and explicit correlations will be stronger for attitudes that are more familiar or beliefs that are important or well elaborated in memory, as opposed to ones that are rarely thought about or believed to be irrelevant (Nosek, 2007). Conceptual correspondence and structural fit between implicit and explicit measures increases the correlations between these two measures. Additionally, due to averaging out noise and the inclusion of more trials, relative scores (e.g., IAT scores) rather than absolute scores (e.g., SC-IAT scores) show stronger implicit and explicit correlations (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005; Payne, Burkley, & Stokes, 2008).

Lack of a perfect correlation between implicit and explicit measures has been cited as evidence for the distinct constructs that these measures assess (Greenwald & Nosek, 2008; Nosek & Smyth, 2007). Yet the divergence could also be due to a number of other factors (e.g., awareness, need for cognition, structural features, for a review see Hofmann, Gschwendner, Nosek, & Schmitt, 2005). Importantly, much research demonstrates the practical value of implicit measures for predicting human behaviour (Perugini, Richetin, & Zogmaister, 2010), especially spontaneous behaviours (for review see Friese, Hofmann, & Schmitt, 2009) and shows that implicit measures can provide information that is distinct from explicit measures (Nosek et al., 2011).

Perugini, Richetin, and Zogmaister (2010) described how implicit measures could contribute to predicting behaviour over and above explicit measures. These include: (1)

separate patterns with implicit measures, but not explicit measures uniquely predicting behaviour (2) additive patterns in which both implicit and explicit measures contribute to predicting behaviour, (3) double dislocation patterns where both measures uniquely predict different types of behaviour (4) moderation patterns where both measures predict behaviour under different conditions, (5) multiplicative patterns where both measures interactively predict behaviour. All these patterns have been shown in the literature. However, the boundary conditions specifying when each will occur are not thoroughly understood, making it difficult to make a priori predictions (Gawronski & De Houwer, 2014).

The most thorough evidence emphasising the value of implicit measures was provided by a meta-analysis of studies using the IAT, which showed it to predict stereotyping or racially prejudicial behaviour better (average $r = .236$) than did explicit self-report measures (average $r = .118$; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; see also Cameron, Brown-Iannuzzi, & Payne, 2012, relating to sequential priming). However, see Oswald, Mitchell, Blanton, Jaccard, and Tetlock (2013), for a more critical view of the predictive validity of the IAT. They showed that that IAT only weakly predicted racial attitudes and stereotypes ($r = .148$) and stated that “the IAT provides little insight into who will discriminate against whom” (p. 188). For a recent defence of the IAT's predictive abilities see Greenwald, Banaji, and Nosek (2015).

If participants are tired, distracted or rushed, they are more likely to respond based on implicit biases than when they have energy, are concentrating, focused or unhurried (Strack & Deutsch, 2004). For example, Friese, Hofmann, and Wänke (2008) found that when participants' self-regulation resources were reduced, they were more likely to respond behaviourally (eating or drinking) in accordance with their implicit attitudes. In contrast, when participants maintained these control resources, their behavioural responses were better predicted by their explicit attitudes. Furthermore, implicit measures have been shown to be better at predicting behaviours of individuals with a preference for intuitive thinking styles,

while explicit measures are better for those with a preference for rational thinking styles (e.g., Richetin, Perugini, Adjali, & Hurling, 2007).

Other examples of the usefulness of implicit measures for predicting behaviour include: (1) countries with stronger implicit biases of associating males rather than females with science and maths, predict larger performance gaps between males and females in these disciplines (Nosek et al., 2009), (2) those with low self-esteem on implicit measures exhibit various defensive behaviours (Jordan, Spencer, Zanna, Hoshino-Browne, & Correll, 2003) (3) more strongly associating the self with death prospectively predicted suicide ideation as well as suicide attempts (Nock et al., 2010; Nock & Banaji, 2007) and (4) higher implicit racial biases predicted increased job interview invitations to racial in-group members (Rooth, 2010). Importantly, neither implicit and explicit measures can be described as a “truer” measure of one’s beliefs, because both predict unique aspects of behaviour (Banaji, Nosek, & Greenwald, 2004). To clarify, explicit measures are generally better at predicting political preferences (Friese, Smith, Koeber, & Bluemke, 2016) and consumer behaviour (Friese, Wänke, & Plessner, 2006), while implicit measures are particularly suited when addressing intergroup attitudes/interactions (Greenwald et al., 2009). The next section aims to explain why this may be the case.

Theoretical basis of implicit measures

The two most prominent theories aiming to explain correlations between implicit and explicit measures are the *motivations and opportunity as determinants* (MODE) model (Fazio & Olson, 2014; Fazio, 1990) and the *Associative Propositional Evaluation* (APE) model (Gawronski & Bodenhausen, 2006, 2011). Both these models make similar predictions but differ in a few subtle ways. Motivation (e.g., social desirability; Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Dunton & Fazio, 1997) and the opportunity (e.g., self-regulation resources; Hofmann, Rauch, & Gawronski, 2007) are the primary determinates

explaining implicit and explicit correlations in the MODE model. In the APE model, however, cognitive consistency (i.e., rejecting affective racial biases in favour of more explicit egalitarian values) is an important factor in explaining implicit and explicit correlations (Brochu, Gawronski, & Esses, 2011; Gawronski, Peters, Brochu, & Strack, 2008).

The MODE model assumes that the same underlying representations are measured using direct and indirect methods, while the APE model assumes they are part of a distinct but mutually reinforcing processes (i.e., associative and propositional processes). The MODE model also assumes that deliberate processing reduces implicit and explicit correlations, while in the APE model deliberating on information that is consistent with an activated association (i.e., implicit bias) will increase implicit and explicit correlations (e.g., Galdi, Gawronski, Arcuri, & Friese, 2012; Peters & Gawronski, 2011). For a full review of the numerous dual process theories of human cognition see Chaiken and Trope (1999) and Strack and Deutsch (2004).

Early theorising around implicit attitude measures assumed that they provide direct access to stable evaluative representations that have their roots in long-term socialisation experiences (see Greenwald & Banaji, 1995). However, new evidence has shown that

implicit attitudes are highly susceptible to contextual influences (for a review see Blair, 2002). Furthermore, the debate surrounding whether implicit biases represent person-based (e.g., personal attitudes; Fazio et al., 1995) or situational-based approaches (e.g., awareness of cultural stereotypes; Devine, 1989) has been going on for some time. More recently, this debate has re-emerged in the form of personal versus extra-personal associations in the IAT (e.g., Nosek & Hansen, 2008; Olson, Fazio, & Han, 2009). However, it has been argued that making a distinction between situational and personal views is not warranted due to the automatic effects of implicit views (Banaji, 2001; Gawronski & LeBel, 2008; Nosek & Hansen, 2008a). It has also been stated that few arguments remain supporting the claim that IAT effects are

causally influenced by extra-personal views (De Houwer et al., 2009). The term “implicit” has often been used synonymously with “unconscious” but one must ask the question: Are implicit measures uncovering unconscious representations? The available evidence challenges the notion that implicit measures offer a window into people’s unconscious processes (e.g., Gawronski, Hofmann, & Wilbur, 2006; Hahn & Gawronski, 2014). To clarify, the majority of participants in Monteith, Voils, and Ashburn-Nardo's (2001) study expressed that they found the incongruent block on the Race IAT (Black-Positive) more challenging, and they felt guilty about it to the extent that they attributed the bias to racial prejudice. Therefore, individuals appear to have much greater introspective access to their mental representation than was originally assumed (Payne & Gawronski, 2010). For example, across five different social groups, participants were surprisingly accurate when predicting their implicit biases (Hahn, Judd, Hirsh, & Blair, 2014), emphasising that participants can have introspective awareness of their implicit bias.

Of note, the mental processes that result in conscious experience for introspection to occur are largely a mystery, and similarly, our knowledge of what implicit measures measure is less mature than our knowledge about what implicit measures do (Nosek et al., 2011).

In summary, this introduction has highlighted some of the most relevant information related to implicit cognition because it is the central theme of this thesis. Almost every intellectual question in social psychology, and some outside it, have been shaped in some way by the methods and theories related to implicit social cognition (Payne & Gawronski, 2010).

Overview of the thesis

Chapter 2 introduces the reader to parasite-stress theory (PST). This theory will be used to explain how disease rates account for differences in racial prejudice across the US as well as across the world. This chapter reports three studies to determine whether environmental occurrences of diseases and being visually primed with diseases can increase an individual’s

prejudice towards a racial out-group. The first two studies use Race IAT data and explicit self-reports made available through Project Implicit (secondary data), while the final study uses an experimental design to determine whether diseases are an important factor for increasing racial prejudice. These studies were limited in that the IAT can only report relative biases.

Chapter 3 presented the use of a recently developed implicit measure (IRAP) that aims to measure implicit biases towards separate categories (absolute biases) allowing the author to overcome the relative limitation of the IAT. However, further studies using the IRAP to measure racial biases after being primed with diseases had to be abandoned, because during testing a major flaw was discovered (a Positive Framing Bias: PFB) in the IRAP. This chapter explains why the PFB occurs and describes a comprehensive study of experimentally manipulating the PFB bias through framing techniques. The findings show that the IRAP overinflates positivity towards items and that experimenters could also inadvertently influence the estimates of implicit attitudes.

Chapter 4 introduces the Simple Implicit Procedure (SIP) which is a new measure designed to overcome the PFB limitation in the IRAP. This chapter presents three studies covering attitudes, stereotypes and self-concept with the aim of validating the SIP. A potential flaw in the SIP was detected and it occurs because participants were generally faster to press the affirming (“Yes”) than the negating (“No”) response key. A functional method to remove this affirming bias is developed and discussed.

Chapter 5 explains across three studies why an affirming bias occurs in the SIP. Also, this chapter identifies other response biases participants can have when completing implicit measures. Specifically, the impact of using nouns, verbs and adjectives on response times in RT tasks is addressed.

Chapter 6 presents a study that uses the SIP and explicit self-reports to measure racial biases towards black and white people pre-and post a disease or terror threat condition. Due to

sampling limitations, sufficient data were collected only from white females. Biases remained stable over time and were explained in terms of tending and befriending responses and sexual strategies.

Chapter 7 describes unanswered questions and further avenues of research that could be conducted using the SIP to address racial prejudice under disease threat.

**Chapter 2: Disease rates are associated
with racial prejudice across the US and
the world**

Abstract

What factors increase or decrease racial tensions? According to the contact hypothesis (Pettigrew & Tropp, 2006), increased exposure to out-groups reduces prejudice towards those groups. Supporting this, black US individuals who have more contact with white people are less prejudiced towards whites. However, inconsistent with the hypothesis, white individuals who have more contact with black people express increased anti-black prejudice (Rae, Newheiser, & Olson, 2015). Here we offer a new account to explain this pattern: we propose that residence in a region with relatively more exposure to infectious diseases leads to avoidance of out-groups because such avoidance may reduce the likelihood of contracting illness (Thornhill & Fincher, 2014). We note that disease rates are typically higher in regions with more black people (Eppig, Fincher, & Thornhill, 2011). It is therefore possible that the increased anti-black prejudice typically shown by white people in regions with large black populations reflects the operation of an adaptive Behavioural Immune System (BIS; Schaller & Park, 2011) which acts against the effects of contact. We show that, consistent with the parasite-stress hypothesis (Thornhill & Fincher, 2014), white individuals ($N > 702,000$) living in US states with higher disease rates display increased implicit (automatic) and explicit (conscious) anti-black/pro-white prejudice. Similarly, black individuals ($N > 149,000$) living in states with higher disease rates show increased implicit and explicit anti-white/pro-black prejudice. These results survive the inclusion of several individual and state-level controls. We also report an analysis of 76 countries; this produced results consistent with the parasite-stress hypothesis. Finally, we show that white participants exposed to disease-related primes show increased implicit and explicit anti-black/pro-white prejudice. Reduction in disease rates may therefore, inter alia, improve relations between groups and foster intergroup integration.

Introduction

Facebook's Mark Zuckerberg has recently pledged to give 99% of his wealth to charity (Goel & Wingfield, 2015). One of the primary uses of this money will be for curing diseases, which is a goal similar to that of the Bill and Melinda Gates Foundation (see <http://www.gatesfoundation.org/>). One major reason for these investments is that infectious diseases have been the leading cause of death worldwide (Jones et al., 2008) throughout much of history. Because of this immense impact on human survival, it is likely that humans have developed strategies/mechanisms to minimise pathogen threat throughout evolution. One such strategy is an immunological response to salient threats. The immune system triggers a physiological response that is activated when a pathogen is detected (reactive response). Other strategies, such as hypervigilance towards out-groups and/or the strengthening of in-group ties (Brown, Fincher, & Walasek, 2016; Diamond, 1999; Reicher, Templeton, Neville, Ferrari, & Drury, 2016) reflect the operation of the behavioural immune system (BIS; proactive response) which evolved to protect an individual from exposure to infectious diseases and potentially threatening stimuli (e.g., avoidance of decaying food or infected individuals, Murray & Schaller, 2016).

Parasite stress theory (PST)

The parasite-stress theory (PST; Thornhill & Fincher, 2014) hypothesis, which is strongly linked to BIS research, predicts that people will tend to avoid apparently infected individuals (Crandall & Moriarty, 1995; Kurzban & Leary, 2001). For example, comparably heightened avoidance behaviours and disgust responses were observed when participants were exposed to individuals with an infectious disease or superficial facial disfigurements compared with healthy controls (Ryan, Oaten, Stevenson, & Case, 2012). Although facial disfigurement can be an indicator of a genetic abnormality, normally no infections can be transmitted because of this abnormality. Therefore, this finding emphasises how humans are biased towards making

Type 1 errors such as incorrectly identifying and responding to non-existent threats (Error Management Theory; Haselton & Nettle, 2006) especially for infection diseases signals because it could increase their chances of survival.

PST also predicts that when diseases are made more salient (e.g., by priming with pictures) people will express increase prejudice towards groups that are strongly associated with diseases (Duncan & Schaller, 2009; Park, Faulkner, & Schaller, 2003; Park, Schaller, & Crandall, 2007). Such predictions have been confirmed. For example, participants in a disease priming condition showed increased prejudice (often on implicit measures such as the IAT) towards the elderly, people with physical disabilities and those who are obese compared to a control slideshow (for a review see Murray & Schaller, 2016).

In addition, PST predicts that people will avoid, and express more negative attitudes towards, dissimilar others, such as people with foreign accents (Reid et al., 2012) or from distant regions (Faulkner, Schaller, Park, & Duncan, 2004; Navarrete & Fessler, 2006). These responses are based on the fact that people who look different (i.e., skin colour), sound different (i.e., accent, language) or have different cultural values (i.e., religion) are more likely to be from a different region and hence might have been exposed to novel diseases. Therefore, foreigners are likely to have increased immunity to some diseases that local people can become infected by and local people can also infect foreigners with diseases they themselves are immune to.

PST is often used to explain why differences in overarching value systems (i.e., collectivism vs. individualism) develop and generally refers to environmental differences (e.g., rates of diseases) to explain cultural variations in beliefs (Thornhill & Fincher, 2014). Several theories have been proposed to explain differing value orientations across societies (e.g., market integration or religion; for a review see Hruschka & Henrich, 2013), but often these theories do not explain the origins of value orientation differences in the first place. Likewise,

why do some people develop conservative belief systems while others develop values that are more liberal? Researchers have shown that differences in personality, moral foundations, genetic factors, family upbringing and motivated cognition, such as a need for certainty, can lead to different ideological beliefs. Unfortunately, none of these research themes point towards why individual differences exist in the first place (e.g., origins of genetic or personality differences associated with ideological differences; Brown, Walasek, & Mullett, 2016). Describing the ultimate causal factors explaining why overarching belief systems develop in the first place is a core goal of PST.

PST hypothesises that people are more likely to develop collectivistic and socially conservative value systems when rates of parasites and infectious diseases are high. In contrast, when parasites and infectious diseases are low, people are more likely to adopt individualistic and socially liberal value systems. The reason for this difference is that there is both a cost and benefit to interacting with out-group members. Engaging with out-group members (openness) results in access to new ideas, mates and resources (benefits). However, these interactions could result in contraction of debilitating diseases (costs) and ultimately lead to death (Brown, Fincher, et al., 2016). Table 2.1 shows the major differences between collectivism and individualism that are relevant to PST.

Table 2.1: Differences between Collectivistic and Individualistic countries taken from Thornhill & Fincher, 2014.

Collectivistic	Individualistic
Developing countries	Developed countries
More conservative	More liberal
More infectious disease	Less infectious disease
More homicide	Less homicide
In-group goals paramount	Personal autonomy and self-fulfilment paramount
Low cognitive ability (IQ)	High cognitive ability (IQ)
More religious	Less religious
Autocratic governance	Democratic governance
Stronger family ties	Weaker family ties
Low rate of innovation	High rate of innovation
High conformity to tradition and norms	Low conformity to tradition and norms
Low socioeconomic status	High socioeconomic status
Restricted/conservative female sexuality	Unrestricted/liberated female sexuality
Greater distinctions between in and out-groups	Fewer distinctions between in- and out-groups
More out-group avoidance and racism (xenophobia)	Less out-group avoidance and racism
High disgust sensitivity	Low disgust sensitivity
Low openness to new experiences	High openness to new experiences
Closed-minded and unimaginative	Creative and curious
Perceptions of a threatening and dangerous world	Perceptions of a more secure world
Intolerance of ambiguity	Tolerance of ambiguity
High contagion concern	Low contagion concern

Using epidemiological data and worldwide cross-national surveys of individualism/collectivism, Fincher, Thornhill, Murray, and Schaller (2008) were the first to find support for PST. They found that greater regional prevalence of pathogens strongly correlated with cultural indicators of collectivism and negatively correlated with individualism. Further evidence has shown that increased pathogen prevalence predicts regional variation (both internationally and across U.S. states) of indicators of collectivism such as stronger family ties, increased ethnocentrism as well as greater religious belief (Fincher & Thornhill, 2012). A potential reason for religious beliefs increasing under high pathogen stress is that they ensure strict adherence to behavioural norms that would reduce disease transmissions (e.g., no premarital sex and not eating certain types of food). Premature death of loved ones, especially children, is more likely in high parasite stress areas and religion has the potential to reduce some of the torment this situation will induce.

Correlational evidence also shows that, across countries as well as U.S states, regions with higher disease prevalence are associated with lower intelligence (Eppig, Fincher, & Thornhill, 2010; Eppig et al., 2011). The potential cause of this correlation is that during development the body has to distribute energy in the most effective way to ensure survival; under high parasite stress, people are more likely to get exposed to an infection and consequently energy would have to be devoted to driving the immune system response, which would in turn reduce the energy available for other areas of development (i.e., cognitive development).

Individualism has often been linked to increased economic growth and Gross Domestic Product (GDP), and Thornhill and Fincher (2014) have shown that a decline in parasites is also strongly linked to these outcomes. Indeed, Murray (2014) has shown that a reduction in parasites predicted scientific and technological innovation, which greatly influences a

country's GDP and economic growth. Importantly, mediation analysis revealed that the link between low disease rates and innovation was mediated by reductions in conforming behaviours. In collectivistic cultures (high parasite), deviating from the norm is greatly disapproved of and this conformity is believed to reduce creativity and innovation. Because social norms (e.g., hygiene and food preparation techniques; Murray & Schaller, 2016) help to reduce the transmission of infections, conforming to social norm likely had evolutionary advantages (Murray & Schaller, 2012).

Both correlational and experimental methods have shown that increased exposure to diseases leads to increases in conformity (e.g., Murray & Schaller, 2012). Furthermore, greater parasite prevalence leads to autocratic governance (Murray, Schaller, & Suedfeld, 2013) and increases (social) conservatism (for a review see Terrizzi, Clay, & Shook, 2014). Higher disease rates have also been shown to be associated with higher number of languages as well as the amount of distinct religions in regions (Fincher & Thornhill, 2008a, 2008b) which is believed to be due to groups' unwillingness to engage with outsiders. Finally, parasites are believed to have an influence on personality and sexuality, resulting in more inhibitory behaviours (e.g., introversion, reduced openness and casual sex), as these decrease the potential for becoming infected with diseases (Schaller & Murray, 2008). Importantly, the correlational results reported above remain significant when controlling for potential confounding variables (e.g., human freedom, economic development, educational attainment, etc.).

Present study

We tested the major predictions of PST using US state-level data (Study 1), country-level data (Study 2) and data from a priming study (Study 3). Study 1 used secondary data from Project Implicit to determine whether people show a greater anti-out-group/pro-in-group bias in regions where disease rates are higher. The dataset included measures of individuals' explicit (conscious) and implicit (automatic associations) attitudes towards both in-groups and

out-groups. An advantage of examining implicit as well as explicit attitudes is that participants may behave in a socially desirable manner when reporting attitudes explicitly, which is especially problematic when socially sensitive topics are concerned (Greenwald et al., 2009). This study aimed to understand why being exposed to more black people increases white respondents' anti-black/pro-white biases (Putnam, 2007; Rae et al., 2015; Taylor, 1998). We first created an exposure-to-whites index for each state. This index was the log of the ratio of number of white people to number of black people in the state, multiplied by an integration index.

In Study 2, we examined whether higher disease rates are associated with higher implicit and explicit anti-out-group/pro-in-group biases across 76 countries. We analysed only data from white participants because data were available from few black respondents. Estimates of average skin colour were used as a proxy for exposure to black people. In Study 3, we used an experimental design to test the hypothesised causal link between diseases and prejudice.

Study 1 (US) and Study 2 (World)

On the basis of PST (Thornhill & Fincher, 2014), we hypothesised that people living in regions with higher disease rates will express greater anti-out-group/pro-in-group biases even after exposure to out-groups is controlled for. In addition, we tested the prediction of the contact hypothesis (Pettigrew & Tropp, 2006) that white individuals exposed to people in countries with darker skin tone will show decreased implicit and explicit prejudice towards black people when holding disease rates constant.

Method

Participants: The sample consisted of volunteers who completed the Race IAT on the Project Implicit website (<https://implicit.harvard.edu/implicit/>) (Nosek, et al., 2007). Analysis across the US was restricted to black and white participants within the 50 US states. The cross-

country analysis was limited to white participants ($N > 787,000$) with a further restriction that at least 30 usable IAT scores and explicit responses were available within each country. These restrictions resulted in 76 countries included in the final analysis. The sample used data collected between 2006 and 2013.

We used standard IAT analytic procedures to remove inappropriate IAT scores (Greenwald et al., 2003). The algorithm removed participants who made errors on $> 30\%$ of trials and had reaction times of $< 300\text{ms}$ and/or $> 10,000\text{ms}$ on $> 10\%$ of trials (approximately 2%). To facilitate reporting, we performed separate analysis on white ($N > 702,000$) and black respondents ($N > 149,000$) within the US (see Appendix 2 for a full description of demographics). The dataset we used is available for public use (<https://osf.io/y9hiq/>; Xu, Nosek, & Greenwald, 2014)

Materials

Implicit bias: All participants completed the Race IAT with “African American” and “European American” as the category labels and “Good” and “Bad” as the valence labels. These labels appeared at the top of the screen. The stimuli used included greyscale pictures of black and white individuals as well as positive (Glorious, Wonderful, Joy, Love, Peace, Pleasure, Laughter, Happy) and negative (Terrible, Evil, Horrible, Agony, Nasty, Awful, Failure, Hurt) words. These stimuli were presented successively to participants at the centre of their screen and on each trial participants were required to sort the stimulus into the appropriate category using the correct key press. If a correct response was given, the stimulus disappeared and a new stimulus appeared after 400ms. If an incorrect response was given, a red “X” appeared directly below the stimulus and both remained until the correct response was given.

In one of the two critical blocks, participants had to press the E key on a computer keyboard if a “good” word or a picture of a white person appeared and press the I key if a “bad” word or a picture of a black person appeared. In the other critical block, participants pressed E

if a “good” word or a picture of a black person was shown and pressed I key for a “bad” word or a picture of a white person. The order of the sorting task (i.e. black-good and white-bad first vs. black-bad and white-good first) was randomised across participants. The basic idea underlying the IAT is that participants will make faster and more accurate responses when those responses are congruent with their current beliefs than when they are not. Participants’ implicit biases were measured using IAT *D* scores (Greenwald et al., 2003). Positive *D* scores indicate an anti-black/pro-white attitude and negative *D* scores indicate an anti-white/pro-black attitude.

Bipolar Explicit bias: Participants used a 7 point Likert scale to respond to the question “Which statement best describes you?”: (1) I strongly prefer African Americans to European Americans – (4) I like European Americans and African Americans Equally – (7) I strongly prefer European Americans to African Americans.

Disease rates across US states and the world: Fincher & Thornhill, (2012) developed a measure of disease rates across the 50 US states. This measure aggregates all infectious diseases reported by the US Centers for Disease Control (CDC; available at www.cdc.gov) for the years 1993 to 2007 for each state, divides the number of diseases by state population, and transforms the result into a z-score. For the cross-country measure of disease rates, the World Health Organization’s (WHO), Infectious Disease DALY (Disability Adjusted Life Years) and non-zoonotic parasite prevalence (human-specific and multi-host disease transmission) within each country were standardised and then summed, producing a combined disease rates measure – see Fincher & Thornhill, (2012) for details and validation.

Control variables: For both the US state and cross-country analysis, five individual-level control variables were used. These included political ideology (1 = strongly liberal to 7 = strongly conservative), religious belief (1 = not at all religious to 4 = strongly religious), gender (dummy coded: 0 = female & 1 = male), age and education level (dummy coded: 0 = as far as

completion of high school, 1 = any educational accreditation after high school). For the US analysis, the state level controls included median income (logged), state inequality (Gini: higher scores = greater inequality), land population density per square mile, race exposure (high scores indicate greater white exposure, lower scores indicate greater black exposure) and whether a state was previously part of the Confederacy. Median income, inequality, population density and race exposure used the American Community Survey 5 year estimates (2008-2012). Based on previous methods of analysis (Alba, Rumbaut, & Marotz, 2005; Rae et al., 2015), we used the logged ratio of white people living in a state relative to black people. This ratio was then multiplied by $1 - (\text{state segregation}/100)$ to create the race exposure estimate. State segregation scores ranged from 0 (complete integration) to 100 (complete segregation) where the value indicates the percentage of black people that would need to move for them to be distributed exactly like white people (Frey & Myers, 2005).

The cross-country control variables included the same individual-level controls. The country level controls included gross national income (GNI) per capita, based on purchasing power parity (PPP) and logged, national level inequality (Gini), land population density per square kilometre, ethnic diversity, and skin colour. GNI, Gini and population density used the World Bank's most recent estimates up to the year 2012. Ethnic diversity measures the probability that two randomly drawn individuals from a country are not from the same ethnic group (see Alesina, Devleeschauwer, Easterly, Kurlat, & Wacziarg, 2003, for methodology). Skin colour within a country was estimated using Chaplin's (2004, updated 2007) skin colour map. The first author and an independent researcher assigned a score based on the maps scale of 1 (darker skin) to 7 (lighter skin) to each of the 86 countries that had 30 usable IAT and explicit scores. The interclass correlation coefficient between the two scores was .96. Discrepancies were resolved and agreed by the two raters. Of note, the full analysis included

only 76 countries because the control variables did not have a value for every country in the analysis.

Analysis: Multilevel analysis was used due to the large sample size available and the many benefits it has over multiple regression (Pollet, Tybur, Frankenhuys, & Rickard, 2014). Multilevel analysis groups individual responses, which provides a much finer analysis because individuals' variability in responding within a region is considered. Furthermore, demographic variables can act as individual level controls rather than having to use only state or country level controls. Participants were grouped by US state or country of residence depending on the analysis being conducted. The SPSS linear mixed model function was used and the model included a random intercept term at the US state level or the country level.

Results

Study 1 – Disease rates across US states

Consistent with PST, multi-level analysis (Table 2.2) revealed that white participants residing in states with higher disease rates showed a greater anti-black/pro-white bias for both implicit ($t = 3.54, p < .01$) and explicit attitudes ($t = 4.51, p < .001$). This finding survived controls for individual-level variables (age, gender, education, political ideology, religious belief) and state-level variables (median income, inequality, race exposure, population density, Confederate state). Also, again consistent with the PST hypothesis, black participants living in states with higher disease rates showed a greater anti-white/pro-black bias (Table 2.3).

Table 2.2: Summary of Multilevel Analysis for Variables Predicting US State Level Implicit/Explicit Scores.

Predictor	White Implicit Attitudes (N=735,119)			White Explicit Attitudes (N=702,815)		
	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>
Disease Rates	0.014	0.004	3.54**	0.054	0.012	4.51***
Political Ideology	0.023	0.000	74.90***	0.104	0.001	143.30***
Religious Belief	-0.012	0.000	-22.57***	-0.056	0.001	-43.85***
Gender	0.023	0.000	23.55***	0.177	0.002	77.67***
Age	0.000	0.000	4.55***	0.001	0.000	12.63***
Education	-0.005	0.001	-4.61***	0.076	0.003	30.46***
Median Income	-0.075	0.050	-1.51	-0.148	0.154	-0.96
State Inequality	0.004	0.190	0.02	0.039	0.591	0.07
Population Density	0.000	0.000	3.48***	0.000	0.000	0.95
Confederate State	-0.005	0.009	-0.51	-0.024	0.029	-0.82
Race Exposure	-0.020	0.013	-1.59	-0.094	0.039	-2.39*

Note: Scores were coded such that higher numbers indicate a greater anti-black/pro-white bias.

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Table 2.3: Summary of Multilevel Analysis for Variables Predicting US State Level Implicit/Explicit Scores.

Predictor	Black Implicit Attitudes (N=155,038)			Black Explicit Attitudes (N=149,551)		
	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>
Disease Rates	-0.017	0.004	-3.93***	-0.050	0.012	-4.27***
Political Ideology	-0.002	0.000	-2.43*	0.048	0.002	20.30***
Religious Belief	-0.012	0.001	-9.34***	-0.050	0.004	-12.73***
Gender	0.015	0.002	6.15***	0.173	0.007	23.81***
Age	-0.002	0.000	-20.57***	-0.012	0.000	-40.11***
Education	0.017	0.002	6.77***	-0.081	0.007	-10.98***
Median Income	0.112	0.058	1.92†	-0.047	0.159	-0.30
State Inequality	-0.178	0.206	-0.87	-0.439	0.561	-0.78
Population Density	-0.000	0.000	-0.04	0.000	0.000	1.00
Confederate State	-0.008	0.009	-0.96	-0.028	0.023	-1.21
Race Exposure	0.010	0.015	0.65	0.107	0.040	2.66*

Note: Scores were coded such that higher numbers indicate a greater anti-black/pro-white bias.

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

This held for both implicit ($t = -3.93, p < .001$) and explicit attitudes ($t = -4.27, p < .001$). In contrast to what would be predicted by the contact hypothesis, white participants living in states where exposure to blacks is higher showed higher explicit anti-black/pro-white biases ($t = -2.39, p = .022$). Yet consistent with the contact hypothesis, black participants living in states where exposure to whites is higher showed weaker anti-white/pro-black biases ($t = 2.66, p = .015$). In Appendix 2, we present alternative analyses using different methods to estimate out-group exposure which largely support parasite-stress theory. In Appendix 2, we also discuss effects of the individual-level controls.

Study 2 – Disease rates across 76 countries

Further, consistent with PST, multi-level analysis (Table 2.4) revealed that white participants residing in countries with higher disease rates showed a greater anti-black/pro-white bias on both implicit ($t = 2.12, p = .041$) and explicit attitude measures ($t = 3.33, p < .001$). This finding survived controls for individual-level variables (age, gender, education, political ideology, religious belief) and country-level variables (Gross National Income, inequality, population density, ethnic diversity, skin colour). Consistent with the contact hypothesis, people living in countries in which the average skin tone was darker showed reduced anti-black/pro-white explicit biases ($t = 3.48, p = .001$). For implicit attitudes this effect was marginally significant ($t = 1.82, p = .077$). Effects of the individual level controls are discussed in Appendix 2.

Table 2.4: Summary of Multilevel Analysis for Variables Predicting Country Level Implicit/Explicit Scores.

Predictor	White Implicit Attitudes (N=787,996)			White Explicit Attitudes (N=826,644)		
	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>
Disease Rates	0.011	0.005	2.12*	0.056	0.017	3.33**
Political Ideology	0.023	0.000	82.00***	0.106	0.000	156.20***
Religious Belief	-0.013	0.000	-25.18***	-0.054	0.001	-45.38***
Gender	0.256	0.000	28.03***	0.180	0.002	83.76***
Age	0.000	0.000	4.46***	0.001	0.000	14.44***
Education	-0.002	0.001	-2.25*	0.078	0.002	33.18***
GNI	0.043	0.022	1.93†	-0.006	0.070	-0.09
GINI	0.001	0.000	1.06	0.003	0.003	0.86
Population Density	0.000	0.000	0.15	0.000	0.000	0.07
Ethnic Diversity	0.006	0.030	-0.20	0.023	0.098	.024
Skin Colour	0.013	0.007	1.83†	0.079	0.023	3.48**

Note: Scores were coded such that higher numbers indicated a greater anti-black/pro-white bias

† $p < .10$, * $p < .05$. ** $p < .01$. *** $p < .001$.

Study 3

We hypothesised that white participants primed with disease images would show increased anti-black/pro-white bias compared to controls (participants primed with furniture and buildings). A terror threat priming condition was also included because previous research has shown that such priming can increase prejudice and conservative worldviews (Van de Vyver, Houston, Abrams, & Vasiljevic, 2016). This extra condition acted as a comparison to the disease condition and allowed us to test whether any threat to a person's life increased prejudice or if the effect (if present) is specific to disease threats. Rather than using a full 15 item scale to measure participants' perceived vulnerability to disease (PVD scale; Duncan, Schaller, & Park, 2009) we instead used a 5 item Private Body Conscious (PBC) scale. The (PBC) scale has previously been shown to moderate the effects of disgust (e.g., Petrescu & Parkinson, 2014; Schnall, Haidt, Clore, & Jordan, 2008).

Method

Participants were recruited through a social media platform dedicated to performing online surveys or experiments (see <https://www.reddit.com/r/SampleSize/>). All participation was voluntary, but each participant could view their implicit bias score at the end of the experiment which acted as an incentive. Data were gathered for slightly over a month (between 3rd March and 6th April, 2016). Overall, 525 participants completed the experiment. However, 74 were removed from the final analysis because "White" was not selected as their race. A further 57 were removed because either they did not respond correctly to the memory question (N=7), they had already previously completed the experiment (N=34), they selected "other" as their gender (N=12), their accuracy on the IAT was below 70% (N=2), or they responded faster than 300ms and/or slower than 10,000ms on >10% of the IAT trials (N=2).

The final sample included 394 participants (224 were male) and each one was randomly allocated to either the control (130 participants), disease threat (138 participants) or the terror

threat condition (126 participants). 231 participants from the US completed the experiment; the remaining participants were mainly from large Western countries such as Canada, Australia and the UK. The mean age of the sample was 24.3 years ($SD = 6.29$), and 338 participants had at least a college diploma. The sample was mainly non-religious/slightly religious ($M = 1.39$, $SD = .81$) and liberal ($M = 2.54$, $SD = 1.48$).

Materials

Demographic information: Participants' gender, age, race, country of residence, state of residence if in the US, educational level, political ideology (1= Strongly Liberal to 7= Strongly Conservative) and religious belief (1= not at all religious to 4= strongly religious) were collected via an online questionnaire.

Private Body Conscious (PBC) scale (Miller, Murphy, & Buss, 1981) is a 5-item measure addressing participants' awareness of internal physical sensations. Items included "I'm very aware of changes in my body temperature" and "I know immediately when my mouth or throat gets dry". Each item was rated on a 6-point Likert scale ranging from 1 (strongly disagree) to 6 (strongly agree). The internal reliability of this scale was .62.

Implicit & Explicit biases: The IAT and the bipolar explicit question used by Project Implicit and described above were matched as closely as possible. Two additional feeling thermometer items were included in the explicit attitude questionnaire. For these questions, participants had to rate how warm or cold they felt towards African Americans and European Americans (0 = coldest feelings, 5 = neutral, 10 = warmest feelings). A relative score was calculated by taking the black feeling thermometer score from the white feeling thermometer score. This measure produced similar findings to the bipolar explicit question and therefore, will not be discussed further. D-scores (Greenwald et al., 2003) were calculated from the raw data of the IAT by applying the D-algorithm. High scores on both the IAT and explicit measures indicate an anti-black bias/pro-white bias.

Disease, terror and control images: The disease images consisted of 30 images of mould, faeces and people with infections. 20 of the images were sourced from previous research that used pathogen primes (Schaller, Miller, Gervais, Yager, & Chen, 2010; Wu & Chang, 2012). 10 of these images had white individuals with chicken pox, cuts or who were coughing or sneezing etc. 10 images of black individuals with infections/diseases were added and closely matched the 10 images of white people⁵. For the control condition, 15 images of buildings and 15 images of single furniture items against a white background were used. For the terror threat condition, 30 images of terrorist attacks (e.g., 9/11, Madrid's ETA bombings) were used and were matched for the proportion of black and white individuals across the set. For the three conditions, the order of the images was the same for each participant, and this order was maintained for the two-time points when images were shown.

A one-way Analysis of Variance (ANOVA) on scores of how unpleasant or disturbing participants found the images revealed a significant difference across the 3 prime conditions (control: $M = 1.11$, $SD = .47$, terror: $M = 4.99$, $SD = 1.21$, disease: $M = 5.04$, $SD = 1.05$), $F(2, 389) = 705.69$, $p < .001$, $\eta_p^2 = .78$. LSD tests showed that both disease and terrorism images were reported as being significantly more disturbing/unpleasant than the control images, $t_s > 32.02$, $p_s < .001$, $d > 35.45$. No significant difference was found between the disease and terrorism conditions, $t(262) = .43$, $p = .71$. Therefore, both threat conditions induced similar aversive reactions, albeit, likely different emotional reactions (e.g., fear vs. disgust).

Procedure: The design included a between-subject variable called prime type which had three levels: control, disease threat and the terror threat. To begin the online experiment,

⁵ A pilot study was carried out that did not include the extra 10 images of black people. Due to this shortcoming, no significant disease priming effect was found at the implicit level.

However, this set of disease primes did significantly increase prejudice at the explicit level.

participants had to click a box which confirmed they were 18 years of age or older and were happy to participate in the experiment based on the information that they had been provided. Next, they completed demographic information and were randomly allocated to one of the three priming conditions. Participants scrolled through their respective images for as long as they wanted but a minimum of 30 seconds elapsed before participants could continue to the explicit questions.

Following these questions, participants viewed the same images previously shown for at least another 30 seconds and were asked if “The images were disturbing and unpleasant?” below all the images. The Race-IAT was then completed. Next, participants responded to a memory question to ensure they viewed the images, as well as a question asking if they had previously completed the experiment. An item asking how recently they have had a cold or flu was included, and they then completed the PBC scale. Finally, they were thanked and debriefed. The full experiment can be viewed at <https://brianpsychexperiments.warwick.ac.uk/tdc.html>

Results

There was a significant effect of prime type (control, terrorism, disease) on IAT D-score, $F(2, 391) = 5.43, p < .01, \eta_p^2 = .03$. Consistent with PST, Fisher’s LSD test showed significantly higher anti-black/pro-white bias with the disease prime ($M = .42, SD = .35$), compared to the control ($M = .29, SD = .41, t(266) = 2.90, p < .01, d = .34$). Similarly, the disease prime showed significantly higher anti-black/pro-white bias compared to the terrorism prime ($M = .29, SD = .35, t(254) = 2.76, p < .01, d = .37$). There was no significant difference between the control and terrorism prime conditions, $t(262) = .12, p > .25$ (see Figure 2.1).

There was also a significant difference between priming conditions for the explicit measure, $F(2, 379) = 3.57, p = .029, \eta_p^2 = .02$. LSD tests showed that there was a significantly stronger anti-black/pro-white bias with the disease ($M = 4.62, SD = .88$) than with the control

prime ($M = 4.37$, $SD = .64$, $t(245.83) = 2.62$, $p = .009$, $d = .34$). Significantly higher anti-black/pro-white biases were also shown with the terror prime ($M = 4.52$, $SD = .81$) compared to the control prime, $t(255.81) = 2.17$, $p = .031$, $d = .29$). The disease and terror threat conditions did not differ, $t(259) = .19$, $p > .25$ (see Figure 2.1). For correlational analysis of the variables used in Study 3 and analysis addressing gender differences, please see Appendix 2.

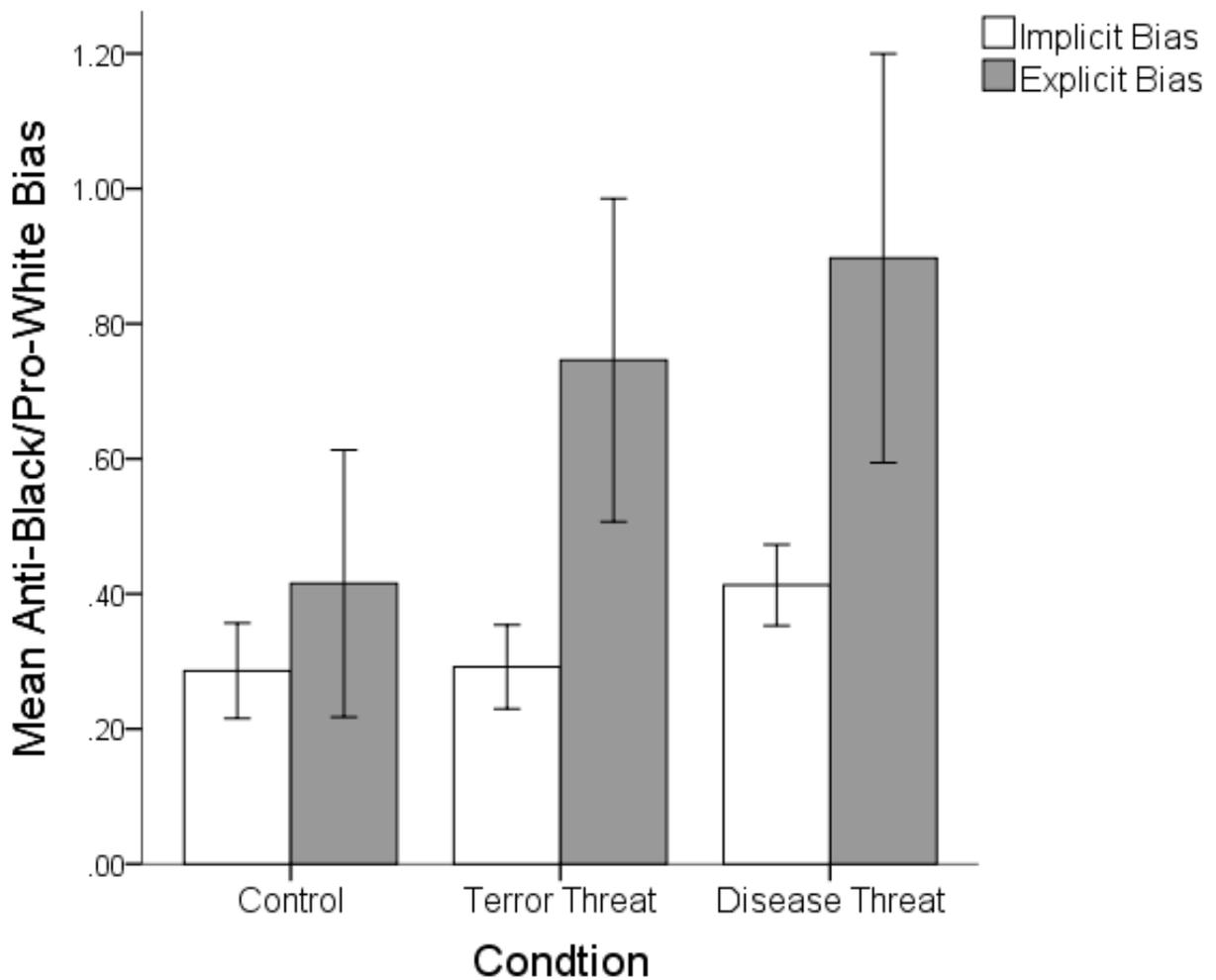


Figure 2.1: Mean implicit and explicit anti-black/pro-white bias across the three conditions. Error bars show 95% confidence intervals. In the disease threat condition, implicit bias (prejudice) is higher than in the terror threat and the control conditions. Explicit bias is higher in both the disease and terror threat conditions compared to the control.

Discussion

Across three studies, our findings are consistent with the hypothesis that living in regions with higher disease rates and being primed with diseases increases anti-out-group/pro-in-group biases. Across the US states, we observed that residents in states with higher disease rates displayed stronger anti-out-group/pro-in-group biases at the implicit and explicit level for both white and black respondents. In many respects, these findings were replicated in our cross-country analysis. Importantly, all the effects reported across the US and the world were robust when controlling for several important individual and state level factors often used to explain prejudice. For example, conflict over limited resources (Baumeister & Bushman, 2010) and greater diversity (Putnam, 2007) have previously been used to explain racial prejudice in the US. However, when we considered median income, inequality (proxy for limited resources) and race exposure (proxy for diversity), disease rates were the best predictor of prejudice.

At the cross-country level when disease rates were held constant, we found that white respondents living in countries that had an average darker skin tone exhibited a reduction in explicit anti-black/pro-white biases. This result also survived the inclusion of numerous control factors. At the implicit level, comparable results (albeit marginally significant results) were observed across the world after the control factors were added to the model. These findings are consistent with our prediction, based on the contact hypothesis, that contact with out-groups is a crucial factor in reducing prejudice particularly at the explicit level.

Across the US states, race exposure was not associated with increased prejudice among black and white respondents at the implicit level. However, on measures of explicit attitudes, even when disease rates were controlled, white respondents with higher exposure to black people were associated with greater anti-black/pro-white biases. In contrast, black respondents exposed to more white people were associated with explicitly weaker pro-black/anti-white biases. These inconsistent findings at the explicit level for both groups, as well as the cross-

country analysis, challenge the notion that exposure to more black people increases prejudice in white individuals. Therefore, if it is the case that exposure to black people increases prejudice, the effect appears to be unique to white residents of the US.

The central aim of this research was to test the novel hypothesis that exposure to disease increases racial prejudice. As hypothesised, we found experimentally that being reminded of diseases/infections increases anti-black/pro-white biases among white individuals. This selective effect was clearly observed at the implicit level because the terror threat prime did not increase prejudice towards black people. From an evolutionary perspective, this outcome is exactly what we would expect given that, unlike with disease cues, there will have been little opportunity to develop detection and defence mechanisms to terrorist threats (if it is even possible to do so).

Regarding the secondary and experimental data, we observed small to moderate effect sizes. However, it has been reported that even small increases in implicit prejudice can have substantial societal effects on groups experiencing persecution (Greenwald et al., 2015). A limitation of the data gathered through Project Implicit is that a high proportion of participants completed the Race IAT for course credits or assessments for school. Therefore, the sample is biased towards students that are liberal and younger than the general population (Nosek et al., 2007). Crucially however, this limitation is unlikely to affect the substantive findings that disease rates increase prejudice at the US state and country level. Furthermore, the randomised experiment in Study 3 overcame this shortcoming. Another limitation is that the relative nature of the implicit and explicit measures used means that we cannot determine if diseases are associated with a strong preference for in-groups or a greater disdain towards out-groups or some combination of these two biases (Greenwald et al., 2003, see Chapter 6 for the description and test of a new potential measure of absolute implicit associations - Simple Implicit Procedure; SIP - that overcomes the relative nature weakness of the IAT.)

Our study underlines the importance of parasite-stress theory and BIS research by demonstrating that the prevalence of infectious diseases is an influential factor in explaining intergroup tensions across the US and the world. Indeed, none of the other US state level factors were consistently associated with racial prejudice. Likewise, across the world, only the country-level factors of disease rates and reduced exposure to people with dark skin tone were consistently related to increased prejudice for white respondents. Therefore, only when disease rates (or the perception of disease rates) are low will increased contact with out-groups be likely to result in a reduction in prejudice. Our research also offers the possibility that disease outbreaks (e.g., recent Ebola and SARS outbreaks) might be an important contributor to heightened prejudice towards ethnic out-groups. Similarly, refugees and illegal immigrants often originate from regions with high disease rates which could be a key factor in explaining the race-motivated attacks or social segregation that these groups sometimes experience. In conclusion, efforts to ameliorate diseases worldwide have the potential to reduce interaction anxiety with out-groups, which could encourage more meaningful engagement between groups.

**Chapter 3: Measuring implicit attitudes:
A positive framing bias flaw in the
Implicit Relational Assessment
Procedure (IRAP)**

Abstract

How can implicit attitudes best be measured? The Implicit Relational Assessment Procedure (IRAP), unlike the Implicit Association Test (IAT), claims to measure absolute, not just relative, implicit attitudes. In the IRAP, participants make congruent (Fat Person-Active: *False*; Fat Person-Unhealthy: *True*) or incongruent (Fat Person-Active: *True*; Fat Person-Unhealthy: *False*) responses in different blocks of trials. IRAP experiments have reported positive or neutral implicit attitudes (e.g., neutral attitudes towards fat people) in cases where negative attitudes are normally found on explicit or other implicit measures. It was hypothesised that these results might reflect a Positive Framing Bias (PFB) that occurs when participants complete the IRAP. Implicit attitudes towards categories with varying prior associations (non-words, social systems, flowers and insects, thin and fat people) were measured. Three conditions (standard, positive framing, and negative framing) were used to measure whether framing influenced estimates of implicit attitudes. It was found that IRAP scores were influenced by how the task was framed to the participants, that the framing effect was modulated by the strength of prior stimulus associations and that a default PFB led to an overestimation of positive implicit attitudes when measured by the IRAP. Overall, the findings question the validity of the IRAP as a tool for the measurement of absolute implicit attitudes. A new tool (Simple Implicit Procedure: SIP) for measuring absolute, not just relative, implicit attitudes is proposed.

Introduction

Implicit attitudes are automatic evaluations that occur outside conscious awareness and are measured without requiring respondents to introspect on their feelings. Explicit attitudes in contrast are the result of deliberate introspection and controlled evaluative judgement (Greenwald & Banaji, 1995). One reason for measuring implicit attitudes is that participants may use self-presentation tactics or respond in a socially desirable manner on explicit self-reports to avoid being perceived as prejudiced. Implicit measures can also be useful in areas where participants might be unwilling to reveal personal psychological attributes or are unaware of these psychological attributes (for a review of implicit attitudes and the tools used to measure them see Gawronski & De Houwer, 2014).

The current gold standard method for assessing implicit attitudes is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) and this tool has been increasingly used in clinically relevant areas (see <https://implicit.harvard.edu/implicit/pimh/>). The IAT has shown promise in predicting self-harm (Randall, Rowe, Dong, Nock, & Colman, 2013), social anxiety disorders (Teachman & Allen, 2007) and suicidal ideation (Harrison, Stritzke, Fay, Ellison, & Hudaib, 2014). The current study questions the validity of a recently developed implicit measure, the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010), which has also been used with vulnerable populations.

Absolute vs. relative measures of implicit cognition

In a typical IAT (e.g., the Fat-Thin IAT) participants are successively presented with various pictures of thin and fat individuals and positive and negative words. In one of the two critical blocks, participants have to press the E key on a computer keyboard if a positive word or a picture of a thin person appears and press the I key if a negative word or a picture of a fat person appears (congruent task). In the other critical block, participants have to press E if a

positive word or a picture of a fat person is shown and they must press the I key for a negative word or a picture of a thin person (incongruent task). The basic idea underlying the IAT is that participants will make faster and more accurate responses when those responses are congruent with their current beliefs than when they are not.

Researchers using the IAT typically find a pro-thin/anti-fat bias (e.g., O'Brien, Hunter, & Banks, 2007). Importantly, the IAT can only measure relative attitudes (e.g., attitudes to fat people relative to attitudes to thin people). It is therefore impossible to determine using the IAT whether this bias is the result of a strong/weak pro-thin bias, a strong/weak anti-fat bias or some combination of the two (see Blanton & Jaccard, 2006; Roddy, Stewart, & Barnes-Holmes, 2011; 2012). Likewise, the IAT cannot be used to determine how interventions that aim to increase or reduce implicit biases have their effect (e.g., a difference in implicit attitudes could be reduced by acting on the 'Thin Person' category, the 'Fat Person' category, or both; see Lai et al., in press).

The IRAP is a recently developed alternative to the IAT that claims to measure attitudes in both absolute and relative terms. Like the IAT, the IRAP is based on latencies of participants' accurate and speeded responses to stimuli. However, rather than categorizing items with the appropriate key press (as in the IAT), participants instead have to press keys that correspond with *True* or *False* displayed on the computer screen. "Correct" responses are based on instructions given to participants before each block of trials begins. For example, participants might be presented with the picture of a 'Fat Person' and the adjective 'Active'. In one condition participants are instructed to respond with a *True* key press (an incongruent response) and in another condition they are instructed to respond with a *False* key press (a congruent response).

The mean latency difference between pressing *True* and pressing *False* across the congruent and incongruent response conditions is claimed to reflect a participant's absolute

positive implicit attitude towards ‘Fat Person’. Similarly, an evaluation of absolute negative implicit attitudes can be obtained by measuring the latency difference to respond *True* to the association of ‘Fat Person’ with negative words (e.g., ‘Fat Person’-‘Unhealthy’), except that the *True* key is now a ‘congruent response’ and the *False* key an ‘incongruent response’ (see Figure 3.1)⁶. When the results from responding *True* or *False* to ‘Fat Person’ with positive words and ‘Fat Person’ with negative words are combined/averaged, the result reveals if the absolute attitude towards ‘Fat Person’ is positive, negative or neutral⁷.

This kind of absolute attitude measurement cannot be achieved using the standard IAT, because the latency with which a person responds to positive words (e.g., ‘Active’) and pictures of fat people in the ‘Fat Person-Positive’ condition is uninterpretable in isolation. The latency only becomes useful if it can be compared with the latency from the condition in which negative words (e.g., ‘Unhealthy’) and pictures of thin people require the same response (i.e., the ‘Thin Person-Negative’ condition).

⁶ Separately measuring both a positive and a negative attitude towards a category can provide a more nuanced understanding of the attitude under investigation. For example, a person can have both a strong positive attitude towards giving blood (i.e., like helping people) but also have a strong negative attitude towards giving blood (i.e., fear of needles). Traditional bipolar attitude measures (e.g., like-dislike) treat positive and negative evaluative processes as reciprocally activated, equivalent and interchangeable which is not always the case (see Cacioppo, Gardner, & Berntson, 1997).

⁷ The IRAP normally includes a comparison group to allow researchers to carry out relative comparisons between groups in order to compare results with those obtained from the IAT.

of particular importance because a number of IRAP publications have reported unexpected or contradictory results that cannot be accounted for easily.

For example, Roddy et al. (2011, 2012; see also Nolan, Murphy, & Barnes-Holmes, 2013) reported that what was driving the pro-thin/anti-fat bias in the IAT was an extremely strong pro-thin bias with participants having a neutral implicit attitude towards fat people. The concept of culture-based socialization that glorifies a slender figure was used to account for these results. However, Bessenoff & Sherman (2000) using a non-relative/absolute lexical decision making task, obtained completely opposite results to those of Roddy et al. (2011, 2012; i.e., they found an anti-fat bias and a neutral/unbiased attitude towards thin people).

In another example, Barnes-Holmes et al., (2010) found that participants associated pictures of a black person with a gun and pictures of a white person with a gun as ‘Safe’ (i.e., they were faster to press *True* rather than *False* to the joint presentation of a picture of a person with a gun and ‘Safe’). This was not what the authors had predicted based on prior evidence that a civilian holding a gun in a neutral context would normally be considered as dangerous rather than safe (i.e., participants should press *False* quicker than *True*)⁸. Other counterintuitive results include findings that a normative sample of participants had an unexpected pro-death implicit bias (Hussey, Daly, & Barnes-Holmes, 2015), and one where religious participants and atheists had unpredicted positive attitudes towards both their in-group and out-group

⁸ It could be argued that people are generally exposed to guns in a safe environment (i.e., guns being carried by law enforcers), resulting in the pro-safe attitude. However, we think this is unlikely because the experiment used pictures of civilians wearing a white t-shirt and this is likely to evoke a fight or flight response especially for the black individuals (e.g., Correll, Park, Judd, & Wittenbrink, 2002).

(O'Shea, 2017a)⁹. Importantly, when the data were compared relatively (like in the IAT) the typical pro-in-group/anti-out-group biases were obtained. In the current study we show that these results can be accounted for on the assumption that participants are adopting a simple default heuristic (a Positive Framing Bias; PFB) when performing the IRAP.

Framing effects and language biases

The power of positive and negative framing has been well documented, particularly in the area of risky decision making (for reviews, see Kühberger, 1998). One of the best-known examples of the framing effect is Tversky and Kahneman's (1981) "Asian disease problem". When presented with choices, if logically equivalent outcomes are phrased in terms of the *number of lives saved*, people will generally select the certain/safe option. In contrast, if outcomes are phrased in terms of the *number of lives lost*, people preferentially select the risky option. Building on this evidence McKenzie and Nelson (2003) found that participants had a bias towards describing a new treatment in terms of the percentage that survived rather than died (i.e., a positively biased outcome description). Furthermore, participants were also biased towards describing a glass as half full rather than half empty (i.e., an optimism bias; e.g., Peterson, 2000). Importantly, it was possible to manipulate these positivity biases by describing the reference points of the treatment outcomes differently or stating where the liquid in the glass had been previously. For example, people were more likely to say the glass was half empty if it had previously been full.

Language both reflects and shapes our representations of the world (Boroditsky, 2011) and language can therefore provide important insights into thought processes and biases. English speakers appear to have a general tendency to describe changes in events or objects as

⁹ Atheists are generally distrusted in societies with a religious majority (e.g., Gervais, Shariff, & Norenzayan, 2011).

increasing rather than decreasing. To clarify, if a person's weight has decreased, we might describe this as 'the person is thinner' (thinness has increased), not 'the person is less fat'. In contrast, a person who has gained weight might be referred to as 'fatter', but not as 'less thin'. Additionally, there is no morpheme in English that is equivalent to the suffix 'er' to indicate that a dimension has decreased (McKenzie & Nelson, 2003). This implies that describing objects in positive (increasing) rather than negative (decreasing) ways is a bias inherent in the English language and therefore our thought processes.

Another strand of psycholinguistic research that emphasizes the importance of positivity is based on markedness (see Haspelmath, 2006, for a critical review) where 'unmarked' is the classification for words used most frequently and which also have the most neutral meaning (Leech, 2006). 'Positive adjectives like *good* and *long* are stored in memory in less complex forms than...their opposites' (Clark, 1969, p.398) and therefore, are described as 'unmarked'. A speaker asking in a restaurant 'How good is the food?' is simply asking for an evaluation of the food and will be satisfied with a positive or negative appraisal. However, when asking 'How bad is the food?' essentially he/she is inquiring about the extent of the food's badness. In this context, "since *good* can be neutralized and *bad* cannot, *good* is said to be 'unmarked' and *bad* 'marked'" (Clark, 1969, p.398). Likewise describing a board as six meters long is acceptable to English speakers but describing it as six meters short is not. Therefore, short would be described as 'marked'. Finally, a morphologically negative word is 'marked' as opposed to a positive one (e.g., honest vs. dishonest; happy vs. unhappy) and therefore, humans are less efficient (in terms of processing speed) at handling negative statements (e.g. Sherman, 1973).

Related studies (e.g., Matthews & Dylman, 2014) have shown that English speakers have a preference to use 'larger' (e.g., more, taller, higher) comparisons to describe the relationship between two magnitudes (e.g., one flag is _____ than the other). Other evidence

shows that people make a concerted effort to dampen down, mute, and even erase negative experiences and that positive illusions promote psychological wellbeing (Taylor & Brown, 1994). Peoples' strong positivity biases have been described as the Pollyanna Principle (for review see Matlin, 2016). Overall, the evidence points towards humans having a bias towards positivity and larger or increasing descriptions (i.e., a Positive Framing Bias; PFB) and suggests that this bias can be manipulated by framing effects.

Present study

The present study tests the hypothesis that the absolute estimates of implicit attitudes obtained from the IRAP are influenced by a PFB. Specifically, people might find it easier to respond *True* to positive descriptions of stimuli than to press *False*. This effect is predicted to occur over and above any effect of congruence between the stimulus and the description that is presented with it, and is expected to lead to an overestimation of the positivity of absolute implicit attitudes. For example, participants may be faster to respond *True* (rather than *False*) when responding whether a stimulus category (e.g., Thin Person/Flowers) is positive (e.g., "Good") than when it is negative (e.g., "Bad"). If a PFB does influence IRAP responding, the IRAP may not be able to measure absolute implicit attitudes in the way intended.

Three main hypotheses were therefore tested:

H1: By default, participants will be more likely to focus on positive rather than negatively framed associations and therefore they will be faster to respond *True* than *False* for categories presented with positive words in the standard IRAP condition.

H2: The way the task is framed will influence estimates of a person's absolute implicit attitudes. If absolute attitudes as measured by the IRAP are susceptible to PFB effects, directly manipulating how the task is framed to participants should influence the results in predictable directions. Specifically, encouraging participants to focus on whether positive associations are true or false should increase estimated 'implicit positive attitudes' towards any category.

Conversely, encouraging participants to focus on whether negative associations are true or false should increase estimates of ‘implicit negative attitudes’. The effect of framing is predicted to be modulated by the strength of pre-existing negative or positive associations. That is, robust positive or negative associations will be less likely to be influenced by framing effects. Tasks that involve weak or absent prior associations will be more strongly influenced by the framing manipulation than those with strong prior associations. To test this we ran four different IRAP tasks, using stimuli which had differing strengths of prior positive and negative associations.

H3: When the IRAP data are analysed in such a way as to obtain relative attitudes (as is done for the IAT), the framing effect will have no influence. This is because any systematic biases will be cancelled out when the relevant conditions are combined. Tasks that use stimuli with weak or absent prior associations will show neutral attitudes; those with strong prior associations will in contrast show the expected pro/anti-bias (e.g., pro-thin/anti-fat).

To test these hypotheses, we manipulated the way in which tasks were framed (no frame-standard, positive frame, and negative framing conditions) for each of four IRAP tasks, which contained stimuli that varied in their strength of prior associations. If a PFB does exist (i.e., if the first two hypotheses were confirmed), the use of the IRAP as a measure of absolute, rather than just relative attitudes would be severely limited.

Method

Participants: The final sample consisted of 60 students from the University of Warwick (mean age = 21.9, $SD = 2.79$), 20 in each condition (standard, positive frame and negative frame). Fourteen participants were replaced (four from the standard, six from the positive frame and four from the negative frame conditions)¹⁰ from the original 60 because the required

¹⁰ Including the original participants did not cause any significant changes to the results reported.

performance criteria were not met. The final sample contained 32 females and comprised 36 Asians, 19 Whites, 3 Mixed Race and 2 Blacks. Participants were recruited via an electronic recruitment system and were paid £4. Participants were tested individually in a small room and the experiment took approximately 40 minutes to complete.

Apparatus and materials

Implicit Relational Assessment Procedure: (IRAP; Barnes-Holmes et al., 2010). Each participant completed four separate IRAPs. These were a Non-word IRAP, a Social System IRAP, a Nature IRAP, and a Weight IRAP. These four different IRAPs were chosen because the category stimuli in each were expected to have varying degrees of prior associations. Both the Nature and Weight IRAP stimuli were predicted to have strong prior associations (see Greenwald et al., 1998 and O'Brien, et al., 2006). The Social System IRAP stimuli were expected to have weak associations¹¹ and the Non-word IRAP stimuli should have no prior associations. All IRAP tests were administered using an Intel Windows 7 laptop with a 15" LCD screen (IRAP software available from <http://irapresearch.org/wp/downloads-and-training/>).

For the Weight IRAP, on each trial one of two category labels was presented at the top of the screen ('Thin Person' or 'Fat Person') and a single positive or negative target stimulus was presented in the center of the screen (e.g., 'Healthy' or 'Ugly'). The Weight IRAP used pictures, and the other three IRAPs used words as the category labels. Two response options ('True' and 'False') also appeared at the bottom left and right (respectively) of the screen (see

¹¹ O'Shea, (2017b) found that 17 UK and 10 Chinese undergraduates had no relative comparison bias in favor or against Capitalism vs. Communism or Socialism. However, see O'Shea (2015) where bankers had extremely strong biases in favor of Capitalism.

Figure 3.1). The remaining IRAP tasks followed the same format but with the condition-specific stimuli presented (see Tables S3.1 and S3.2 in Appendix 3 for the full stimulus sets).

All category labels and target stimuli were matched on word length, phoneme length and word frequency using Brysbaert and New (2009) film subtitle database. The positive and negative target stimuli for the Social System IRAP were determined by asking 50 participants (via Amazon Mechanical Turk) to generate positive and negative words that they associated strongly with Capitalism and Socialism. The most frequently reported words were selected for use in the current study. Stimuli for the nature IRAP were chosen from past implicit research (e.g., Greenwald et al., 1998). The 12 images for the weight IRAP were taken from Nolan et al. (2013) and consisted of three pictures of women and three pictures of men, before and after they had lost a significant amount of weight. These images were controlled on a number of dimensions (e.g., picture angle, cropping, clothing and background). The most appropriate positive and negative target words were chosen from studies of implicit weight biases (Roddy et al., 2010, 2011).

Design and procedure

The study used a mixed 4 (trial type: category 1-positive words, category 1-negative words, category 2-positive words, and category 2-negative words) \times 4 (IRAP task: Non-word, Social System, Nature and Weight) \times 3 (framing condition: standard, positive, negative) design, with IRAP task and IRAP trial type as within-subjects factors and framing condition as a between-subjects factor.

We illustrate the procedure using the Weight IRAP task. Participants were required to respond in a predefined way as specified before each block of trials. For example, participants might be instructed ‘On this block please respond as if Thin Person is Positive and Fat Person

is Negative'. This would be described as a pro-Thin block of trials (congruent responding¹²) and participants were required to respond *True* to the stimulus combination 'Thin Person – Positive Word' and 'Fat Person – Negative Word', and to respond *False* to 'Thin Person – Negative Word' and 'Fat Person – Positive Word'. After participants completed this pro-Thin block they were required to respond in the opposite manner on the next block and were thus instructed: 'On this block please respond as if Thin Person is Negative and Fat Person is Positive' (incongruent responding). On these pro-Fat trials participants had to respond *False* to 'Thin Person – Positive Word' and 'Fat Person – Negative Word' and to respond *True* to 'Thin Person – Negative Word' and 'Fat Person – Positive Word'.

This procedure enabled the experimenter to measure attitudes based on four separate trial types (Thin Person – Positive Word, Thin Person – Negative Word, Fat Person – Positive Word and Fat Person – Negative Word) by measuring the mean latency difference in responding *True* vs. *False* in pro-Thin and pro-Fat blocks. To respond, participants pressed either 'D' or 'K' on the keyboard. The D and K keys corresponded to the left – right positions of the response options on the screen respectively. The locations of the two response options interchanged quasi-randomly from left to right among trials, with the constraint that they could not remain in the same position three times in succession.

To complete the IRAP sequence, participants completed six test blocks which alternated between requiring pro-Thin or pro-Fat responses across blocks. Initial block (pro-Thin or pro-Fat) was counterbalanced across participants. Each block consisted of 24 trials,

¹² For the Non-Word IRAP and the Social System IRAP the blocks that are referred to as congruent or incongruent responding were arbitrarily selected. However, for the Nature and Weight IRAP congruent responding was defined in terms of the blocks people were expected to find easier due to prior associations in memory.

made up of the 6 positive and 6 negative target words presented twice in the presence of the two category stimuli (i.e., ‘Thin Person’ or ‘Fat Person’). Trials were presented quasi-randomly, such that the same trial type could not be repeated across two successive trials. If participants pressed the correct response on a trial the screen cleared for 400ms before the next trial was presented. If an incorrect response was given a red letter (X) appeared on screen and remained until the participant pressed the correct response key. After participants completed a block of trials their mean accuracy and median response latency scores were displayed on the screen. The first participant was assigned to the standard condition, the second to the positive framing condition and the third to the negative framing condition. This order was then repeated and maintained for the remaining participants.

Standard framing condition: For the standard condition visual instructions before each block were alternated across participants. For example, one participant would be shown, ‘On this block please respond as if **Thin Person** is Positive and **Fat Person** is Negative’ and ‘On this block please respond as if **Thin Person** is Negative and **Fat Person** is Positive’ and the next participant would always begin with **Fat Person** as the first category described in each sentence.

Positive framing condition: In the positive framing condition visual explanations of how to perform the task were always framed in a positive way prior to beginning each block of trials (e.g., ‘On this block please respond as if **Thin Person** is Positive and **Fat Person** is Negative’ or ‘On this block please respond as if **Fat Person** is Positive and **Thin Person** is Negative’)¹³. These instructions were dependent on participants completing a pro-Thin or pro-

¹³ Since English speakers process sentences from left to right the leftmost instructions were likely be processed first.

Fat block. Unlike in the standard framing condition, each participant in this condition was also presented with a standardized verbal explanation from the experimenter.

The script was as follows:

“A method or strategy that will help you complete this task is to keep ‘**Thin Person**’ and ‘Positive word’ in your mind and base all other responses off that¹⁴. For example, when a ‘**Thin Person**’ and a Positive word appears, press *True* and if this does not occur press *False* (such as ‘**Thin Person**’ and Negative word). Then use this strategy to gauge how to respond to the other category by responding in the opposite manner (such as ‘**Fat Person**’ Positive words *False*; ‘**Fat Person**’ Negative word *True*) To emphasize ‘**Thin Person**’ is Positive, ‘**Thin Person**’ is Positive”

After participants completed the first practice block they were told the following:

“Now on this block put ‘**Fat Person**’ is Positive into your mind ‘**Fat Person**’ is Positive, ‘**Fat Person**’ is Positive”

Participants were told the following on each successive IRAP:

“For this task use the same strategy as the last time by keeping ‘**XXX**’ and Positive word in your mind.”

Negative framing condition: The negative framing condition was identical to the positive except that the underlined words in the verbal script above were replaced with their antonyms and the visual description before each block always described the negative instructions first (e.g., ‘On this block please respond as if **Fat Person** is Negative and **Thin**

¹⁴ Category words in bold were replaced depending on which pro-block was presented first and which of the four versions of the IRAPs was being conducted. Sentences in brackets were only explained to participants who were confused and required more detailed instructions.

Person is Positive’ or ‘On this block please respond as if **Thin Person** is Negative and **Fat Person** is Positive’).

The remaining three IRAP tasks (Non-word, Social System and Nature) followed the same procedure but with the condition-relevant stimuli substituted. The order of the four IRAPs was randomized within the standard framing condition with the same set of random orders used for the positive and negative framing conditions. For the first IRAP participants were required to complete a minimum of two practice blocks. This was to ensure that participants were accustomed to the IRAP’s procedure. If a participant received first pro-Fat then pro-Thin conditions on the practice blocks this sequence was maintained for the 1st and 2nd test block, the 3rd and 4th, and the 5th and 6th test block.

To proceed to the test blocks participants had to achieve an accuracy of 79% or above and a median response latency of less than 2,200ms on two consecutive practice blocks. All participants met the practice block criteria but if these criteria were not maintained throughout each test block on the four IRAP tasks, participants’ data were removed and replaced with data from new participants. This resulted in 14 participants being replaced. When the four IRAPs were completed, participants provided demographic information, received their payment and were thanked and debriefed.

Results

The primary data obtained from the IRAP tasks are raw latency scores defined as the time in milliseconds that elapsed between the onset of the stimulus and the correct response being made by the participant. The DV was participants’ mean *False* minus *True* reaction time difference for each of the four trial types (i.e. Thin Person-Positive Words, Thin Person-Negative Words, Fat Person-Positive Words, Fat Person-Negative Words) across the congruent (e.g., pro-thin) vs. incongruent (e.g., pro-fat) blocks. Analysing the trial types/categories separately provided the absolute results while averaging all the trial types provided the relative

results. Following standard procedures to control for individual variation, (Barnes-Holmes et al., 2010a) each participant's response latencies were transformed using an adaption of the Greenwald et al., (2003) D-algorithm. The steps involved in calculating both the absolute and relative D-IRAP scores can be found in Appendix 3. Initially a stimulus Category \times Word Valence \times IRAP task within-subjects ANOVA was performed on the *False* minus *True* absolute D-IRAP scores for each of the four trial types in the standard condition. This analysis allowed us to test the first hypothesis: H1: participants will have a faster *True* (vs. *False*) response for categories presented with positive words. Inclusion of the factors stimulus Category and IRAP task also allowed us to consider the generality of the effect of Word Valence.

Following this, a 4 (trial type: category 1-positive words, category 1-negative words, category 2-positive words, category 2-negative words) \times 4 (IRAP task: Non-word, Social System, Nature, Weight) \times 3 (framing: standard, positive, negative) mixed ANOVA was performed on the of absolute D-IRAP (estimate) scores¹⁵. This analysis was carried out to test H2 (how the task is framed will influence the estimates of a person's implicit attitudes). A further analysis combined both the positive word and negative word D-IRAP (estimate) scores for each of the eight categories (i.e., Fat Person, Thin Person, Flower, Insect etc.) in the standard condition. This method provides the average absolute D-IRAP (estimate) scores for each category and matches the standard way of calculating the IRAP's results. This analysis

¹⁵ The absolute D-IRAP (estimate) scores were found by reversing the *False* minus *True* D-IRAP scores for categories presented with negative words. Categories presented with positive words were not reversed. This was carried out so that scores above zero represent a positive attitude and those that are below the zero mark represent a negative attitude. The absolute D-IRAP (estimate) scores were used to test H2.

was used to test further the argument that the PFB has an influence on peoples' responses and to examine the extent to which this can lead to potentially inflated estimates of implicit attitudes¹⁶. Finally, tests were carried out to determine whether framing influenced participants' *relative* implicit attitudes (H3). In all analyses, whenever Mauchly's sphericity assumption was violated the Greenhouse-Geisser correction was applied.¹⁷

Test of H1: The influence of the PFB on prior associations in the standard condition

False minus *True* D-IRAP scores for the standard framing condition were analysed using a 2 (Category) \times 2 (Word Valence) \times 4 (IRAP task) within-subjects ANOVA. This revealed a significant main effect of Valence: $F(1,19) = 67.05, p < .001, \eta^2 = .78$, as predicted by H1. There was also a significant Category \times Valence two-way interaction: $F(1,19) = 12.26, p < .005, \eta^2 = .392$, and a significant IRAP \times Category \times Valence three-way interaction: $F(3,57) = 5.77, p < .005, \eta^2 = .233$. No other main effects or their interactions were significant (all $F_s < 1.57, p_s > .21$). As shown in Figure 3.2, *True* responses were faster than *False* responses on positive word trials while for negative word trials there were no differences between pressing *True/False* supporting H1. The significant three-way interaction arises because *True* responses were faster than *False* responses for Category 1 stimuli across all four IRAPs. In contrast, for Category 2 stimuli *True* responses were only faster than false responses

¹⁶ Separate analysis of each of the four IRAPs (Non-word, Social System, Nature, and Weight) can be found in Appendix 3.

¹⁷ Preliminary analyses revealed the sentence sequence instructions (e.g., **Thin Person** described first or **Fat Person** described first) in the standard condition and the order of the IRAP blocks presented (e.g. beginning with either a pro-**Thin** or a pro-**Fat** block) did not influence the critical IRAP effect in subsequent analyses.

for the Non-word and Social System IRAPs (i.e., those in which the stimuli had relatively weak prior associations).

This pattern of results is assumed to reflect differences in the strength of prior associations across the four IRAP types. For the two IRAPs with the weakest prior associations (Non-word and Social System) there was little influence of whether the description was congruent or incongruent with the stimulus being presented. Thus, for those IRAPs there was little difference between scores for Category 1 versus Category 2 stimuli but there was an overall effect of Word Valence. That is, *True* vs. *False* responses were faster on positive word trials. In contrast, for the IRAPs with stronger prior associations (Nature and Weight), there was an effect of stimulus congruence. That is, congruent responses (e.g., positive word – flower, negative word - insect) were faster than incongruent responses (positive word – insect, negative word – flower). As for the Non-word and Social System IRAPs, the effect of the PFB was to increase the speed of *True* responses on positive word trials. This resulted in an increased difference between positive word and negative word trials for positive stimuli (e.g., Flowers/Thin) and a corresponding smaller difference for negative stimuli (Insects/Fat).

Test of H2: The influence of the PFB on estimates of absolute implicit attitudes

The absolute D-IRAP (estimate) scores were analysed with a 4 (trial type: category 1-positive words, category 1-negative words, category 2-positive words, category 2-negative words) \times 4 (IRAP task: Non-word, Social System, Nature, Weight) \times 3 (framing: standard, positive, negative) mixed ANOVA, with trial type and IRAP task as the within-subjects factors and framing condition as the between-subjects factor. This analysis revealed a significant main effect of trial type, $F(3, 171) = 64.52, p < .001, \eta^2 = .53$, but not of IRAP task, $F(2.62, 149.34) = .40, p = .76, \eta^2 = .01$. Importantly, the trial type \times IRAP task interaction was significant, $F(7.02, 399.99) = 10.26, p < .001, \eta^2 = .15$. As shown in Figure 3.3 (top left), for stimuli with absent or weak priori associations (i.e., the Non-word IRAP and the Social System IRAP),

elevated implicit attitudes were found (see Cat1-Positive and Cat2-Positive). For stimuli with stronger prior associations (i.e., the Nature & the Weight IRAPs) this pattern was weaker.

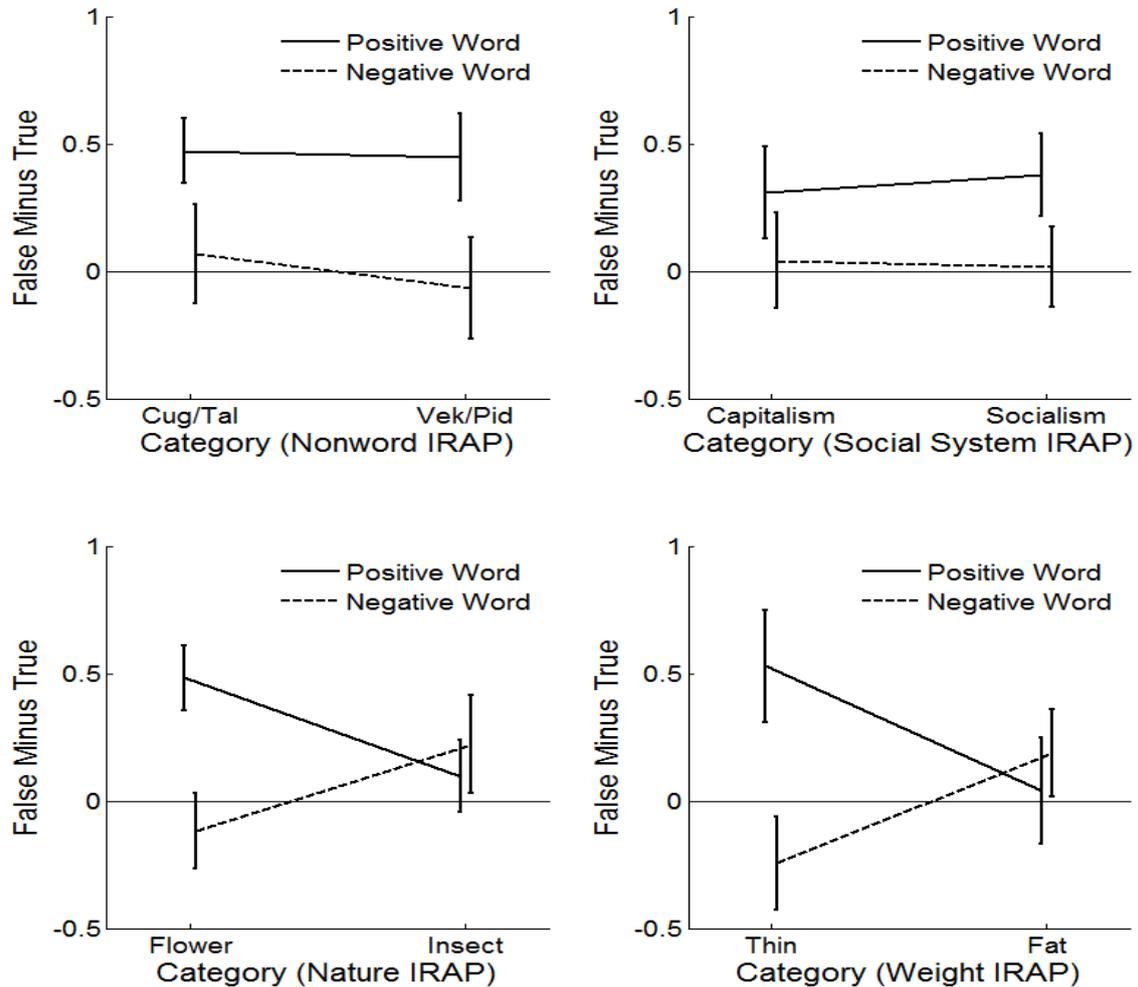


Figure 3.2: Mean *False Minus True* D-IRAP scores as a function of Category, Word Valence and IRAP task. Values above zero indicate that responses for pressing *True* were faster than responses for pressing *False* while values below zero indicate a faster *False* vs. *True* response bias. Error bars with 95% confidence intervals have been included.

A significant main effect was found for framing, $F(2, 57) = 42.40, p < .001, \eta^2 = .60$. As shown in Figure 3.3 (top right) the positive framing condition produced elevated estimates of implicit attitudes ($M = .31$) whilst negative framing resulted in reduced estimates of implicit attitudes ($M = -.09$). The standard framing condition showed estimates of implicit attitudes that were in between the estimates obtained in the positive and negative framing conditions ($M =$

.16). Across the four trial types the framing manipulation had a similar effect — a framing \times trial type interaction was not found, $F(6,171) = 1.02, p = .41, \eta^2 = .04$. Crucially, the framing \times IRAP task interaction was significant $F(6,171) = 9.50, p < .001, \eta^2 = .25$. Figure 3.3 (bottom left) shows that framing influenced the estimates of participants' implicit attitudes and that this influence was weaker for strong prior associations. Lastly the three way interaction was not significant $F(18, 504) = .62, p = .89, \eta^2 = .02$. The overall pattern of results suggest that: i) participants had positive implicit attitudes (estimates) across all the IRAP tasks, ii) framing had a strong effect on the estimates of implicit attitudes, and iii) the effect of framing was stronger for associations with weaker rather than stronger prior associations.

Test of H2: Combining the positive and negative trials for each category in the standard condition

In order to match the standard way the IRAP results are reported in the literature we combined/averaged the absolute D-IRAP (estimate) scores for the positive and negative word associations in each category (see Figure 3.3; bottom right). For example, 'implicit attitudes' for 'Thin Person' were calculated by averaging the D-IRAP (estimate) scores from the 'Thin Person-Positive Words' and 'Thin Person-Negative Words' trials. One-sample t-tests were conducted on the resulting values which revealed that the scores for the Non-word and Social System trials (absent/ weak prior associations) were significantly greater than ($t_s > 2.95, p_s < .01$) or marginally greater (for Capitalism; $t = 1.90, p = .07$) than zero. In the Nature IRAP there was an expected pro-flower bias ($t = 7.24, p < .001$). The apparent neutral attitude towards insects ($t = -1.07, p = .30$) is likely due to the tendency for participants to frame the IRAP task in a positive way, with this PFB then offsetting the negative bias people usually possess for insects. This is similar to the Weight IRAP in which a strong positive attitude was observed for thin people ($t = 4.43, p < .001$) but a neutral attitude was found for fat people ($t = -1.03, p =$

.32). These findings further underline the problem of using the IRAP's absolute results because they are likely to overestimate the positivity of a person's implicit attitudes.

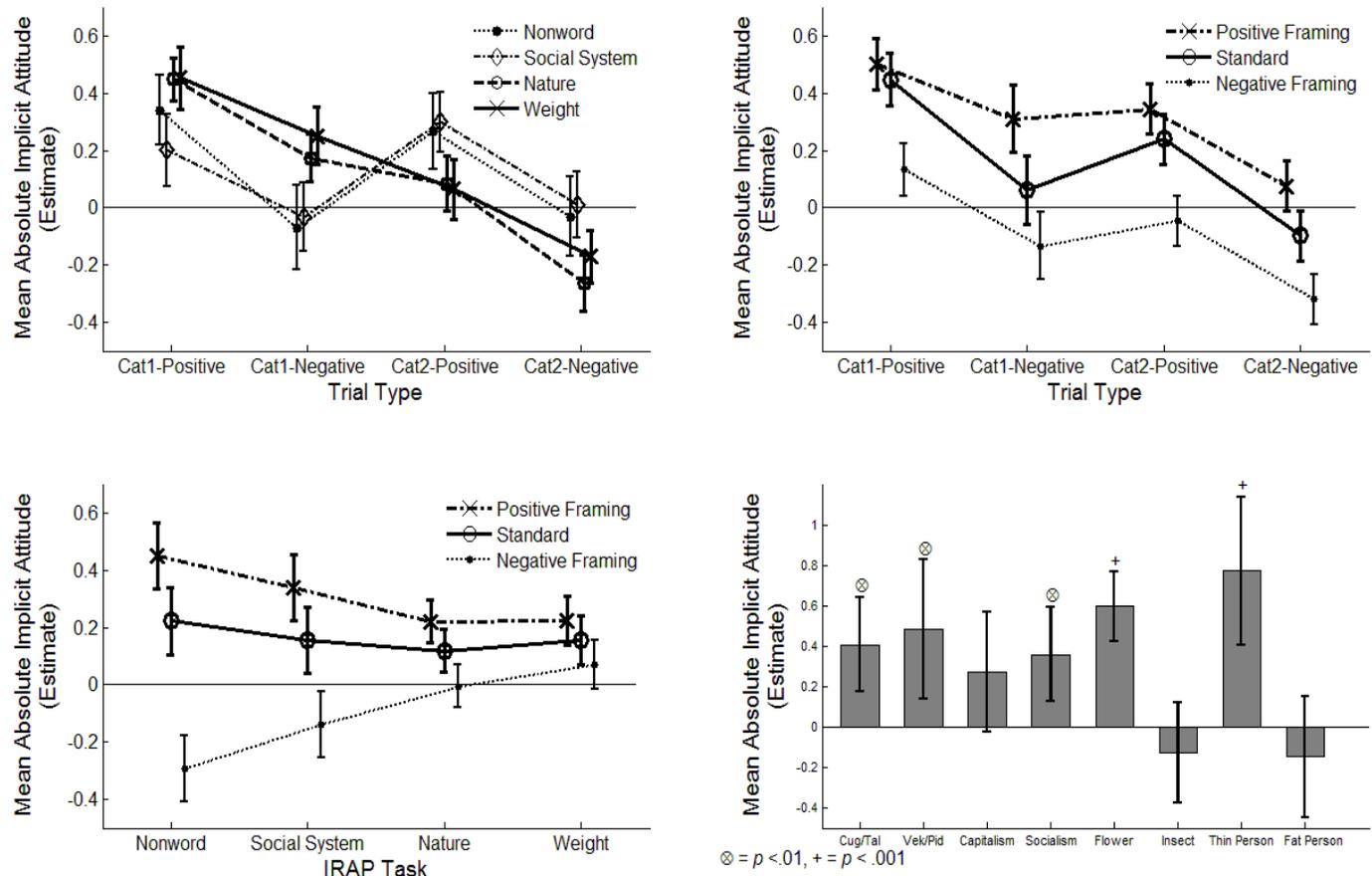


Figure 3.3: (Top Left): The mean absolute D-IRAP (estimate) trial type × IRAP task interaction. Points above the zero line indicate a positive attitude towards the category under investigation. This occurs when *True* responses are faster than *False* responses for categories presented with positive words and also when *False* response are faster than *True* response for categories presented with negative words. Points below the zero line indicate a negative attitude. This occurs when *False* responses are faster than *True* response for positive word associations and also when *True* response are faster than *False* response for negative word associations. Error bars that cross the zero mark indicate a statistically neutral attitude; those that do not cross zero indicate a significant positive or negative attitude. (Top Right) The mean absolute D-IRAP (estimate) trial type × framing condition interaction. (Bottom Left) The mean absolute D-IRAP (estimate) IRAP task × framing condition interaction. The IRAP tasks are ordered from no prior associations (Non-word IRAP; leftmost) to strong prior associations (Weight IRAP; rightmost). (Bottom Right) The mean combined/averaged absolute D-IRAP (estimate) results for each category. Error bars with 95% confidence intervals have been included.

Test of H3: The influence of the PFB on the relative IRAP results

The relative overall D-IRAP scores were analysed using a similar method to that used to analyse IAT data to assess if there were differences across the standard, positive and negative framing conditions for each of the four IRAP tasks. For all the IRAP tasks there were no significant differences across the three framing conditions (Non-word = $F(2, 57) = .35, p = .70, \eta^2 = .01$; Social System = $F(2, 57) = .19, p = .83, \eta^2 = .01$; Nature = $F(2, 57) = .30, p = .74, \eta^2 = .01$; Weight = $F(2, 57) = .28, p = .78, \eta^2 = .01$; see Figure 3.4).

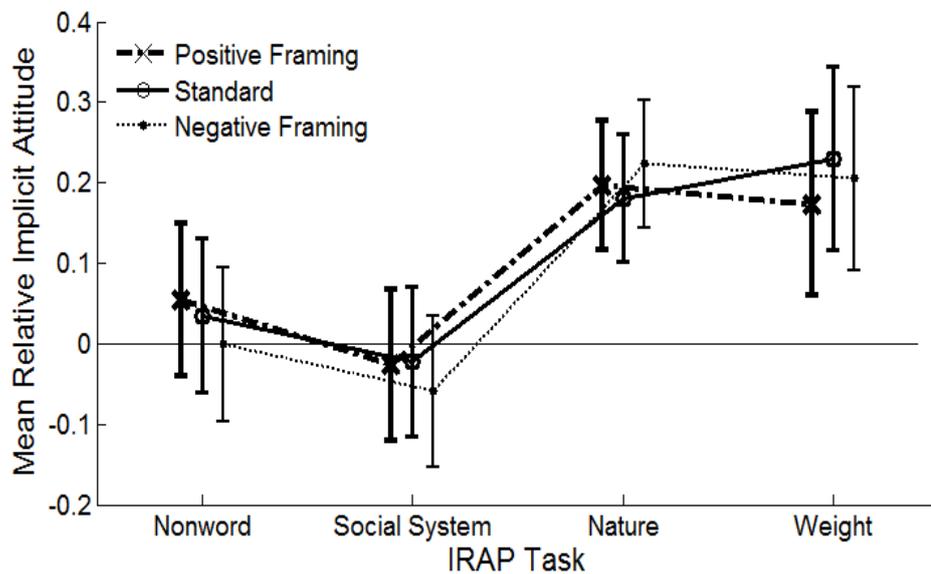


Figure 3.4: The mean relative D-IRAP score for the four IRAP tasks. Error bars with 95% confidence intervals have been included.

In each of the framing conditions for the Non-word and Social System IRAP tasks (absent/ weak prior associations), the D-IRAP scores did not differ significantly from zero (all $t_s < 1.51, p_s > .15$), indicating neutral attitudes to the Non-words as well as Capitalism relative to Socialism. In each of the framing conditions for the Nature and the Weight IRAP tasks (strong prior associations), the D-IRAP scores were a significant distance away from zero (all

$ts > 3.30, ps < .01$), indicating the predicted pro-flower/anti-insect bias and the pro-thin/anti-fat bias.

Discussion

The main aim of the present study was to determine if default (positive) framing biases might compromise the validity of the IRAP (Barnes-Holmes et al., 2010). According to previous research (Matlin, 2016; McKenzie & Nelson, 2003) people have a general and default bias to frame events and objects in a positive way (a Positive Framing Bias; PFB) and it is possible to manipulate this bias through framing. Based on the procedure and structure of the IRAP we reasoned that such a PFB might have the effect of biasing estimates of absolute attitudes. Indeed, such a bias might account for previous results which seem either counterintuitive (e.g., Hussey et al., 2015), or inconsistent with prior research (e.g., Roddy et al., 2011; 2012). We suggest that the effect of a positivity bias in this situation would lead to more rapid *True* responses to categories presented with positive words than *False* responses for the same stimuli. This in turn would lead to inflated estimates of positive attitudes towards the stimulus set being examined. If this is indeed the case, the validity of the IRAP as a measure of absolute implicit attitudes would be severely limited.

We set out to test three hypotheses: (H1) participants hold a default (as measured in the standard condition) Positive Framing Bias (PFB) which results in a decrease in the time required for them to press *True* to categories presented with positive words than to press *False*. (H2) framing the response task in different ways should influence the estimates of absolute implicit attitudes obtained by the IRAP, with weak prior associations being more malleable, and (H3) framing biases or direct manipulations would have no influence when a relative category analysis of the data is carried out.

The effect of a default Positive Framing Bias (H1)

In the standard condition the absolute IRAP scores were positively biased across all four IRAPs. In these IRAPs participants were generally faster at pressing *True* rather than *False* when associating any category with positive words. The implication of this finding is that researchers using the IRAP might incorrectly interpret their results as showing that participants have a positive (absolute) attitude towards abstract categories (e.g., social systems or atheism: (O’Shea, 2017a). In contrast, the ‘true’ attitudes could be either neutral or even negative and so this finding calls into question the validity of the IRAP procedure. Of note, even the IRAP tasks that used stimuli with strong prior associations (Nature and Weight) were influenced by this PFB and, as predicted, a reduction in positivity was seen for categories that usually show a negative bias when assessed by other measures (e.g., insects and fat people), resulting in neutral absolute attitudes as measured by the IRAP.

These findings could account for at least some of the unusual results that have been previously published using the IRAP. For example, as detailed earlier, Roddy et al. (2011, 2012) reported a neutral bias towards fat people whereas we might expect a negative attitude. Similarly, Barnes-Holmes et al. (2010) reported that participants evaluated black and white non-security related civilians with a gun as ‘Safe’ whereas we might expect that they would be viewed as dangerous. Both these results could be explained as an artefact caused by a PFB. That is, participants are likely to be simplifying the task by picking just one of the two possible associations they have to perform on a block (usually the positive) and basing all responses on that association¹⁸. This bias in turn leads to an overestimation of the absolute positive attitude

¹⁸ After participants completed the current experiment those in the standard framing condition were asked to write down any strategies they used to complete the task. This question was not needed for the participants in the positive and negative framing condition because they were

towards a stimulus set because, as we have shown, *True* responses to positive associations are faster than *False* responses.

The influence of framing on estimates of absolute attitudes (H2)

We also determined the effect of directly manipulating the framing of the task on estimates of absolute implicit attitudes. If the IRAP results are influenced by framing, then directly manipulating the frame should have a related influence on the ‘implicit attitude’ (estimate) scores obtained when an absolute analysis of the data is calculated. As argued previously, we would expect that weak or absent prior associations would be more likely to be influenced by the framing effect than those with strong prior associations. This prediction was also confirmed. For the Non-word and Social System IRAP there was a strong effect of framing. That is, in the positive framing condition scores were elevated across all trial types whereas in the negative framing condition scores were reduced. In contrast, for the Nature and Weight IRAP the framing effects were smaller, particularly in the negative framing condition.

These outcomes suggest that the reason that using the absolute results from IRAP trial types are problematic is that responses are influenced strongly by how a person frames each block of trials. It is noteworthy that the way in which the associations are phrased before beginning an IRAP block (especially the sequence) are practically non-existent in published

specifically given a strategy to follow. Of the 10 participants in the standard framing condition who made any references to focusing on mainly the associations, all mentioned the positive associations and none made reference to a focus on the negative associations. The remaining 10 could not describe any strategies they used other than remembering the two possible associations they had to perform on each block of trials. This suggests that when participants did spontaneously choose an explicit strategy they chose one based on positive associations.

IRAP reports. Yet we have shown that they can have an influence on the results obtained, particularly if a researcher primes a participant in some way to focus on either the negative or positive associations. In summary, the IRAP claims to determine absolute implicit attitudes. However, as we have shown, framing effects can influence the absolute value of the scores obtained, making them unreliable at best. Such framing biases can arise naturally and by default (as in the standard condition) to frame objects positively (the PFB). In addition, differences in the way the IRAP task is explained to the participants can have large influences (as in the positive and negative framing conditions).

The influence of framing on the relative results (H3)

The third hypothesis was that the framing effect would not have an influence on relative attitudes (when the data were analysed using the standard IAT methodology). Use of stimuli with weak or no prior associations was expected to lead to neutral attitudes, while stimuli with strong prior associations were predicted to lead to findings showing the expected pro/anti-bias (e.g., pro-thin/anti-fat). Again, all these predictions were confirmed across the four IRAPs. When analysed in a relative manner we found the Non-word IRAP and the Social System IRAP indicated neutral attitudes towards the various category labels. For the Nature and Weight IRAPs, the results were consistent with previous IAT studies (i.e., pro-flower/anti-insect; pro-thin/anti-fat). This suggests that that the IRAP can still be used when a relative analysis of the data is to be carried out.

One possible benefit of using the IRAP as a relative measure is that it appears to be difficult for participants to fake their responses (McKenna, Barnes-Holmes, Barnes-Holmes, & Stewart, 2007) but see Hughes et al., (2016) for more critical findings. The major disadvantages of the IRAP as a relative measure are that it is more complicated, takes longer to complete and has a higher attrition rate than the IAT (Golijani-Moghaddam et al., 2013).

Developing a new absolute measure of implicit attitudes

In addition to the framing biases revealed here for the IRAP, other proposed absolute methods also suffer from problems. As discussed earlier, there are problems with both the Go/No-Go task and the Extrinsic Affective Simon Task for measuring absolute implicit attitudes (Bar-Anan & Nosek, 2014; De Houwer & De Bruycker, 2007). Initial results obtained using the single concept IAT variations (e.g., Karpinski & Steinman, 2006) have shown promise in measuring absolute implicit attitudes. However, the IAT and its derivatives are not without their limitations. For example, the sequence or order (Klauer & Mierke, 2005) in which participants carry out the critical blocks has been shown to influence the IAT results. For example beginning with the Fat Person-Positive and Thin Person-Negative block results in a weaker pro-thin/anti-fat IAT effect than when the presentation order is the other way around. However, it is possible that including more practice trials after the first critical block reduces this problem (Nosek, Greenwald, & Banaji, 2005). Nonetheless, it remains to be determined whether the PFB influences the results of the single concept IATs in a similar way to those of the IRAP.

Diversifying the tools used to measure implicit attitudes is likely to enhance our understanding of them. Consequently, the authors are currently developing a new implicit tool called the Simple Implicit Procedure (SIP) which was inspired by the present work. In brief, this tool has parallels to the IRAP but, as the name suggests, rather than participants having to respond to two opposing associations during blocks, the SIP will only involve one association per block. For example, before each block in the IRAP participants could be told ‘On this block please respond as if Thin Person is Negative and Fat Person is Positive’, but during the SIP they could be told ‘On this block please respond as if Thin Person is Negative’ which is much simpler as it only requires one association to be kept in memory.

We anticipate that this change will reduce participants' PFB because they will be exposed to equal frequencies of instructions that are framed positively and negatively¹⁹. Keeping just one association in memory will reduce participants' cognitive load and hence their need to use a simplifying strategy/method to make the task easier. To ensure that participants do not focus solely on the positive and negative words and ignore the category stimuli, an inhibition/alternative response key will be used in some of the trials. For example, participants will have to press the Space Bar when 'Fat Person' and any valenced word appears on a 'Thin Person' block.

The SIP will also bring the ability to measure a single concept (e.g. self-esteem) without the need for a comparison group. Additionally, this tool could act as both an absolute and relative measure by simply adding in a comparison group/category. This feature would enable researchers to obtain a general overview of participants' attitudes through the relative comparisons but a more complete understanding of implicit attitudes could be observed with the absolute results. The order or sequence effect apparent in the IAT (Klauer & Mierke, 2005) is not expected to have as much of an influence when using the SIP. For example, in a Fat-Thin SIP participants will respond in one of four possible blocks of trials at any point (i.e. Thin

¹⁹ In the IRAP participants have to press "True" or "False" to positively valenced word and "True" or "False" to negative valenced words depending on which "Pro" block they are carrying out. The SIP, however, will require participants to press "True" to the positive valenced words in the positive association blocks and "True" to the negatively valenced words in the negative association blocks. The mean latency difference between pressing "True" on different block will be measured. Likewise, "False" is pressed to negative words in the positive association blocks and to positive words in negative association blocks. Using this procedural set up should remove the PFB people have.

Person-Positive, Thin Person-Negative, Fat Person-Positive, Fat Person-Negative). Each participant will experience a random sequence of these four blocks in the SIP (24 possible sequences) which represents an improvement over the two possible sequences available for the IATs. Pilot studies are currently being carried out to test the validity of the SIP.

Consideration of special populations

Although the current research implies that people in general have a PFB, it is possible that the results would not generalize to specific groups, namely those with depression or other emotional disorders. There is a wealth of evidence showing that these groups are in fact biased towards negative thought processes (e.g., Browning, Holmes, & Harmer, 2010). It would be illuminating to see if those with depression are more likely to focus on negative associations, resulting in a Negative Framing Bias (NFB) when conducting the IRAP. Indeed, (Kosnes, Whelan, O'Donovan, and McHugh (2013) provided evidence for this suggestion using the IRAP with a sub-clinically depressed sample and a normative one. The authors claimed they were measuring participants' responses to future thinking (i.e., measuring the difference between pressing *True* and *False* to 'I expect' or 'I don't expect' which were presented with positive and negative words).

While the experimenters interpreted their results as the normative sample having positive future thinking and the subclinical having negative future thinking, the current findings suggest an alternative interpretation. That is, the statements 'I expect' and 'I don't expect' are unlikely to evoke strong prior associations, similar to the Non-word IRAP and Social System IRAP above. Therefore, the cognitive heuristics or framing strategies a participant engages in (e.g., a normative sample having a PFB and a depressed sample having a NFB) may be the driving factor behind Kosnes et al.'s (2013) results.

Conclusion

The current study has provided the first empirical evidence that people have a tendency to spontaneously frame opposing associations in a positive way (Positive Framing Bias; PFB) when put under time pressure. This bias can be accentuated when participants are encouraged to focus on the positive associations and reversed when they are prompted to focus on the negative associations. These findings seriously question the validity of the IRAP as a tool for determining absolute implicit attitudes.

**Chapter 4: The Simple Implicit
Procedure (SIP): A new method for
measuring implicit cognition**

Abstract

Diversifying the tools used to measure implicit attitudes will enhance our understanding of the response biases that such tools measure and will enable us to control for method-specific variations (i.e., use of sorting tasks vs. affirming and negating tasks). Here the Simple Implicit Procedure (SIP) is introduced as a tool for capturing individual's implicit attitudes and stereotypes. In the SIP, participants must respond correctly to statements using affirming ("Yes") and negating ("No") responses. If participants are asked to "respond as if Flowers are Positive", "Yes" is the correct response when a flower and a positive word appear together, while "No" is the correct response for a flower and a negative word pair. Likewise, participants must respond in the opposite manner to the statement "respond as if Flowers are Negative". This procedure allows separate biases towards flowers and positive words, and flowers and negative words, to be measured. A contrasting category (i.e., insects) can also be included, allowing for the measurement of relative attitudes. Using the SIP across two attitude domains with near universal evaluative differences (i.e., flowers vs. insects and carers vs. criminals), attitudes towards the self, and a socially sensitive stereotype domain, we find that measures of implicit attitudes obtained using the SIP correlate with explicit measures, have acceptable internal reliability and show added predictive validity over the explicit measure in the socially sensitive domain. Furthermore, unlike the Implicit Association Test, the SIP is not subject to problematic order and practice effects. The SIP may therefore be useful in clinical settings where use of control groups is not ethical. These results show the unique benefits that the SIP can offer to attitude researchers.

It has been almost 20 years since the introduction of the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) and it is still the most widely used (Nosek et al., 2011) and reliable measure for estimating an individual's implicit social cognitions (Bar-Anan & Nosek, 2014; Greenwald et al., 2009). The appeal of the IAT is due in part to its simple procedural set up and the ease with which participants can understand and complete the task. In a typical IAT (e.g., the Flower–Insect²⁰ IAT) participants are successively presented with various pictures of flowers and insects, and positively and negatively valenced words. In one of the two critical blocks, participants must press the “E” key on a computer keyboard if a positive word or a picture of a flower appears and press the “I” key if a negative word or a picture of an insect appears (congruent block). In the other critical block, participants must press “E” if a positive word or a picture of an insect is shown and press the “I” key for a negative word or a picture of a flower (incongruent block).

This task set up aims to measure biases participants have in associating concepts (flowers and insects) with valenced words. The stronger the association, the faster and more natural the sorting task will feel (congruent block), while weaker associations will result in a slowing down of processing times due to the unaccustomed pairings in memory (incongruent block). These response biases are often referred to as implicit attitudes which are “introspectively unidentified (or inaccurately identified) traces of past experience that mediate favourable or unfavourable feelings, thought or actions towards social objects” (Greenwald & Banaji, 1995, p. 8). Explicit attitudes, however, are the result of controlled evaluative judgement and thoughtful introspection.

In all areas of psychological research, particularly the area of implicit cognition, investigations into psychological phenomena depend on the tools of measurement (Gawronski, De Houwer, Reis, & Judd, 2014). The present research develops a new tool called the Simple

²⁰ The flower-insect example will be used through the introduction.

Implicit Procedure (SIP) which aims to offer an alternative to associative measures of implicit attitudes like the IAT.

Associative versus propositional measures of implicit cognition

The IAT is limited in the sense that it can only measure attitudes towards one category relative to another category (e.g., flower vs. insect; see Blanton & Jaccard, 2006). To overcome this limitation, the IAT has been adapted in various ways to create new measures of implicit attitudes such as the Single Category (SC)-IAT (Karpinski & Steinman, 2006), the Sorting Paired Features Task (SPFT; Bar-Anan et al., 2009) and Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001). These absolute/unipolar measures can estimate implicit attitudes towards separate categories (separate bias towards flowers and insects). However, these measures cannot determine whether a bias is driven by a stronger positive or negative evaluation towards a category because to do so the response latencies for associating a category both with positive words and with negative words are required.

All of these tasks have their advantages and disadvantages (Bar-Anan & Nosek, 2014; Greenwald & Sriram, 2010) but all measure associations between concepts, with the consequence that they ignore how the concepts are related. For example, “I am good” and “I want to be good” both involve an association between “I” and “good” but differ with respect to the type of relation/proposition involved. These two beliefs are of course different and will likely predict different behavioural outcomes, something that associative measures may not detect (De Houwer, Heider, Spruyt, Roets, & Hughes, 2015).

New lines of research on implicit attitudes have questioned the idea that implicit cognition is confined to associations and suggests that implicit cognition may also involve rudimentary propositional/relational abilities (De Houwer, 2014; Hughes, Barnes-Holmes, & Vahey, 2012; for empirical evidence supporting this suggestion see Remue, De Houwer, Barnes-Holmes, Vanderhasselt, & De Raedt, 2013, and Remue, Hughes, Houwer, & Raedt,

2014). This novel way of thinking about implicit cognition led to the development of a task that aims to measure propositions between concepts and not just associations.

This task is called the Implicit Relational Assessment Procedure (IRAP; (Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010). In the IRAP participants must respond correctly on the basis of opposite statements about two categories (i.e., “respond as if flowers are positive and insects are negative”) using affirming (“Yes”) and negating (“No”) relational response options, rather than participants having to sort items into the correct categories as in the IAT. For example, based on the “as if” statement above, when a picture/word of a flower and a positive word appears, “Yes” is the correct response, but if a negative word appears with a flower, “No” is the correct response. Similarly, participants must respond according to the rule “insects are negative”. When these two stimuli appear together, “Yes” is the correct response, and when an insect appears with a positive word, “No” is the correct response. This pro-flower/anti-insect block is expected to be congruent with participants’ prior experiences of positive associations with flowers and negative associations with insects.

On a subsequent incongruent pro-insect/anti-flower block, participants must respond using the correct affirming and negating responses on the basis of the relational statement “respond as if flowers are negative and insects are positive”. Therefore, a “Yes” response will be made when an insect and a positive word, or a flower and a negative word, appear on the screen. A “No” response should be used when an insect and a negative word, and a flower and a positive word appear on the screen. Due to the alternating responding across pro-flower/anti-insect blocks and the pro-insect/anti-flower blocks, the IRAP allows for the assessment of four separate (absolute/unipolar) response biases (Flower–Positive, Flower–Negative, Insect–Positive, Insect–Negative). These four response biases/trial types are found by measuring the mean latency difference in pressing “Yes” vs. “No” across the pro-flower/anti-insect blocks and the pro-insect/anti-flower blocks (see Figure 4.1 for a graphical representation).

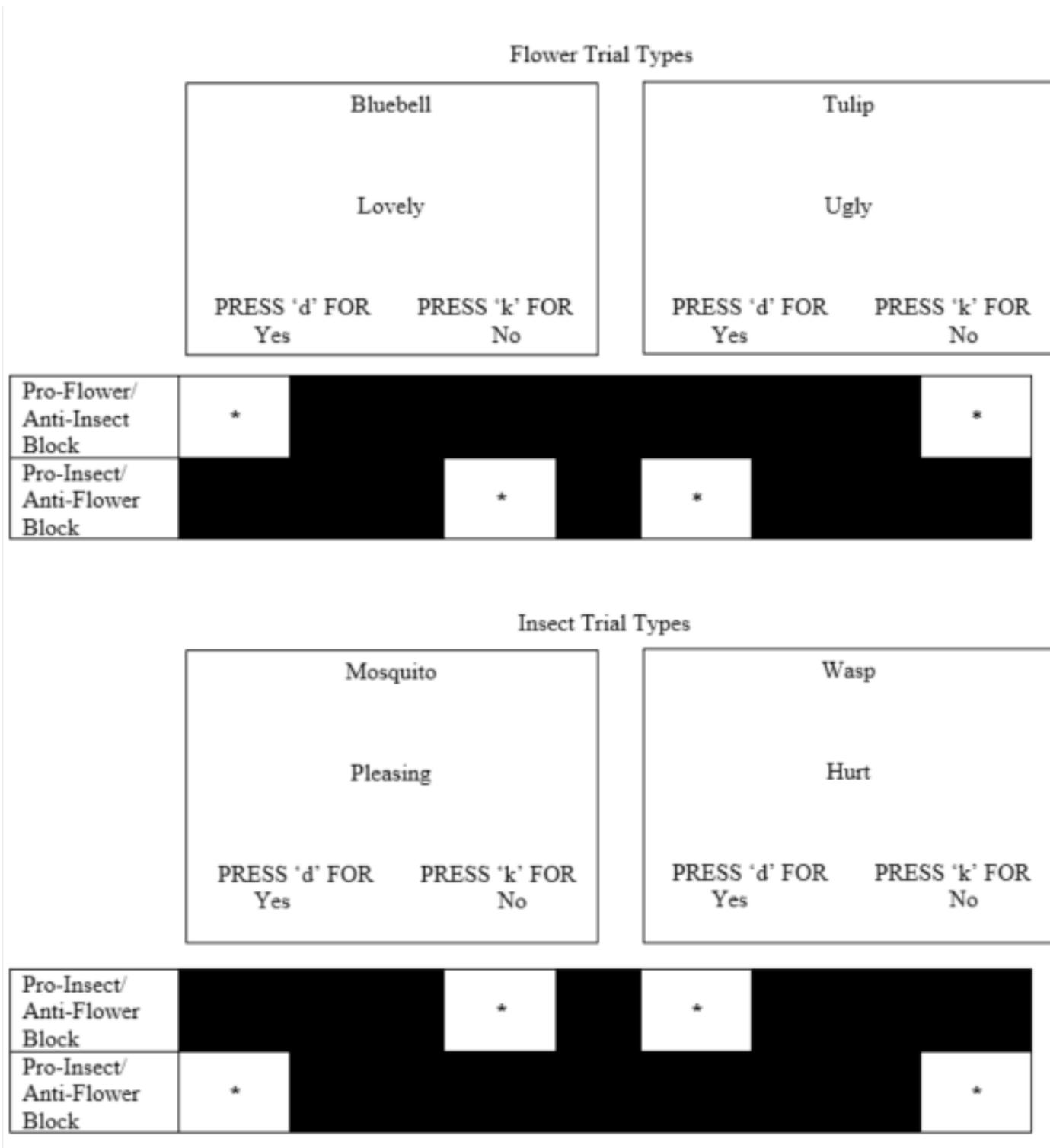


Figure 4.1: Screen shot examples of the four Implicit Relational Assessment Procedure trial types. The category exemplars (a flower or an insect), target word (Enjoy, Cheer, Poison, etc.) and response options (Yes and No) appear simultaneously on each trial. Asterisks indicate the correct response option to press on either block. For each trial type, latency differences are compared between the correct “Yes” and “No” responses across the pro-flower/anti-insect blocks and the pro-insect/anti-flower blocks.

The advantage of the IRAP over other associative implicit measure can be exemplified with the gender-career IAT. In this task, which aims to measure stereotypes, the four categories are: (1) male names, (2) female names, (3) career words and (4) family words. Participants are faster to sort males with careers and females with family than they are to sort females with careers and males with families, which shows a relative male-career/female-family bias on the IAT (Nosek, et al., 2007). With the GNAT, SPFT or the ST-IAT, researchers can test if participants have a bias in associating males with a career or a family, and a separate bias in associating females with a career or a family. Going a step further, the IRAP can measure separately the bias participants have in relating males with a career, males with a family, females with a career and females with a family. This enhanced ability is due to the propositional nature of the IRAP which compares “Yes” versus “No” response latencies in each of the 4 separate trial types²¹.

Limitations of the IRAP

The added specificity that the IRAP has in determining the source of an implicit bias unfortunately comes with some costs. When participants undertake the IRAP, it is immediately apparent that they find the task substantially harder than the IAT and the IRAP has been reported to have high attrition rates (Golijani-Moghaddam, Hart, & Dawson, 2013; Hughes & Barnes-Holmes, 2013). This limitation is likely to be due to the complicated nature of the IRAP which requires participants to keep two opposing relations in memory and to the strictness of the performance criteria ($\geq 80\%$ median accuracy & $\leq 2,000\text{ms}$ median response latency) that

²¹ A similar measure based on the IRAP’s propositional response options is the Relational Responding Task (RRT; (De Houwer, Heider, Spruyt, Roets, & Hughes, 2015). Although the RRT could technically measure separate (absolute) implicit attitudes, the initial publication did not report these details.

must be maintained throughout the task (Barnes-Holmes, Murphy, et al., 2010). The experimenter also needs to impart detailed instructions to the participants to ensure they can complete the task correctly which constrains its utility for gathering data online (De Houwer et al., 2015).

However, more problematically, research by O’Shea, Watson, and Brown (2016)²² demonstrated that great caution should be used when interpreting the zero/neutral point of the four absolute IRAP trial types due to a “default” positive framing bias (PFB) that participants have when completing the IRAP. The PFB occurs precisely because participants are asked to store two relational statements in memory (e.g., “respond as if flowers are negative and insects are positive”). To make the task easier, participants appear to use a cognitive heuristic by storing just one of the relations in memory (i.e., normally the positive relation, e.g., insects are positive) and base all their responses on this relation, regardless of whether it is congruent or incongruent with their prior history of relating these items together.

This bias is illustrated by the findings from O’Shea et al., (2016) that across four different IRAP tasks participants were generally faster to press ‘Yes’ rather than ‘No’ when relating any category with positive words (the four domains were: non-words, capitalism/socialism, flower/insects and thin/fat individuals). This results in an over-inflation of estimates of positive biases towards any category, even those that have negative prior associations such as insects. In addition, using other relational response options (e.g., “True” and “False”; “Confirm” and “Deny”) did not remove or reduce the PFB. Positivity biases that are inherent to the English language (e.g., Dodds et al., 2015; Matlin, 2016; Matthews & Dylman, 2014) and a human propensity toward optimism (e.g. Peterson, 2000) were used to explain this PFB.

²² Reported in Chapter 3

O'Shea et al., (2016) also successfully manipulated the PFB by getting participants to frame the IRAP task either positively or negatively. In these conditions, participants were instructed to focus on only the positive or negative relation in the statement before each block of trials. In the positive framing condition, the positivity of participants' estimates of implicit attitudes across the four domains increased compared to the control. In contrast, for the negative framing condition, estimates of implicit attitudes were reversed compared to those in the positive framing condition, resulting in more negative attitudes across the four domains. The authors recommended that the IRAP's relative comparison results should also be reported (i.e., flowers vs. insects biases) because these yielded the predicted results in line with previous IAT findings. This finding occurs because the PFB gets cancelled out when calculating the relative results.

It has been argued that it is advantageous for the IRAP to be sensitive to positivity biases inherent to languages (Hussey et al., 2015) because the IRAP was designed to capture such patterns in verbal or relational responding. If this assertion is correct, then describing the absolute results in the IRAP as positive, negative or neutral is meaningless, because the estimates of implicit attitudes are automatically shifted towards the positive pole. Therefore, results on the IRAP that apparently show neutral implicit attitudes towards a concept (e.g., O'Shea, 2015; Roddy, Stewart, & Barnes-Holmes, 2010) might, in fact, be reflecting extremely negative attitudes.

The IRAP remains a novel way of estimating implicit attitudes through its propositional methodology, as well as its enhanced abilities to determine where biases lie through the measurement of four separate absolute trial types. However, it is imperative that the IRAP task is simplified to make it easier for participants to complete and to reduce the number of instructions required. More simplicity would help with developing an online version, enabling easier access to participants that are not just WEIRD (Western, Educated, Industrialised, Rich,

Democratic; Henrich, Heine, & Norenzayan, 2010). This greater range of people would increase the generalisability of findings. Furthermore, the PFB in the IRAP contaminates the accurate measurement of absolute implicit attitudes. Therefore, a simplified version of the IRAP that retains the propositional aspect but is not subject to the PFB would be a substantial advance. This possibility motivated the development of the Simple Implicit procedure (SIP).

The Simple Implicit Procedure: SIP

The SIP was designed to remove the PFB apparent in the IRAP and, as the name suggests, simplicity in the procedural set up was a priority. The SIP aims to achieve these two goals by requiring participants to store only one statement in memory before each block, instead of requiring participants to remember two relational statements as in the IRAP. To clarify, before each block in the IRAP participants are told one of two response rules (e.g., (1) “On this block respond as if Flowers are Negative and Insects are Positive” and (2) “On this block respond as if Flowers are Positive and Insects are Negative”). In the SIP, participants are instead told one of four response rules (i.e., (1) “On this block respond as if Flowers are Positive”, (2) “On this block respond as if Flowers are Negative”, (3) “On this block respond as if Insects are Positive”, (4) “On this block respond as if Insects are Negative”).

Having four separate response rules requires participants to keep only one statement in memory on any given block of trials. Keeping just one statement in memory on all blocks of trials reduces participants’ cognitive load and hence their need to use a simplifying strategy/method to make the task easier. This procedure will remove the PFB because equal numbers of positive and negative statements will be stored in memory when participants complete the SIP.

The SIP is visually and structurally like the IRAP (see Figure 4.2). If participants are asked to “respond as if Flowers are Positive”, “Yes” is the correct response when a flower and a positive word appear together, while “No” is the correct response for a flower and a negative

word pair. Likewise, participants must respond in the opposite manner to the statement “respond as if Flowers are Negative”.

The SIP also has an additional “space bar” response option which acts as an inhibition/alternative response to accommodate the contrasting category on each block of trials (e.g., insects on the pro/anti-flower blocks). This “space bar” response is necessary in the SIP to ensure participants relate the category stimuli with the positive and negative target stimuli. If there were no “space bar” responses, participants could focus solely on the positive and negative target words/stimuli and ignore the category stimuli. Hence, the SIP would not be measuring the speed at which participants relate flowers with positive and negative words but rather the speed at which they respond to positive and negative words. To summarise, participants must press “space bar” when “Insect” and any target word (i.e., both positive and negative) appears when they are using response rule 1 (“On this block respond as if Flowers are Positive”) and 2 (“On this block respond as if Flowers are Negative”). Likewise, participants must press the “space bar” when “Flower” and any target word appears when they are completing response rule 3 (“On this block respond as if Insect are Positive”) and 4 (“On this block respond as if Insect are Negative”).

Throughout the following studies, an appropriate contrast category (e.g., Carer vs. Criminal, I am vs. Others are, Male Names vs. Females Names) was always used to allow relative results to be calculated as in the IAT. It is important to note that a contrasting category is not necessarily needed as the “space bar” response could be pressed for any item (e.g., cars, professions, celebrities etc.) that is not a flower if participants are responding to rule 1 or 2 above. See Figure 4.3 for the correct responses on each trial when completing one of the four response rule blocks.

The order or sequence effect that occurs in the IAT (i.e., completing the pro-flower/anti-insect block first rather than second can influence the magnitude of the IAT effect; Nosek,

Greenwald, & Banaji, 2007) is not expected to be a problem when using the SIP. This problem will be greatly ameliorated or eradicated in the SIP for the group level analysis because each participant will complete a random sequence of the four response rules (i.e., 24 different sequences possible). Nevertheless, order effects in the SIP could still occur at the individual level but extremely large samples will be required to determine the influence order has. If only one category is used (no contrast category) then it is likely that a similar order effect to the IAT will still occur in the SIP because only one of two random sequences are possible.

Flower Trial Types								
	<div style="border: 1px solid black; padding: 10px; text-align: center;"> Lavender Lovely Yes Space No </div>				<div style="border: 1px solid black; padding: 10px; text-align: center;"> Bluebell Ugly Yes Space No </div>			
Pro-Flower Block	*						*	
Anti-Flower Block			*		*			
Insect Trial Types								
	<div style="border: 1px solid black; padding: 10px; text-align: center;"> Wasp Pleasing Yes Space No </div>				<div style="border: 1px solid black; padding: 10px; text-align: center;"> Spider Poison Yes Space No </div>			
Pro-Insect Block	*						*	
Anti-Insect Block			*		*			

Figure 4.2: Screen shot examples of the four Simple Implicit Procedure trial types. The category exemplars (a flower or an insect), target word (Enjoy, Cheer, Evil, etc.) and response options (Yes, Space and No) appear simultaneously on each trial. Asterisks indicate the correct response option to press on each block. For each trial type, latency differences are compared between the correct “Yes” and “No” responses across the pro-flower and the anti-flower blocks, and the pro-insect and anti-insect blocks.

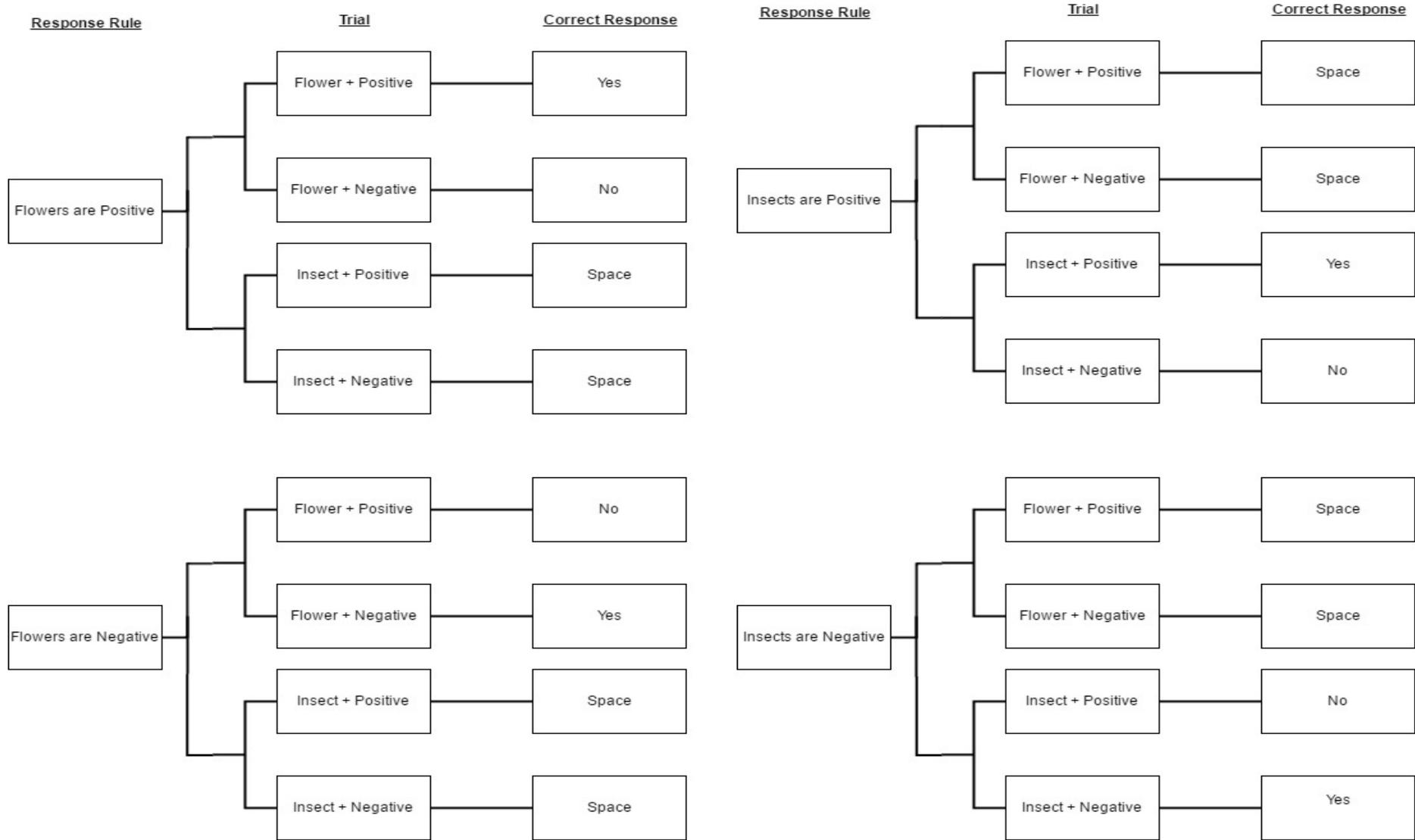


Figure 4.3: Diagram showing the correct response option to press on each trial in the four response rule blocks.

Overview of studies

In the three studies that follow, the reliability and the validity of the SIP as well as its susceptibility to practice or experience effects was examined. These studies addressed attitudes across four domains (1) flowers vs insects, (2) non-words (i.e., *cug* vs *vek*), (3) carer vs. criminals and (4) the self (i.e., “I am” vs. “others are”). One stereotype domain (gender–career vs family) was also examined. For each SIP, the mean “No” minus “Yes” RT differences for each of the four separate trial types (e.g., Flower–Positive, Flower–Negative, Insect–Positive, Insect–Negative) are reported. These allow for an improved understanding of where implicit biases lie (i.e., is a bias driven by the category stimuli being related to either positive words, negative words, or both). In addition, the scores from what we term “compacted” trial types (e.g., the average of the Flower–Positive and Flower–Negative trial types) are reported. It is hypothesised that the compacted trial type results will be less affected by noise in participants’ RTs. This improvement is due to the compacted RT scores being averaged across twice as many trials as the separate trial types. The relative results are also reported (i.e., biases towards flowers relative to biases towards insects). Throughout, internal reliability is reported for each SIP and the correlations between implicit beliefs and criterion variables are also reported (i.e., explicit measures; see De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009, for how to validate implicit measures).

Study 1

Study 1 reports the findings from four separate SIPs assessing implicit and explicit biases towards: (1) flowers and insects (Nature) SIP, (2) carers and criminals (Character) SIP, (3) self-other (Self) SIP, and the (4) gender-career/family (Gender) SIP. It is hypothesised that the SIP’s 4 separate response rules, each with only one statement on a block of trials, will remove the PFB that occurs in implicit measures that use response rules with two statements on a block of trials (i.e. the IRAP and potentially the RRT). This elimination of the PFB in the

SIP would be observed if estimates of implicit biases are in line with the following expectations: for the absolute/unipolar/separate²³ implicit scores, positive biases are observed for flowers, carers and the self on both the positive target word trials (i.e., faster to press “Yes” than “No”) and the negative target word trials (i.e., faster to press “No” than “Yes”). Negative biases are observed for insects and criminals on both the positive target word trials (i.e., people will be faster to press “No” than “Yes”) and negative target word trials (i.e., people will be faster to press “Yes” than “No”). Similarly, on the compacted absolute trial types, pro-flower, pro-carer, pro-self, anti-insect, and anti-criminal bias are expected. These expectations were based on research by Greenwald et al., (1998), Karpinski and Steinman, (2006), Nosek and Banaji, (2001) and public opinion (Gallup, 2017). No predictions were made for the implicit bias scores for the “Other” category in the Self SIP.

Regarding the Gender SIP, the target stimuli are career and family words instead of positive and negative target words. It was expected that participants will be faster at relating both male and female names with career words as well as family words (i.e., faster to press “Yes” than “No”) on the separate trial types (i.e., Male-Career, Male-Family, Female-Career, Female-Family). The compacted (absolute) male trial type scores were predicted to show faster “Yes” responses when relating males with career than family but for females, faster “Yes” responses would be shown when relating females with family than career.

For the relative implicit scores, pro-flower/anti-insect biases, pro-carer/anti-criminal biases, pro-self/anti-other biases and male-career/female-family biases were predicted (Nosek, Smyth, et al., 2007). Internal reliability was expected to be in line with most implicit measures but would likely not be as good as the IAT (Nosek, et al., 2007) due to each participant having to complete a number of different SIPs in one sitting. The scores of the relative implicit and explicit measures in the Nature and Character tasks were expected to show a positive

²³ These three terms will be used interchangeably throughout this chapter.

correlation. Due to the near universal evaluative differences of people expressing positivity towards flowers and carers and negativity towards insects and criminals, and the greater variability in implicit attitudes between individuals, the absolute implicit and explicit measures were not predicted to show strong correlations. For the Self SIP, positive correlations between implicit and explicit attitudes were expected, and explicit attitudes were expected to predict scores on a validated self-esteem scale (Rosenberg, 1967) better than implicit attitudes. However, for the gender-career SIP, it was expected that participants will explicitly express egalitarian gender roles, while implicit biases will be more in line with traditional views of males being the breadwinner and females being the housewife (e.g., Rudman & Phelan, 2010). Therefore, the explicit measure will not correlate with the implicit measures. Because of this divergence, implicit attitudes are expected to predict scores on a scale addressing traditional gender justification (Jost & Kay, 2005) better than explicit attitudes.

It is hypothesised that: (1) no differences in implicit scores will be seen if images or words are used as the category stimuli, but (2) using more stimuli in each category will lead to stronger apparent implicit biases. It is also hypothesised that (3) a “space bar” response is necessary to accurately estimate implicit biases because it forces participants to view the category stimuli and associate them with the positive and negative words. Hence, without a “space bar” response in the SIP, the task is not measuring implicit biases but instead response biases towards positive and negative words. Therefore, an overall neutral bias would be expected when no “space bar” responses are used. To test this hypothesis, scores from a SIP tasks with no “space bar” responses will be compared with scores from the same SIP task that uses “space bar” response options. Lastly, it is hypothesised that (4) having the “Yes” and “No” response options switch location on the screen throughout a SIP task (i.e., randomising) leads to more variability in participants responding (i.e., lower internal reliability) compared to a SIP

task with stationary “Yes” and “No” response options. Each of the four SIPs used tests one of these four hypotheses.

Method

Participants: The sample consisted of 64 first year psychology students (55 females) from the University of Warwick (UK) who received course credit for completing the current experiment. The mean age of the sample was 19.1 ($SD = .97$) and the majority (42 participants) were white British.

Materials

Demographic Information: Participants’ age, gender, nationality and ethnicity were gathered using an online questionnaire.

Simple Implicit Procedure (SIP): The SIP was programmed using Blitzmax (www.blitzbasic.com/Products/blitzmax.php) and ran on an Intel Windows 7 desktop attached to a 22” LCD screen at a resolution of 1440×900 pixels. A single trial consisted of three elements: (1) On the top of the screen a category stimulus appeared, (2) in the centre of the screen a positive or negative target word/stimulus appeared (see Table 4.1 for all the stimuli used in the four SIPs) and (3) “Yes”, “Space” and “No” response options appeared at the bottom of the screen. “Space” always remained at the bottom centre of the screen, while “Yes” was at the bottom left and “No” was at the bottom right of the screen for half of the participants. The other half had “Yes” on the right and “No” on the left of the screen throughout the task (counterbalanced across participants) (see Figure 4.2). An “E” keypress corresponded to the option on the bottom left of the screen and an “O” keypress corresponded to the response option on the bottom right of the screen and “Space” corresponded to the “space bar” key.

Explicit Semantic Differential Scale (SDS): For the Character and Nature SDS items, a Likert scale was used that ranged from -3 (Strongly Disagree) to +3 (Strongly Agree), where each of the category stimuli were used to form a sentence/question with both the word

“Positive” and “Negative” (e.g., “Doctors are Negative”, “Bluebells are Positive”). For the Self SIP, a similar Likert scale was used for questions that connected “I am” with all the target words in that task (e.g., “I am hated”, “I am smart”). No explicit questions were collected for the “Other” category because of the abstract nature of who the other would constitute (Karpinski, 2004). Similarly, the gender SDS questionnaire used a Likert scale where “Males” and “Females” formed coherent sentences that included each of the career and family stimuli (e.g., “Males want to marry when the time is right”, “Females are effective in corporations”).

Table 4.1: The stimuli used in the four SIPs in Study 1.

Nature SIP		Character SIP		Self SIP		Gender SIP	
Category Labels							
Flower	Insect	Carer	Criminal	I am	Others are	Male Names	Female Names
Category Stimuli							
Bluebell	Mosquito	Doctor	Murderer	Me	They	Jack	Chloe
Daffodil	Spider	Nurse	Rapist	My	Them	Thomas	Emily
Tulip	Wasp	Volunteer	Pedophile	Mine	Their	James	Megan
Lavender	Cockroach	Therapist	Terrorist	Self Myself	People Him/Her	Joshua Daniel Harry	Charlotte Jessica Lauren
Target Labels							
Positive	Negative	Positive	Negative	Positive	Negative	Career	Family
Target Stimuli							
Enjoy	Unpleasant	Helpful	Evil	Smart	Stupid	Executive	Home
Cheer	Poison	Nice	Bad	Beautiful	Ugly	Management	Parents
Happy	Evil	Kind	Nasty	Good	Bad	Professional	Children
Lovely	Damage	Friendly	Wrong	Nice	Awful	Corporation	Cousin
Friend	Ugly	Generous	Terrible	Competent	Useless	Salary	Marriage
Pleasing	Hurt	Gentle	Dangerous	Loved	Hated	Office Business	Wedding Relatives

All negative items were reverse coded such that higher scores indicate a more positive bias. For the Character and Nature SIP, the categories connected with positive and negative scores were averaged to create a compacted absolute explicit SDS score. A relative/bipolar SDS score was obtained by subtracting the negative category compacted score (criminal/insect) from the positive category compacted score (carer/flower) for each participant. For the self-esteem questionnaire, the self-positive and self-negative items were averaged to create a compacted absolute self SDS score. Because explicit attitudes were not gathered towards the “Other” contrast category, a relative explicit SDS score could not be calculated. However, the self-esteem questionnaire used an additional feeling thermometer item where participants responded to the question: “How warmly do you feel towards yourself (0 = very cold, 100 = very warm)”. For the Gender SIP SDS, a positive score on both the career and family items indicate a pro-career or pro-family bias, while negative scores indicate an anti-career or anti-family bias. To calculate the compacted SDS for male names, the male-family items were subtracted from the male-career items for each participant. Similarly, to calculate the compacted SDS for female names, the female-family items were subtracted from the female-career items for each participant. The relative SDS was calculated by subtracting the compacted absolute female SDS from the compacted absolute male SDS.

Rosenberg Self-Esteem Scale (Rosenberg, 1965) is a ten-item scale measuring an individual’s global self-worth. Participants used a four-point Likert scale (Strongly Disagree, Disagree, Agree, Strongly Agree) to respond to the questions.

Gender System Justification Scale (Jost, & Kay, 2005) is an eight-item scale measuring individual’s tendency to legitimise gender inequality. A 9 point Likert scale (1 = strongly agree to 9 = strongly disagree) was used for the item responses (see https://warwickpsych.qualtrics.com/SE/?SID=SV_agEITuIdboAbPeJ for the items used in the explicit questionnaires).

Design and procedure

Each participant completed four separate SIPs. For the Nature SIP, a 4 (SIP trial type: Flower-Positive, Flower-Negative, Insect-Positive, Insect-Negative) \times 2 (stimulus: pictures, words) mixed design was used. For the four SIPs, trial type was always a within-subject factor, and on the Nature SIP, stimulus was a between-subject factor. For the Character SIP a 4 (SIP trial type) \times 2 (stimulus: 1 stimulus, 4 stimuli) mixed design was again used, with stimulus as the between-subject factor. This was also the case for the Self SIP, with the between subject factor having two levels (response; “space bar” response, no “space bar” response). Lastly, the Gender SIP used a 4 (SIP trial type) \times 2 (response: stationary, randomising) mixed design with the response as a between subject factor.

Each participant completed a randomised order of the four SIPs. Within each of the four SIPs, participants were assigned to one of the two conditions. Assignment to a condition was counterbalanced across participants. In the Nature SIP, one of the conditions used names of different kinds of flowers and insects and the other condition used images of the same flowers and insects. For the Character SIP, one condition had only one stimulus in the carer and criminal category. Within this condition, the first participant completed the Character SIP with “Doctor” and “Murderer” as the category stimuli, the second participant with “Nurse” and “Rapist”, the third with “Volunteer” and “Paedophile”, and the fourth participant with “Therapist” and “Terrorist”. This order was then repeated for the remaining participants in this condition. For the other condition in the Character SIP, the participants completed the task with four stimuli in both the carer and criminal categories.

For the Self SIP, the “space bar” response was removed in one of the conditions to test whether the results in the no “space bar” condition differed from the results in the condition that used the “space bar” response. Without a “space bar” response, participants did not need to look at the category label because the stimuli from the contrasting category were never used.

Only the “Yes” & “No” response options were used and therefore, participants only had to look at the valence word to determine the correct response. The Gender SIP had a condition where the “Yes” and “NO” response options remained stationary, and in the other condition, the response options alternated location with each other. This switching occurred in a quasi-random fashion with a requirement that a switch must take place after three trials but a switch could also happen from one trial to the next or after two trials.

Participants completed the experiment individually in a well-lit room. After giving informed consent, each participant was then directed towards the computer to complete their first SIP. Participants read the following on-screen instructions; “In this task, you will have to respond to statements in the appropriate fashion as fast and as accurately as possible” and the experimenter explained how the task was to be performed. For each block, participants had to respond in accordance with one of four rules across a block of trials. For brevity, only the Character SIP’s rules will be explained here; the other three SIPs followed a similar procedure but with their domain specific stimuli.

For the Character SIP, the response rules were: (1) “On this block respond as if the CARER is POSITIVE” (pro-carer), (2) “On this block respond as if the CARER is NEGATIVE” (anti-carer), (3) “On this block respond as if the CRIMINAL is POSITIVE” (pro-criminal) and (4) “On this block respond as if the CRIMINAL is NEGATIVE” (anti-criminal). One of these four response rules was presented to the participants before they started a block of trials. For example, if a participant was completing the block “respond as if the CARER is NEGATIVE”, when a negative word and a word referring to a carer appeared on screen, the correct response was ‘YES’. If a positive word and a word referring to a carer appeared, the correct response was ‘NO’. If a word referring to a criminal was shown, regardless of the target word (i.e., positive or negative) that also appeared on the screen, the correct response was “Space”. Likewise, if a participant was completing the block “respond as if the CRIMINAL is

POSITIVE”, when a positive word and a word referring to a criminal appeared, the correct response was ‘YES’. If a negative word appeared with the word of a criminal, the correct response was “NO”. If words referring to a carer was shown, the correct response was “Space”. Each participant was exposed to a random sequence of the four response rule blocks.

Each participant had to successfully complete the practice blocks which were composed of a minimum of four blocks of trials. Each practice block had 10 trials, which included 4 “Space”, 3 “Yes” and 3 “No” response trials. Trials were presented quasi-randomly, such that the same trial was not repeated across two successive trials. If a correct response was made, the screen was cleared for 400ms before the next trial’s stimuli appeared. If an incorrect response was made a red “X” appeared below the target word and all the stimuli remained until the correct response was selected. Following each block, participants’ median accuracy and response latency were presented and if these scores across the 4 practice blocks averaged to $\geq 80\%$ accuracy and $\leq 1,800\text{ms}$ response latency, they continued to the test blocks. These criteria were used to ensure that participants understood and complied with the block’s response rules.

If participants did not meet the above criteria, the 4 practice blocks were repeated. If they failed to meet the accuracy and response latency criteria on the second attempt the study ended, but this never occurred for any of the participants. This procedure likely led participants to believe that they would eventually have to repeat blocks of trials if they did not maintain the speed and accuracy criteria throughout the SIP. However, this penalty only occurred for the practice blocks but it nevertheless potentially contributed to the high accuracy and quick responses from all the participants throughout the SIPs.

After completing the practice blocks successfully, participants were informed on the screen to “press space bar to begin the test blocks”. For the test blocks, participants completed 16 blocks which were comprised of a random sequence of the 4 response rule blocks for each

participant. This unique sequence was repeated until a participant had completed all 16 blocks. Each test block contained 18 trials consisting of 6 “Spacebar,” 6 “Yes” and 6 “No” response trials. When all the 16 blocks were completed, participants were informed on screen the task was over and instructed to call the experimenter. The experimenter then set up the next SIP for the participant to complete, followed by the third and then the fourth SIP. Each of these SIPs had the same response criteria, number of blocks and trials etc. as the Character SIP apart from the stimuli used.

After participants completed the four SIPS, they were directed towards the online questionnaire to complete the explicit questions. The order in which participants completed the Nature, Character, Self and Gender explicit SDS was randomised and the order of the items within each of the questionnaires were also randomised. Next, the Rosenberg Self-Esteem Scale was completed followed by the Gender System Justification Scale. Following these scales, a number of other questionnaires that were unrelated to the current experiment were completed. Lastly, participants were thanked and debriefed. To complete all these sections, it took participants between 75 – 90 minutes.

Results

The primary data obtained from the SIP are raw latency scores defined as the time in milliseconds that elapsed between the onset of the stimulus and the correct response being made by the participant. The dependent variable was participants’ mean “No” minus “Yes” reaction time (RT) latency for each of the four trial types (i.e., Category 1-Positive, Category 1-Negative, Category 2-Positive, Category 2-Negative) across the pro-Category 1 and anti-Category 1 blocks as well as the pro-Category 2 and anti-Category 2 blocks. For the Category 1-Positive and Category 2-Positive trial types, values above zero indicate a positive bias and negative values indicate a negative bias. For both the Category 1-Negative and Category 2-Negative trial types, scores were multiplied by -1. The purpose of this was to match the valence

output of the Category 1-Positive and Category 2-Positive trial types such that scores above zero indicate a positive bias and scores below zero indicate a negative bias. For example, if participants were faster to press “No” than to press “Yes” to Category 1–Negative words across the pro-Category 1 and anti-Category 1 blocks, that would indicate participants have a positive bias towards Category 1-Negative words.

To minimise the impact that individual differences such as age, cognitive ability and motor skills can have, an individual’s response latencies were transformed using an adaption of the Greenwald et al., (2003) D-algorithm. This D-algorithm is normally applied to IAT response latency scores and similar transformations to response latency scores in the IRAP are also applied (see Barnes-Holmes, et al., 2010). The steps involved in calculating both the absolute and relative D–SIP scores are as follows:

(1) Only response latency data from the 16 test blocks were used; (2) latencies above 10,000ms were discarded; (3) if latencies from more than 10% of a participant’s trials throughout the 16 test blocks were less than 30ms, that participant was removed from the analysis; (4) for each SIP task, 16 individual standard deviations were calculated for each trial type across the 16 test blocks (Category 1-Positive, Category 1-Negative, Category 2-Positive, Category 2-Negative \times 4 = repeating blocks); (5) 16 mean latencies were calculated for both the “Yes” and “No” responses for each trial type across the 16 test blocks; (6) difference scores were calculated for each trial type by subtracting mean latencies of “Yes” responses from mean latencies of “No” responses in each test block pair (i.e., pro- Category 1 block and anti–Category 1 block; pro-Category 2 block and anti-Category 2 block); (7) each difference score was then divided by its corresponding standard deviation from step 4, yielding 16 D-SIP scores, one score for each trial type across the 16 test blocks; (8) four overall trial type D-SIP scores were calculated by averaging the four scores for each of the four trial types across the blocks (these calculations revealed the absolute/non-relative separate trial type results); (9) averaging

the positive and negative trial type for each category showed the absolute/non-relative compacted trial type results); (10) To compute the relative comparison, equivalent to that of the IAT, an overall relative D-SIP score was calculated by subtracting the compacted Category 2 score from the compacted Category 1 score.

The Nature, Character and Self SIP used the above analytic procedures. For the Gender SIP, the task did not measure general attitudes using positive and negative words but instead measured stereotypes with words referring to careers and family. Because the categories were not connected with negative words, none of the trial types were reverse coded. The four overall trial type D-SIP scores (steps 1-8) were calculated using the same methods as above. For step 9, the male compacted absolute scores were calculated by subtracting the Male-Family trial type from the Male-Career trial type. Similarly, the female compacted absolute scores were calculated by subtracting the Female-Family trial type from the Female-Career trial type. Step 10 was the same as above where the relative overall D-SIP score was calculated by subtracting the compacted female score from the male compacted score.

The results from the Nature SIP are reported first, followed by the Character SIP, then the Self SIP and finally the Gender SIP. For each SIP, the findings from the 4 (trial type) \times 2 mixed ANOVA are presented. One-sample t-tests are carried out on the four trial scores, the two compacted trial types scores as well as the relative results to tests which are a significant distance above or below zero. These results indicate whether implicit bias scores towards an object are positive, neutral or negative. Next, the SIP's internal reliability is reported by using odd even split-half reliability analysis. To obtain the odd and even SIP scores, two separate scores were calculated in the same way as for the overall relative D-SIP, except that the algorithm described previously was applied separately to odd trials and to even trials. Several correlational analyses were also used to test whether the scores on implicit and explicit measures were associated. If the scores show a positive correlation, these findings indicate the

measures are assessing similar constructs. However, if the scores are negatively correlated, these findings indicate the measures are assessing distinct constructs. Lastly, for the Self and Gender SIP, hierarchical logistical regression analyses will be used to determine if the SIP has increased predictive validity over the explicit SDS.

Nature SIP: A 4 (trial type: Flower–Positive, Flower–Negative, Insect–Positive, Insect–Negative) \times 2 (stimulus: words, images) mixed ANOVA showed a significant main effect of trial type, $F(3, 186) = 104.54, p < .001, \eta^2 = .63$, a non-significant main effect of stimulus, $F(1, 62) = .15, p = .70, \eta^2 = .00$ and no significant interaction between trial type and stimulus, $F(3, 186) = 1.98, p = .119, \eta^2 = .03$. Because no significant difference between the stimulus types was found, these two conditions were combined and are reported together. Figure 4.4 shows the scores for four trial types, including the scores for compacted flower and insect trial types and the overall D-SIP score.

Follow-up one-sample t-tests were used to determine what was causing the main effect of trial type. The scores on the Flower–Positive trials, $t(63) = 14.67, p < .001$, the Insect–Positive trials, $t(63) = 7.58, p < .001$, and the compacted Flower trials, $t(63) = 10.38, p < .001$, were significantly above zero, while the compacted Insect trials scores did not differ from zero, $t(63) = .045, p = .965$, indicating a neutral attitude. The scores on the Insect–Negative trials, $t(63) = -7.80, p < .001$ was significantly below zero, while the scores on the Flower–Negative trials did not differ from zero, $t(63) = -.920, p = .361$. As expected, the relative score (overall D-SIP) was significantly above zero, $t(63) = 7.06, p < .001$ consistent with a pro-flower/anti-insect bias. To test for internal reliability, the correlation between the odd and even trials was significant, $r = .298, n = 64, p = .017$.

Correlations between the relative implicit and relative/bipolar explicit scores are reported here and in subsequent SIPs. To make use of the absolute/unipolar scores for both the implicit and explicit measures an individual's Category 1 (Flower) and Category 2 (Insect)

implicit and explicit attitude scores are included in the same correlation (e.g., each participant has two correlation points for both their implicit and explicit scores in the analysis). The term “unipolar relative” is used to describe this setup. Based on this unipolar relative setup, two additional correlation scores are also reported. Each of these correlations report scores that alternate from using Category 1 (Flower) implicit and explicit scores to Category 2 (Insect) implicit and explicit scores across each participant. For example, analysis 1 (cut 1) will include, participant 1 = Category 1, participant 2 = Category 2, participant 3 = Category 1, etc. While analysis 2 (cut 2) will include, participant 1 = Category 2, participant 2 = Category 1, participant 3 = Category 2, etc.

The unipolar relative correlations are expected to create more variation across responses and hence stronger correlations between the implicit and explicit variables are expected to be shown. This expectation is especially likely for attitudes that have a near-universal evaluative difference (i.e., almost everyone prefers flowers to insects and carers to criminals). Correlations between the compacted Category 1 and Category 2 absolute trial types and the corresponding absolute/unipolar explicit measure will also be reported.

The correlation between the relative bipolar implicit and explicit measures was not significant, $r = .056$, $n = 64$, $p = .660$. The compacted absolute Flower, $r = .149$, $n = 64$, $p = .240$, and Insect, $r = .110$, $n = 64$, $p = .386$, implicit and explicit scores correlated in the expected direction but they did not reach conventional significance levels. There was a significant positive correlation between the unipolar relative implicit and explicit scores, $r = .528$, $n = 128$, $p < .001$, as were the correlations for cut 1, $r = .465$, $n = 64$, $p < .001$ and cut 2, $r = .554$, $n = 64$, $p < .001$.

To summarise, these findings suggest that using words or images in the SIP produce similar results. The initial results also suggest that perhaps an affirming/positivity bias (not PFB) is still occurring in the SIP because participants were faster to press a “Yes” rather than

a “No” response when relating insects with positive words. This affirming/positivity bias appears to artificially increase scores on the compacted trial types leading to estimates of neutral biases towards insects, even though a negative bias was expected. The internal reliability was found to be reasonable for an implicit measure and although the bipolar and absolute compacted implicit and explicit correlations were not significant, the correlations were in the expected direction. Importantly, the unipolar relative correlations (i.e., each participant with two data point, cut 1 and cut 2) between the implicit and explicit measures were moderately-strongly related.

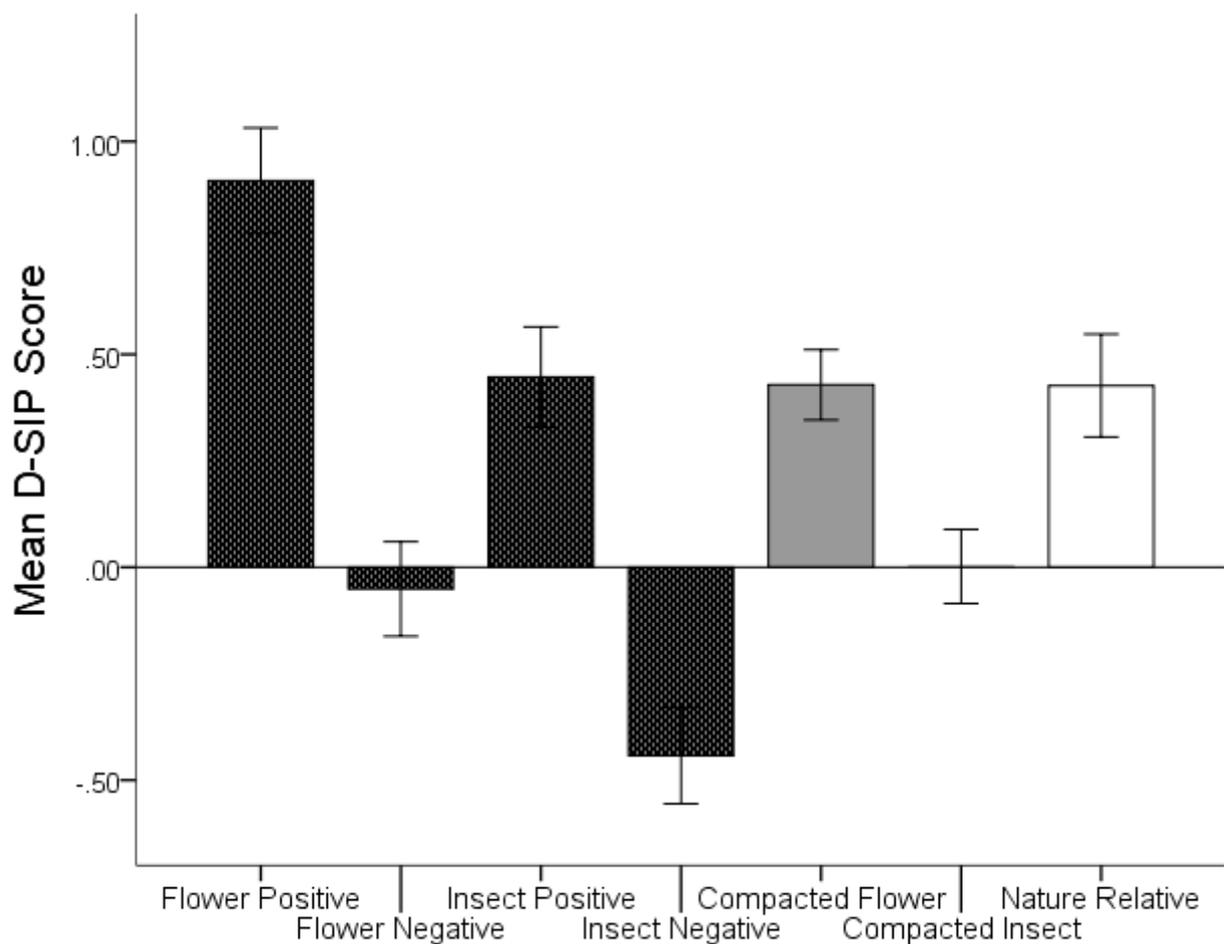


Figure 4.4: The four individual trial types, the two compacted trial types and the relative D-SIP score for the Nature SIP. 95% confidence internal error bars are included. Error bars that cross the zero mark indicate a statistically neutral attitude; those that do not cross zero indicate a significant positive or negative attitude.

Character SIP: The 4 (trial type: Carer–Positive, Carer–Negative, Criminal–Positive, Criminal–Negative) \times 2 (stimulus: one, four) mixed design showed a significant main effect of trial type, $F(3, 186) = 87.97, p < .001, \eta p^2 = .59$. A non-significant main effect of stimulus, $F(1, 62) = 3.53, p = .065, \eta p^2 = .05$ and a significant interaction between trial type and stimulus was shown, $F(3, 186) = 3.18, p = .032, \eta p^2 = .05$. As shown in Figure 4.5, the interaction occurred because there was no difference between the 1 stimulus and the 4 stimulus conditions for Carer–Positive and Carer–Negative, but more negative biases were observed in both the Criminal–Positive and Criminal–Negative trial types in the 4 stimulus condition compared to the 1 stimulus condition.

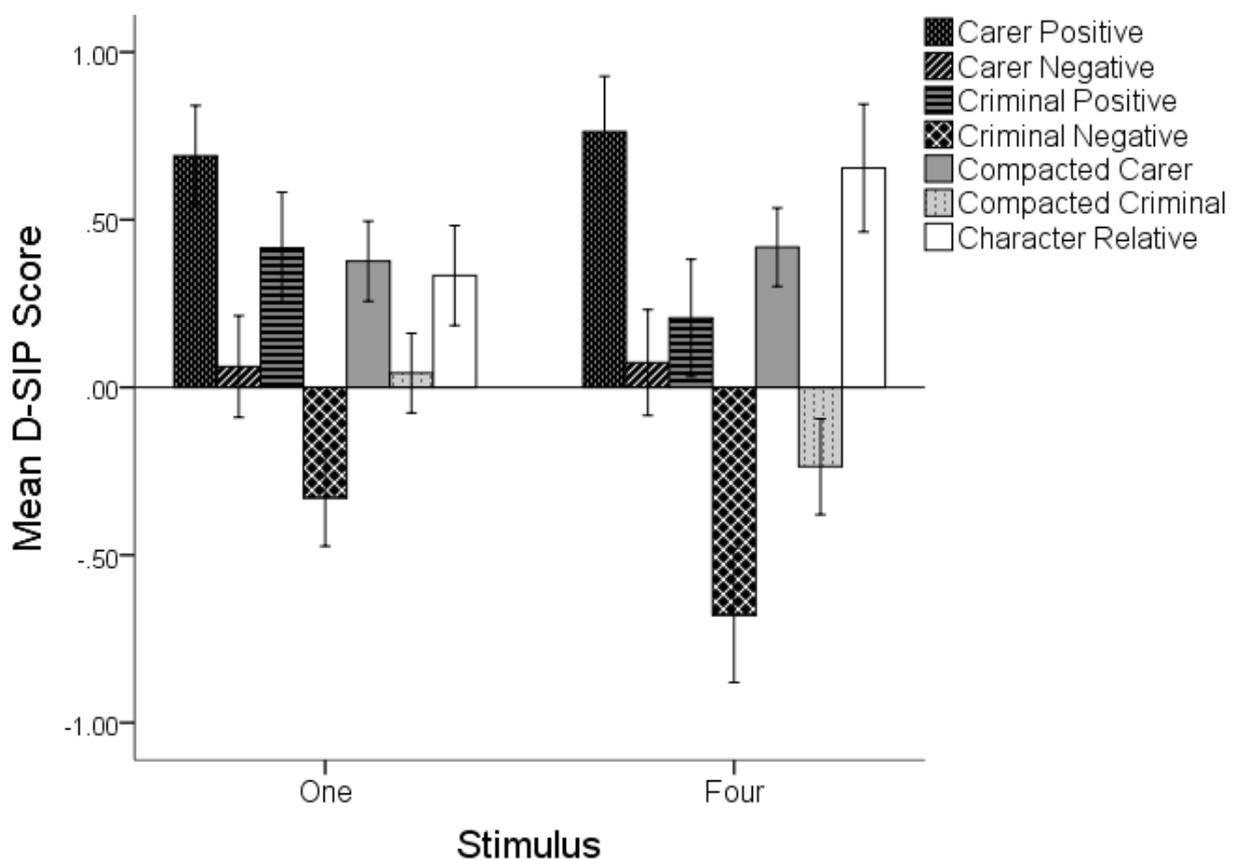


Figure 4.5: The four individual, the two compacted and the overall D-SIP for the two conditions in the Character SIP. Error bars with 95% confidence intervals have been included.

For brevity, we will report the results for the seven one-sample t-tests for only the 4-stimulus condition. These one-sample t-tests will also indicate why the main effect of trial type occurred. Carer-Positive, $t(31) = 9.40, p < .001$, Criminal-Positive, $t(31) = 2.41, p = .022$, and compacted Carers trial type, $t(31) = 7.30, p < .001$ means were significantly above zero. Criminal-Negative, $t(31) = -6.93, p < .001$, and compacted Criminal scores, $t(31) = 3.37, p = .002$, were significantly below zero, while Carer-Negative scores did not differ from zero, $t(31) = .95, p = .35$. The overall D-SIP was significantly above zero, $t(31) = 7.01, p < .001$, consistent with a strong pro-carer/anti-criminal bias.

The correlation between the odd and even trials to test for internal reliability in the 1 stimulus condition was not significant, $r = .262, n = 32, p = .148$. However, for the 4 stimulus condition a significant effect between the odd and even trials was shown, $r = .428, n = 32, p = .014$. The correlations between implicit and explicit scores for the 1 stimulus condition were as follows: the correlation between relative bipolar implicit and explicit scores was significant, but in the opposite direction to expectations, $r = -.370, n = 64, p = .037$. For both the compacted Carer, $r = -.084, n = 64, p = .646$, and Criminal, $r = -.181, n = 64, p = .322$, implicit and explicit scores, non-significant negative correlations were found. The correlation between the unipolar relative implicit and explicit score was significant, $r = .404, n = 64, p = .001$, as was the correlation for cut 1, $r = .504, n = 32, p = .003$, but not cut 2, $r = .284, n = 32, p = .115$.

The correlation between implicit and explicit scores for the 4 stimulus condition were as follows: the correlation between the relative bipolar implicit and explicit scores was not significant, but it was in the predicted direction, $r = .088, n = 32, p = .632$. For both the compacted absolute Carer, $r = .009, n = 32, p = .959$, and Criminal, $r = .111, n = 32, p = .544$, implicit and explicit scores, non-significant effects were found but the effects were in the predicted direction. The correlations between the unipolar relative implicit and explicit scores

was significant, $r = .669$, $n = 64$, $p < .001$, as was the correlation for cut 1, $r = .606$, $n = 32$, $p < .001$ and cut 2, $r = .729$, $n = 32$, $p < .001$.

In summary, the results confirm that there is still an apparent affirming bias in the SIP as can be seen in the criminal positive trial type, where participants were unexpectedly faster to respond with an affirming response (“Yes”) rather than a negating response (“No”), when relating/associating criminals with positive words. Using four stimuli rather than just one for the categories appears to improve the accuracy of the results, especially for the negative category (criminal) as can be seen in the compacted Criminal trial type. The 4-stimulus condition also improved the internal reliability and stronger correlations between the implicit and explicit measures were shown compared to the 1 stimulus condition.

Self-SIP: A 4 (trial type: Self-Positive, Self-Negative, Other-Positive, Other-Negative) \times 2 (response: “space bar”, no “space bar”) mixed ANOVA revealed a significant main effect of trial type, $F(3, 186) = 47.708$, $p < .001$, $\eta p^2 = .45$ and a non-significant main effect of response, $F(1, 62) = 1.20$, $p = .278$, $\eta p^2 = .02$. The interaction between trial type and response was significant, $F(3, 186) = 4.94$, $p = .003$, $\eta p^2 = .07$, and it occurred because all the trials in the no “space bar” response condition were closer to zero/neutral than the trials in the standard SIP with a the “space bar” response (see Figure 4.6). The neutral overall D-SIP score in the no “space bar” condition as shown in Figure 4.6 was in line with expectations.

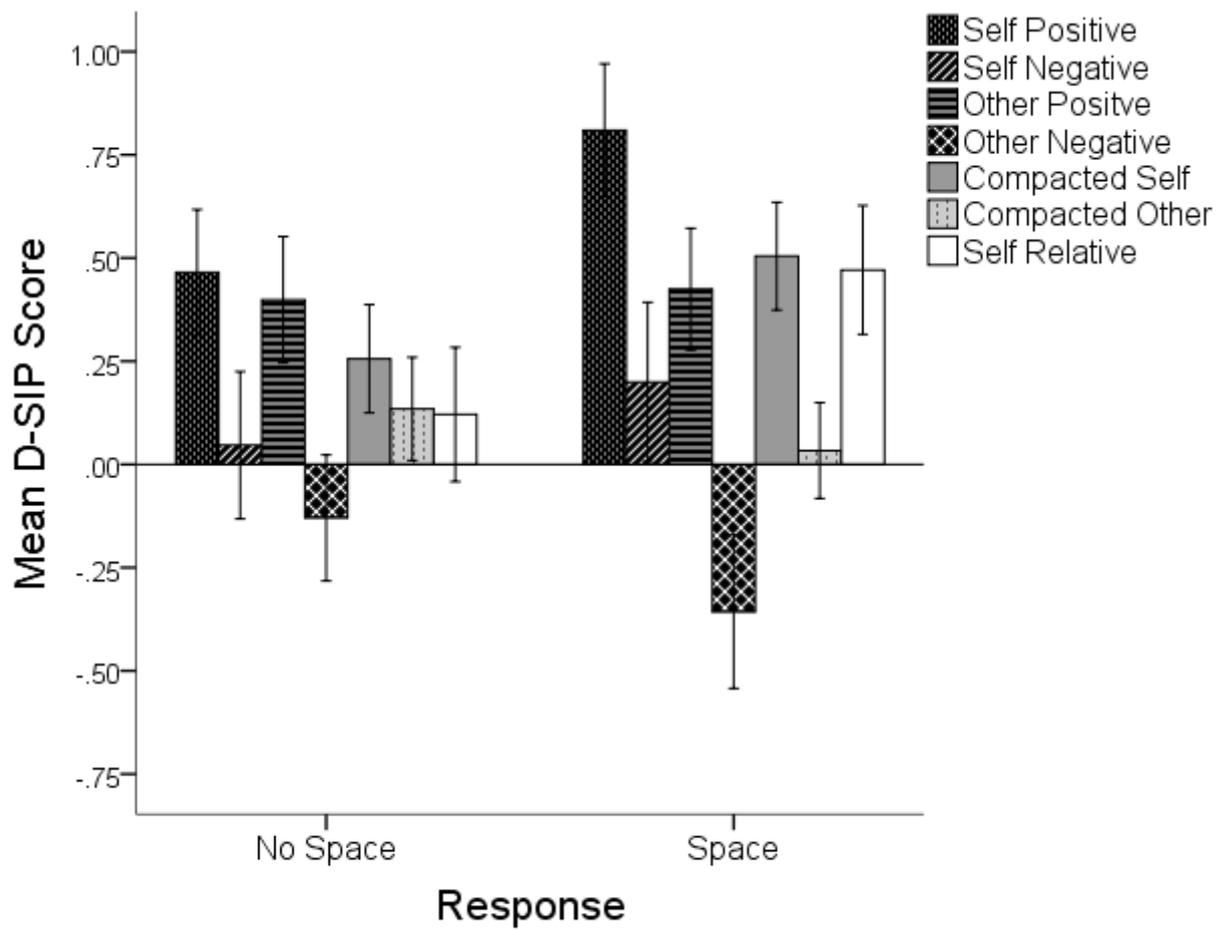


Figure 4.6: The four individual, the two compacted and the overall D-SIP for the two conditions in the Self SIP. Error bars with 95% confidence intervals have been included.

The subsequent Self SIP results will be reported using only the “space bar” version because it is measuring implicit bias while the no “space bar” condition is simple measuring response bias towards words. For the seven one-sample t-tests, scores were significantly above zero for the Self–Positive trial type, $t(31) = 10.25, p < .001$, the Self–Negative trial type (faster to press “No” than “Yes” when relating the self with negative words), $t(31) = 2.11, p = .043$, the Other–Positive trial type, $t(31) = 5.87, p < .001$, and the compacted Self trial type, $t(31) = 7.87, p < .001$. The Other–Negative trial type score was significantly below zero, $t(31) = -3.93, p < .001$, and the compacted Other trial type score did not differ significantly from zero, $t(31) = 0.59, p = .559$. The overall D-SIP score showed a pro-self/anti-other bias because it was significantly above zero, $t(31) = 6.16, p < .001$.

The internal reliability across the odd and even trials was not significant, $r = .104$, $n = 32$, $p = .570$. However, when split half reliability analyses were carried out on the Self and Other trials separately, we found a significant correlation between the odd and even scores in the more meaningful Self trial types, $r = .582$, $n = 32$, $p < .001$ and a non-significant correlations in the less meaningful Other trial types, $r = .073$, $n = 32$, $p = .690$. The correlation between the compacted Self SIP trial type and the explicit Self SDS was significant, $r = .395$, $n = 32$, $p = .025$. The compacted Self SIP and the feeling thermometer correlation approached significance, $r = .334$, $n = 32$, $p = .062$, and the correlation between the compacted Self SIP and the Rosenberg self-esteem scale was non-significant but it was in the predicted direction, $r = .187$, $n = 32$, $p = .306$.

Based on research by Karpinski and Steinman (2006), a composite score was created by averaging together each participant's explicit Self SDS item score, their self feeling thermometer (logged) score and their score on the Rosenberg scale. This composite score significantly correlated, albeit marginally, with participants' compacted Self SIP score, $r = .349$, $n = 32$, $p = .051$. We tested whether the SIP had increased predictive validity over the explicit SDS by running a hierarchical logistical regression analysis. We first entered the Self explicit SDS into the model, and it was a significant predictor of scores on Rosenberg scale, $B = .366$, $p < .001$, accounting for 76% of the variance. When the compacted Self trial type was added into the model, it did not improve it, $B = -.192$, $p = .306$.

These results indicate the SIP could be useful as a tool for assessing implicit self-esteem. The internal reliability of the compacted Self SIP trial type performed better than most other implicit measures used in previous research (e.g., Bosson, Swann, & Pennebaker, 2000). The improved performance of the SIP compared to other implicit measures used to examine the self is also apparent from the compacted Self SIP trial types scores which correlated with the explicit Self SDS scores. This significant correlation is likely due the implicit and explicit

items matching closely (Hofmann, Gawronski, et al., 2005). In addition, finding strong internal reliability for the Self trials and weak reliability for the Other trials might indicate that the Other category used in relative implicit measures like the IAT could be the reason for their low predictive validity (see also Karpinski & Steinman, 2006). These findings also emphasise the importance of the “space bar” response in the SIP to measure implicit attitudes accurately.

Gender SIP: A 4 (trial type: Males–Career, Male–Family, Female–Career, Female–Family) \times 2 (response: stationary, randomising) mixed ANOVA showed a significant main effect of trial type, $F(3, 186) = 87.97, p < .001, \eta^2 = .59$, a non-significant main effect of response, $F(1, 62) = 1.43, p = .237, \eta^2 = .02$, and the interaction between trial type and response was also non-significant, $F(3, 186) = 2.46, p = .064, \eta^2 = .04$. The interaction approached significance because all the trial types were comparable to each other, apart from the Male–Career trial type which showed a stronger pro-career bias in the stationary condition than in the condition where the location of “Yes” and “No” response option was randomised in each trial.

Since no significant difference was shown between the two conditions, they were combined and reported together. A one-way ANOVA conducted on the 4 separate trial types, the two compacted trial types and the overall D-SIP showed that all these scores were significantly above zero (Males-Career: $t(63) = 11.29, p < .001$, Males-Family: $t(63) = 5.03, p < .001$, Females-Career: $t(63) = 10.00, p < .001$, Female-Family: $t(63) = 10.60, p < .001$, compacted Males: $t(63) = 4.47, p < .001$), except for the compacted Female trial type score which did not differ significantly from zero, $t(63) = -1.56, p = .123$. The overall D-SIP score, $t(63) = 17.34, p < .001$, suggests that participants had a strong bias in relating males with careers and females with families (i.e., males-career/female-family bias; see Figure 4.7).

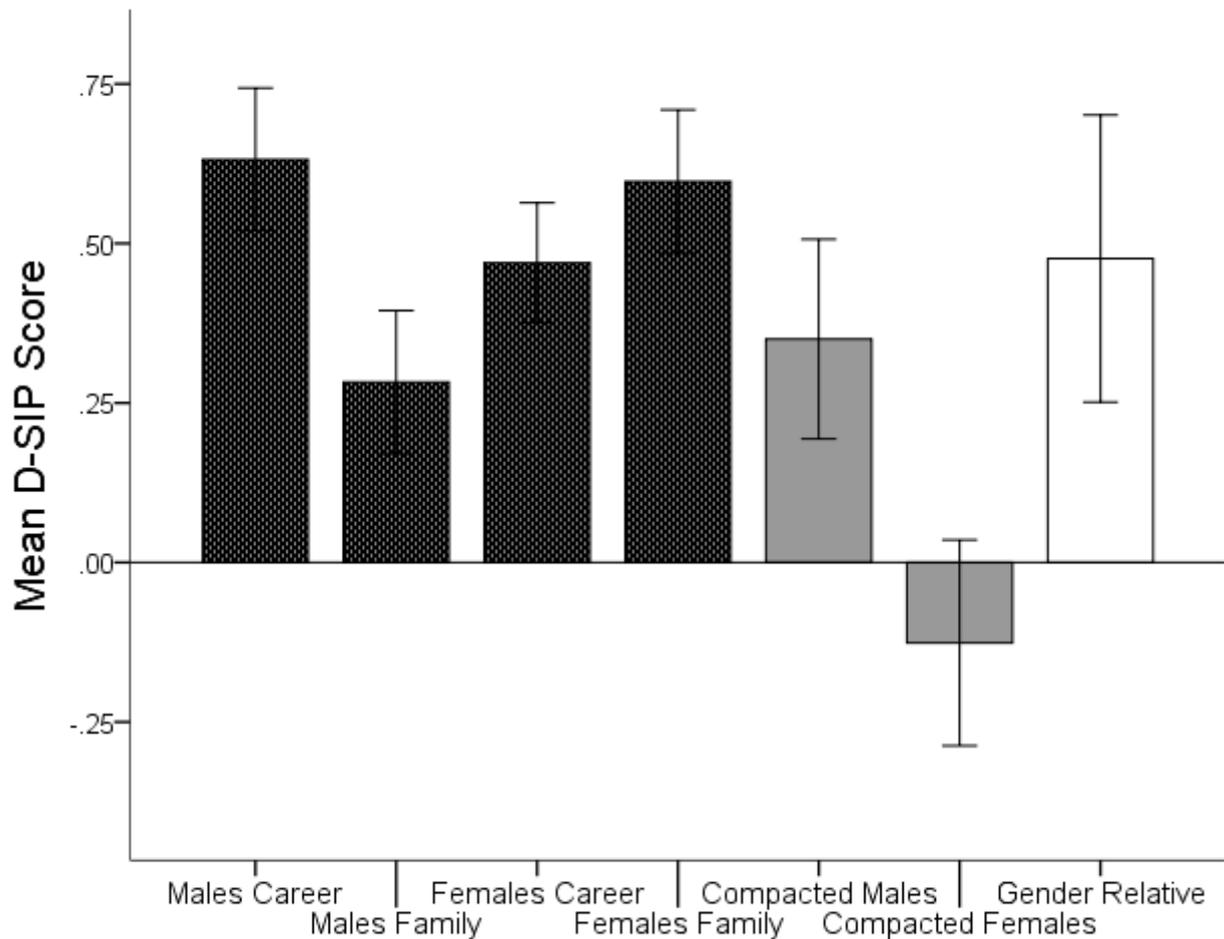


Figure 4.7: The four individual, the two compacted and the overall D-SIP for the Gender SIP. Error bars with 95% confidence intervals have been included.

Testing the internal reliability, the odd and even trials were positively correlated, but a significant result was not obtained, $r = .201$, $n = 64$, $p = .110$. Because the response option randomisation could have reduced the internal reliability in the SIP, separate split-half correlational analyses were carried out on the data that used the randomising and stationary response options. For the SIP that used stationary responses, the internal reliability approached significance, $r = .337$, $n = 32$, $p = .059$, while the internal reliability of the randomising response condition was substantially worse, $r = .049$, $n = 32$, $p = .792$. However, these two-correlation coefficients did not significantly differ from one another, ($p = .13$, one tailed). No significant correlation was found between the overall D-SIP implicit and explicit SDS scores, $r = -.122$, $n = 63$, $p = .341$, the compacted male scores, $r = -.094$, $n = 63$, $p = .296$, and compacted female

scores, $r = -.169$, $n = 63$, $p = .186$. The correlation between the unipolar relative implicit and explicit scores was also non-significant, $r = -.094$, $n = 126$, $p = .129$. Of note, the direction of all these effects were numerically negative, indicating a miss match between implicit and explicit stereotyping.

When we correlated the scores from Gender System Justification Scale with the compacted SIP trial types scores, the overall D-SIP score, the unipolar explicit SDS and the relative explicit SDS scores, we found that Gender Justification was significantly correlated with the relative explicit questionnaire, $r = .296$, $n = 63$, $p = .018$, and the compacted female trial type, $r = -.254$, $n = 63$, $p = .045$. All other relationships were non-significant, $r_s < .191$, $n = 63$, $p_s > .133$. This finding indicates that as gender system justification increases, participants more strongly implicitly associate females with family rather than careers and show a stronger male-career/female family relative explicit bias. Using a hierarchical logistical regression analysis with the relative explicit measure entered first into the model, we found that it proved to be a significant predictor of gender system justification, $B = .524$, $p = .018$ and accounted for 30% of the variance. When the compacted female trial type was added into the model, it significantly improved the model, $B = -.556$, $p = .020$, resulting in an R^2 change of 7.9% and the new model accounted for 41% of the variance.

These results suggest that using stationary “Yes” and “No” response options are likely more appropriate because this condition showed better internal reliability (albeit non-significantly) compared to the randomising response option condition. Most of the participants ($N > 25$) in the randomising condition also reported verbally to the experimenter that they found the constant switching much more challenging compared to the other SIPs that used stationary response options. Although no significant correlations between the implicit and explicit SDS scores were found in the current SIP, this was predicted based on the divergence between participants expressing egalitarian gender roles while implicitly supporting traditional

views. The compacted female trial types showed added predictive validity over the relative explicit SDS when predicting gender system justification which emphasises the importance of the added specificity that the SIP can offer.

Discussion

Using four SIPs we showed that the procedure has strong potential for capturing/measuring implicit beliefs. This can be seen through the appropriate internal reliability, even though participants completed four SIPs across different domains in one sitting. In addition, all the correlations were in the expected directions and the SIP correlated strongly with the unipolar relative explicit SDS for three out of the four domains. For the socially sensitive domain, the SIP correlated with a measure examining justification of gender inequality and the SIP had improved predictive validity over the explicit SDS.

No differences in implicit bias scores were found when the participants were shown either images or pictures as the category stimuli. We can infer from this finding that researchers could mix the use of visual and verbal category stimuli within the SIP's block of trials. Previous research using the IAT has shown that using words rather than images results in a stronger pro-flower/anti-insect bias (e.g., Carnevale, Fujita, Han, & Amit, 2015; Nosek, Banaji, & Greenwald, 2002) but this effect is modulated by using an image instead of a word as the category labels (e.g., similar or stronger pro-flower/anti-insect biases are shown in the IAT when people sort images into the correct image on the top of the screen than if they sort words into the correct word on the top of the screen; Meissner & Rothermund, 2015). This finding implies that mixing visual and verbal stimuli in the IAT is not ideal. The likely reason that using images or words in the SIP does not influence the findings is because participants must store only one of the four response rules in memory and cannot use stimuli on the screen to assist with the task.

Using one stimulus rather than four stimuli in each category did not show results in line with expectations (i.e., negative implicit biases towards criminals were weaker). However, the results showed more pronounced biases in the 4 stimulus condition. This finding is likely due to participants being reminded of more people that fit into the criminal category and hence, an additive effect might occur for implicit biases scores when more category stimuli are included in the SIP. A study systematically increasing the number of stimuli in each category would be useful to test if a linear relationship occurs between stronger implicit biases and more stimuli included in the SIP. A linear relationship would be problematic for the SIP because researchers could simply influence individuals' scores by increasing or decreasing the number of category exemplars.

A "space bar" response should always be used in the SIP to ensure participants relate/associate the category stimuli with the positive and negative words. However, we still have not determined the optimal number of "space bar" responses required to ensure participants do not ignore the category stimuli. Because "space bar" responses are never used in the analysis, determining the least amount necessary would be useful to shorten the overall time it takes to complete the SIP.

Finally, stationary response options should be used in the SIP because it makes it easier for participants to complete the task and improves the internal reliability. Overall, the Gender SIP showed the worst internal reliability compared with the other three SIPs. This reduced reliability could be due to the higher number of category and target stimuli used in the Gender SIP (i.e., 26 stimuli in the Gender SIP vs 20 stimuli in the Nature and Character SIP). The SIP has removed the PFB by using four separate response rules which use equal numbers of single positive and negative statements that must be stored in memory before each block of trials. Regardless of this enhanced procedure, it appears there is still another positivity/affirming bias in the SIP that overestimate the positivity of implicit biases scores towards objects (i.e.,

participants showed positive biases towards criminals and insects in the positive word trial types).

New research by O'Shea, Brown, and Watson (2017) has shown that participants are faster to sort affirming (e.g., "Yes") than negating (e.g., "No") stimuli (see Chapter 5). Furthermore, they showed that participants were more likely to associate affirming words with positive words and negating words with negative words as well as being overall faster when sorting affirming and positive words than negating and negative words. These findings would account for why participants are faster to press "Yes" than "No" when relating insects/criminals with positive words. The next study develops a method to overcome these response biases.

Study 2

Study 2 assessed implicit and explicit biases towards: (1) novel/non-words (Non-Word SIP), (2) flowers and insects (Nature SIP), and (3) carers and criminals (Character SIP). The findings from Study 1 indicate there is an affirming/positivity bias in SIP which appears to lead to overestimates of the positivity of implicit biases towards objects. One potential method of overcoming the SIP's affirming/positivity bias is as follows. One would expect participants to have neutral attitudes to novel/non-words. If neutral implicit attitudes are not detected to such stimuli, then it could be assumed that the detection of any apparently non-neutral attitude is most likely reflecting a (affirming/positivity) response bias. Therefore, this information from the Non-Word SIP can be used to correct for the response biases in other SIP tasks (i.e., Nature and Character SIP) by correcting for an individual's data which accounts for biases towards novel/non-words. Therefore, for the Non-Word SIP each of the four separate trial types are corrected/normalised until they show neutral attitudes and these correction metrics are then used to correct/normalised the results in each of the four trial types in the Nature and Character SIP.

After this correction, we expect the absolute/unipolar implicit scores to show positive biases towards flowers and carers on both the positive and negative target trials, while negative biases will be observed for insects and criminals on both the positive and negative target trials. Similarly, on the compacted trials, pro-flower, pro-carer, anti-insect and anti-criminal biases are expected. For the relative/bipolar implicit and explicit scores, pro-flower/anti-insect biases and pro-carer/anti-criminal biases are predicted. Internal reliability is expected to be in line with the Nature and Character SIP from Study 1 due to participants completing three SIPs in one session. Correlations are also expected to be similar to the Nature and Character SIP in study 1 (i.e., they will positively correlate). However, significant correlations are not expected between the Non-Words SIP and the explicit self-report due to participants mainly rating the non-words as neutral.

A final aim of this study is to provide a preliminary test of whether there are practice/experience effects after participants complete two SIPs (weaker biases are shown in the IAT after participants complete it once or twice; Greenwald et al., 2003). Study 1 could not test practice/experience effects because each participant completed the four SIPs in a random order. In the current experiment, in contrast, half the participants completed the Nature SIP first and the other half completed the Nature SIP last (3rd SIP). This design allows preliminary tests of whether weaker implicit scores are shown on the last SIP completed compared to the first SIP.

Method

Participants: 28 participants completed the current study (15 females). The mean age of the sample was 22.3 ($SD = .51$) with mostly Chinese participants ($N = 16$) and the remaining being white British/European. Participants were sourced using an online recruitment platform available through the University of Warwick. Due to this method of recruitment, all participants

were students from various levels and educational degrees at the University. Each participant was paid £5 for approximately 40 minutes of their time.

Materials

Demographic Information: Each participant's age, gender, nationality and ethnicity was gathered using a paper and pen questionnaire (see Appendix 4).

SIP: The SIPs used here were designed in the same way as those reported in Study 1. For example, the same number of practice (4) and test blocks (16) had to be completed and each block was composed of the same "Space", "Yes" and "No" response options. The same criteria as the SIP in Study 1 were also used; average of $\geq 80\%$ accuracy and $\leq 1,800$ ms response latency must be met across the practice blocks to continue to the test blocks. Study 2 used all the same Nature stimuli as in study 1 but for the Character SIP, the category stimuli only included doctor, nurse, murderer and rapist. In the Non-Words SIP, Cug and Vek were used as the Non-Words for half of the participants (time point 1), while the other half used non-words with "E/e" (Fykes, Beith, Chegn, Glure, Elths, Yimed) and "O/o" (Ointh, Tholc, Dofth, Shrox, Volph, Squov) (time point 2). The positive and negative words used in both non-word conditions were: Good, Happy, Positive, Freedom, Care, Lucky, Problems, Died, Negative, Hated, Bad and Sick.

Explicit Questionnaire: Using a paper questionnaire (see Appendix 4), participants completed a feeling thermometer which ranged from 0 (very cold or unfavourable feeling) to 100 (very warm or favourable feeling) for how they felt towards carers, criminals, insects, flowers, Cug and Vek (time 1), words with "E/e" and words with "O/o" (time 2). Scores on all these feeling thermometers were recoded into a -4 to +4 format with higher scores indicating more positive attitudes. Each thermometer was used as an absolute/unipolar result. In addition, three relative scores were created for each individual: (1) scores from the criminal thermometer were subtracted from the carer thermometer scores, (2) scores from the insect thermometer

were subtracted from the flower thermometer scores, (3) scores from the Vek/“O” thermometer were subtracted from the Cug/“E” thermometer scores.

Design and Procedure

Three separate 4 (SIP trial type: Category 1 - Positive, Category 1 - Negative, Category 2 - Positive, Category 2 - Negative) \times 2 (time: point 1, point 2) mixed designs were used to analyse data from (1) the Non-Word SIP, (2) the Nature SIP (3) and the Character SIP. The SIP trial type was always a within-subject factor and time was a between-subject factor. For time point 1 the Character SIP was the first task that each participant completed followed by the Non-Word SIP that used Cug and Vek. Lastly, each participant completed the Nature SIP. For time point 2, the position of the Character SIP and Nature SIP was reversed so that the Character SIP was always the last task participants completed and the Nature SIP was always the first. The Non-Word SIP was again completed between the Nature and Character SIP, however, this time, participants had to respond to the “E/e” and “O/o” non-words instead of Cug and Vek.

Participants completed time point 1, individually in a small, well-lit room. Time point 2 was carried out a month later using the same room. Each SIP took approximately 11-12 minutes to complete²⁴. When all three SIPs were completed, participants completed the demographic information and the explicit questionnaire. Following all these tasks, participants were thanked, debriefed and compensated for their time.

²⁴ The amount of time to complete a SIP task could be greatly reduced by having 8 or perhaps even 4 test blocks rather than 16. However, the number of trials should be increased if only 4 or 8 test blocks are used.

Results

Similar analytic procedures to Study 1 were used. Initially, the results from the Non-Words SIP are presented using the SIP standard analytic methods. Following this presentation, the four separate trial types in the Non-Words SIP are transformed to neutral to control for any response biases. These correction metrics from the Non-Word SIP are then used to correct for each individual's response biases on each of the four trial types on the Nature and Character SIP. The Non-Word SIP and the corrected Nature and Character SIP's results will be reported using the same format as Study 1.

Non-Word SIP: The 4 (trial type: Non-Word 1–Positive, Non-Word 1–Negative, Non-Word 2–Positive, Non-Word 2 – Negative)²⁵ × 2 (time: Cug & Vek, words with “E/e” and “O/o”) mixed design showed a significant main effect of trial type, $F(3, 78) = 47.26, p < .001, \eta^2 = .65$, a non-significant main effect of time, $F(1, 26) = .86, p = .36, \eta^2 = .03$ and no significant interaction between trial type and time, $F(3, 78) = .44, p = .73, \eta^2 = .02$. One-sample t-tests showed that the score from the Non-Words 1–Positive trials, $t(27) = 8.60, p < .001$, Non-Words 2–Positive trials, $t(27) = 11.09, p < .001$ were a significantly above zero (i.e., participants were faster to press “Yes” than to press “No” to any non-word related with positive words). Non-Word 1–Negative scores, $t(27) = -2.24, p = .033$, and the Non-Words 2– Negative scores, $t(27) = -4.08, p < .001$, were both significantly below zero (i.e., participants were faster to press “Yes” than to press “No” to any non-word related with negative words).

The compacted Non-Words 1, $t(27) = 4.32, p < .001$, and compacted Non-Words 2 trial type scores, $t(27) = 3.84, p = .001$, were both significantly above zero. This finding indicates that overall participants were the fastest at pressing “Yes” to positive words. The relative score was neither significantly above or below zero, $t(27) = 0.86, p = .40$. This suggests that participants had a neutral relative implicit attitude towards the non-word stimuli but the

²⁵ Non-Words 1 = Cug/“E/e” and Non-Word 2 = Vek/“O/o”.

apparent positive and negative biases on the separate trial types are likely due to an affirming bias (see Figure 4.8). The positive biases on the two compacted trial types indicate that there is also an overall positivity bias when using the SIP.

The split-half internal reliability between odd and even trials was significant, $r = .378$, $n = 28$, $p = .048$. The correlation between the relative implicit and explicit scores was not significant, $r = .209$, $n = 28$, $p = .286$. Both the compacted absolute Non-Words 1, $r = .330$, $n = 28$, $p = .087$, and Non-Words 2, $r = .093$, $n = 28$, $p = .639$, implicit and explicit measure scores were not significantly correlated but again they were in the expected direction. The correlation between the unipolar relative implicit and explicit scores was also non-significant, $r = .202$, $n = 56$, $p = .136$, as was the correlation in cut 1, $r = .076$, $n = 28$, $p = .700$, but their directionality was in the expected direction. For cut 2, a significant correlation was shown, $r = .422$, $n = 28$, $p = .025$.

There was an overall response bias with participants being faster at pressing “Yes” rather than “No” to both positive and negative words in the Non-Words SIP, even though neutral biases would have been expected towards non-words (i.e., no difference in pressing “Yes” or “No” to non-words). This finding indicates a clear affirming bias when the SIP is used. To remove this affirming bias, we averaged the Non-Word 1-Positive and Non-Word 2-Positive scores and subtracted this number from the scores in the Non-Word 1-Positive and Non-Word 2-Positive trial types. Likewise, we averaged the Non-Word 1-Negative and Non-Word 2-Negative scores and subtracted this number from the scores in the Non-Word 1-Negative and Non-Word 2-Negative trial types. These two averaged scores from (1) the two positive and (2) the two negative Non-Word trial types were used to transform/normalise the Nature and Character SIP biases. Importantly, rather than this correction being done at a group level, each individual’s non-word response bias scores were used to correct for their response biases in the Nature and Character SIP. As seen in Figure 4.9 the response bias corrections

resulted in neutral biases towards non-words on all the trial types, including the compacted trial types and the overall D-SIP.

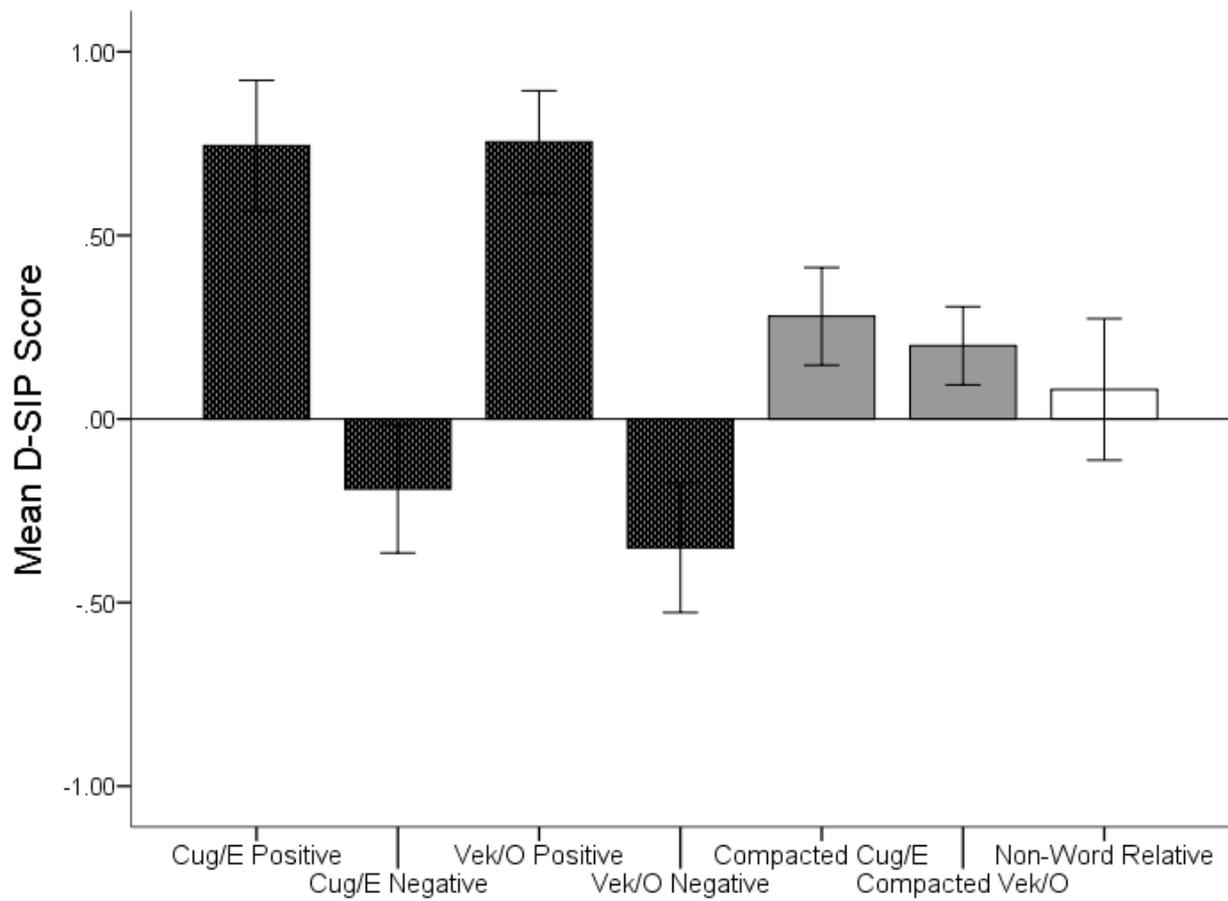


Figure 4.8: The uncorrected Non-Words SIP. Error bars with 95% confidence intervals have been included.

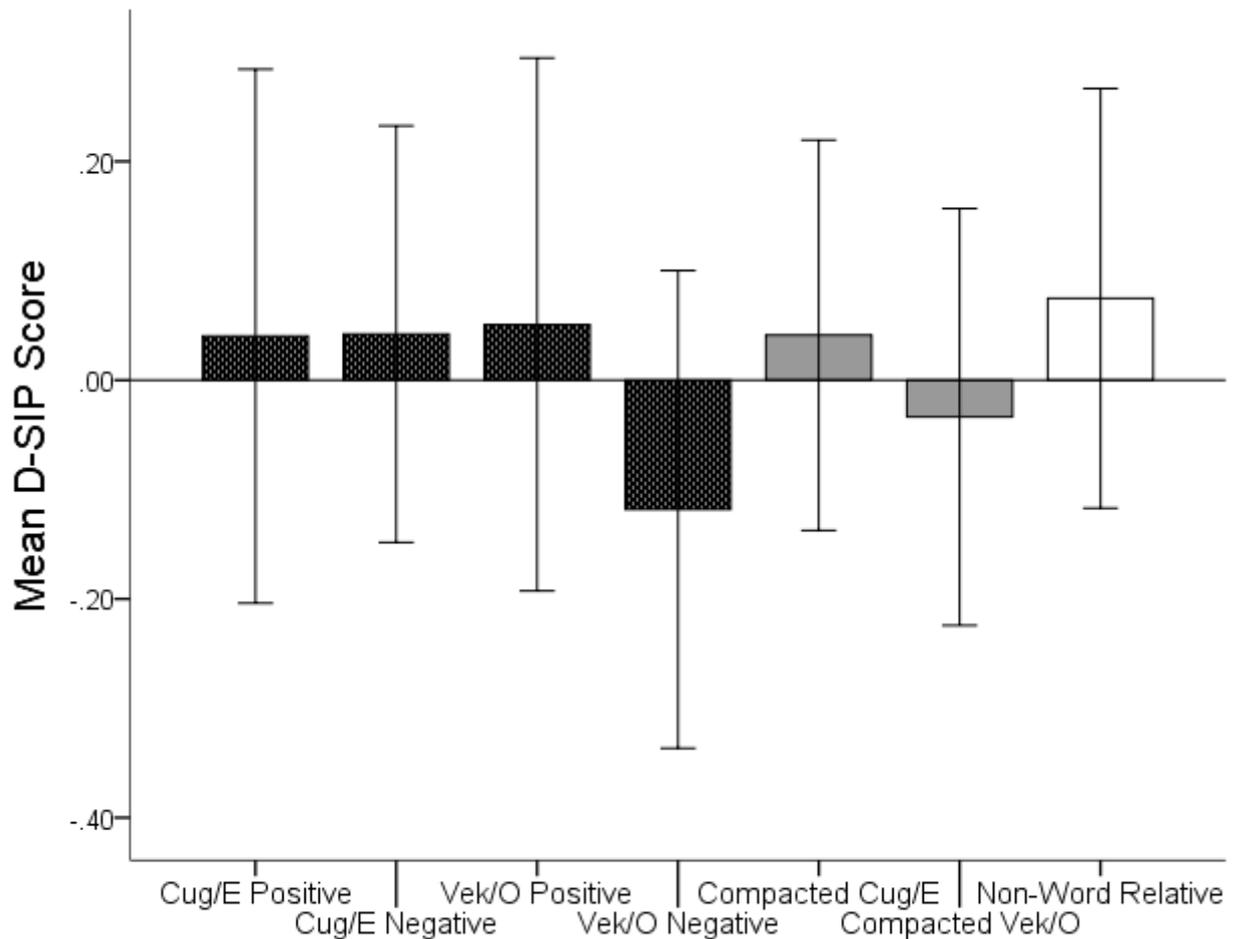


Figure 4.9: The corrected Non-Words SIP. Error bars with 95% confidence intervals have been included.

Nature SIP: Initially, the data from each individual from the four absolute trial types were corrected using the same scores that were used to correct the data in the Non-Words SIP. A 4 (trial type: Flower–Positive, Flower–Negative, Insect–Positive, Insect–Negative) \times 2 (time: point 1 = Nature SIP competed last, point 2 = Nature SIP completed first) mixed design revealed a significant main effect of trial type, $F(3, 78) = 7.66, p < .001, \eta p^2 = .23$, a non-significant main effect of time, $F(1, 26) = .33, p = .571, \eta p^2 = .01$ and no significant interaction between trial type and time, $F(3, 78) = 1.86, p = .170, \eta p^2 = .07$. Similar results were also found on the uncorrected data and therefore will not be reported. Importantly, the uncorrected scores from the four separate trial types and compacted trial types were similar to Study 1.

For the corrected results, one-sample *t*-tests showed that the scores on the Flower–Negative, $t(27) = 2.30, p = .029$, and the compacted Flower trial type, $t(27) = 2.27, p = .032$, were significantly above zero. Both the Insect–Positive, $t(27) = -2.85, p = .008$, and the Insect–Negative, $t(27) = -3.21, p = .003$, including the compacted Insect trial type scores, $t(27) = -5.37, p < .001$, were significantly below zero. The Flower–Positive trial type score did not differ significantly from zero, $t(27) = .26, p = .800$. The relative D-SIP score showed a strong pro-flower/anti-insect bias, $t(27) = 6.33, p < .001$ (see Figure 4.10).

The split-half internal reliability between odd and even trials, was in the predicted direction but was not significantly correlated, $r = .240, n = 28, p = .218$. It should be noted that when a Spearman–Brown correlation was used, a significant correlation was shown, $r = .412, n = 28, p = .030$ ²⁶. The correlation between the relative implicit and explicit scores was in the predicted direction but it was not significant, $r = .255, n = 28, p = .190$. Both the compacted absolute Flower, $r = .260, n = 28, p = .181$, and compacted absolute Insect, $r = .200, n = 28, p = .308$, implicit and explicit scores did not significantly correlate but again they were in the expected direction. The correlation between the unipolar relative implicit and explicit scores was significant, $r = .525, n = 56, p < .001$, as were the correlations in cut 1, $r = .589, n = 28, p = .001$, and cut 2, $r = .462, n = 28, p = .013$.

In summary, the correction to remove the affirming biases in the SIP seem to be effective. This effectiveness is especially apparent because strong negative attitudes were seen towards insects on the two separate trial types, including the compacted Insect trial type. A positive bias was observed on the compacted Flower and the Flower–Negative trial type but the neutral finding on the Flower–Positive trial type might suggest that the response bias correction

²⁶ Throughout, Pearson’s correlation was used and generally this analysis showed comparable results to Spearman–Brown’s correlational analysis.

may be over compensating on this trial type. Promisingly, the relative unipolar implicit and explicit measures were moderately-to-strongly correlated and acceptable internal reliability was shown when Spearman-Brown's correlation was used. Finally, previously completing two other versions of the SIP (Non-Words and Character SIP) did not influence or reduce participants' biases towards flowers and insects compared with those who initially completed the Nature SIP.

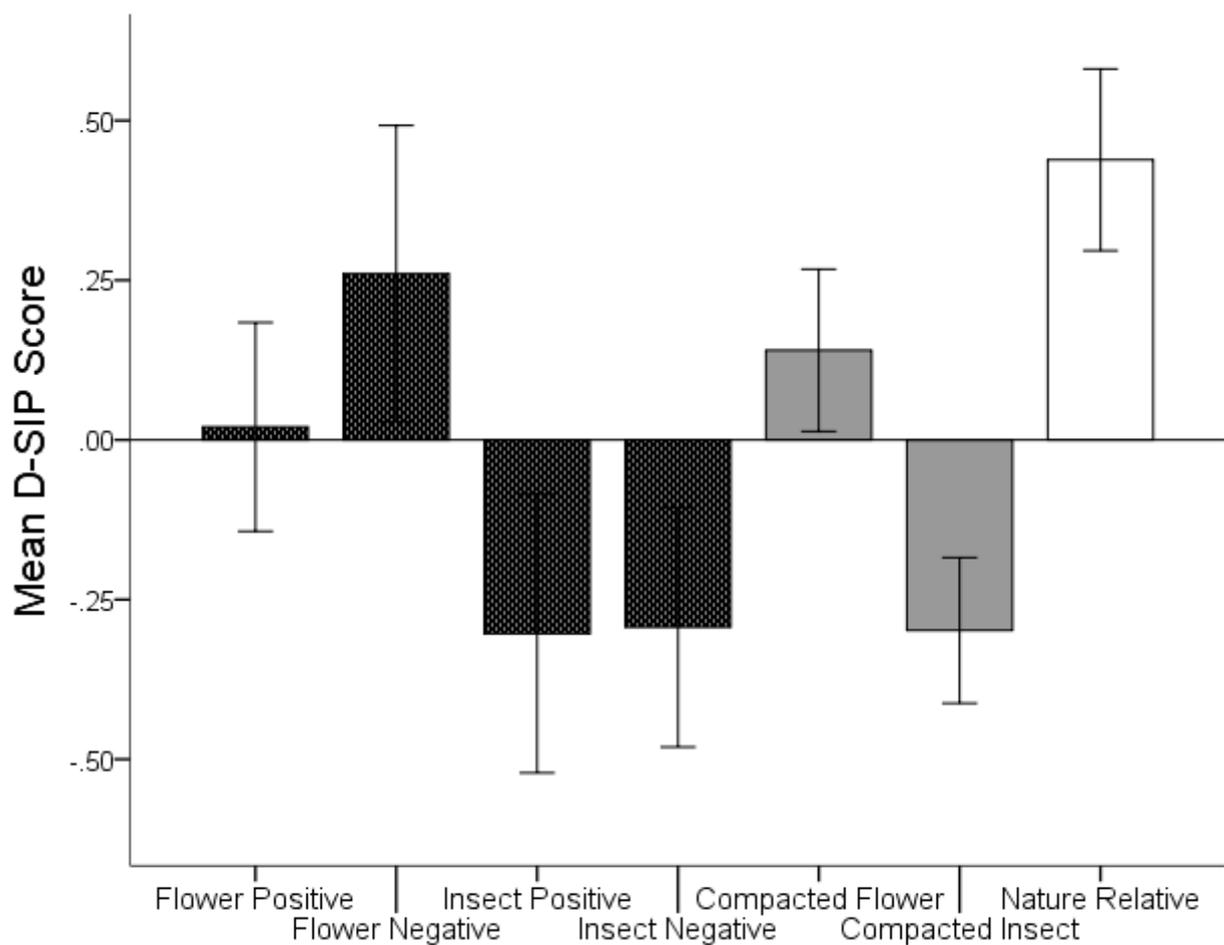


Figure 4.10: The corrected Nature SIP results. Error bars with 95% confidence intervals have been included.

Character SIP: After the correction was applied to the Character SIP to remove the affirming bias in the SIP, a 4 (trial type: Carer-Positive, Carer-Negative, Criminal-Positive,

Criminal–Negative) \times 2 (time: point 1 = Character SIP competed first, point 2 = Character SIP completed last) mixed design revealed a significant main effect of trial type, $F(3, 78) = 10.72$, $p < .001$, $\eta^2 = .29$, a non-significant main effect of time, $F(1, 26) = .77$, $p = .390$, $\eta^2 = .03$ and no significant interaction between trial type and time, $F(3, 78) = .73$, $p = .501$, $\eta^2 = .03$. Similar findings were also shown on the uncorrected data but the scores on the four separate, including the compacted trial types were similar to those obtained in Study 1.

One-sample t-tests showed that the scores on Carer–Negative, $t(27) = 2.54$, $p = .017$, and the compacted Carer trial type, $t(27) = 2.71$, $p = .011$, were significantly above zero. Criminal–Positive, $t(27) = -5.42$, $p < .001$, and the compacted Criminal trial type scores, $t(27) = -5.11$, $p < .001$, were significantly below zero. The Criminal–Negative trial type score was below zero but not significantly, $t(27) = -1.82$, $p = .080$. The Carer-Positive trial type did not significantly differ from zero, $t(27) = .43$, $p = .674$. Finally, the relative D-SIP score was a significantly above zero, $t(27) = 6.935$, $p < .001$, consistent with a strong pro-carer/anti-criminal bias (see Figure 4.11).

The split-half internal reliability between odd and even trials proved to be significant, $r = .382$, $n = 28$, $p = .045$. The correlation between the relative implicit and explicit scores was not significant but the correlation was in the predicted direction, $r = .122$, $n = 28$, $p = .537$. Both the compacted absolute Carer, $r = .231$, $n = 28$, $p = .237$, and compacted absolute Criminal, $r = .027$, $n = 28$, $p = .893$, implicit and explicit scores were not significantly correlated but again the correlation was in the expected direction. The correlation between the unipolar relative implicit and explicit scores was significant, $r = .558$, $n = 56$, $p < .001$, as were the correlations in cut 1, $r = .479$, $n = 28$, $p = .010$, and cut 2, $r = .640$, $n = 28$, $p < .001$.

Again, the affirming biases in the SIP in Study 1 seem to be removed after an individual-level correction is done based on the non-word data. The expected negative biases are observed on the absolute Criminal trial types and a strong negative bias is shown on the

compacted Criminal trial type. Positive biases are also observed towards Carers as seen on the compacted Carer trial type. The neutral finding on the Carer-Positive trial type might again be suggesting that the response bias correction on this trial type may be too pronounced. Strong correlations were observed between the relative unipolar implicit and explicit measures and acceptable internal reliability was shown. Similarly, to the Nature SIP above, these results also indicate that previously completing other versions of the SIP does not change implicit bias scores.

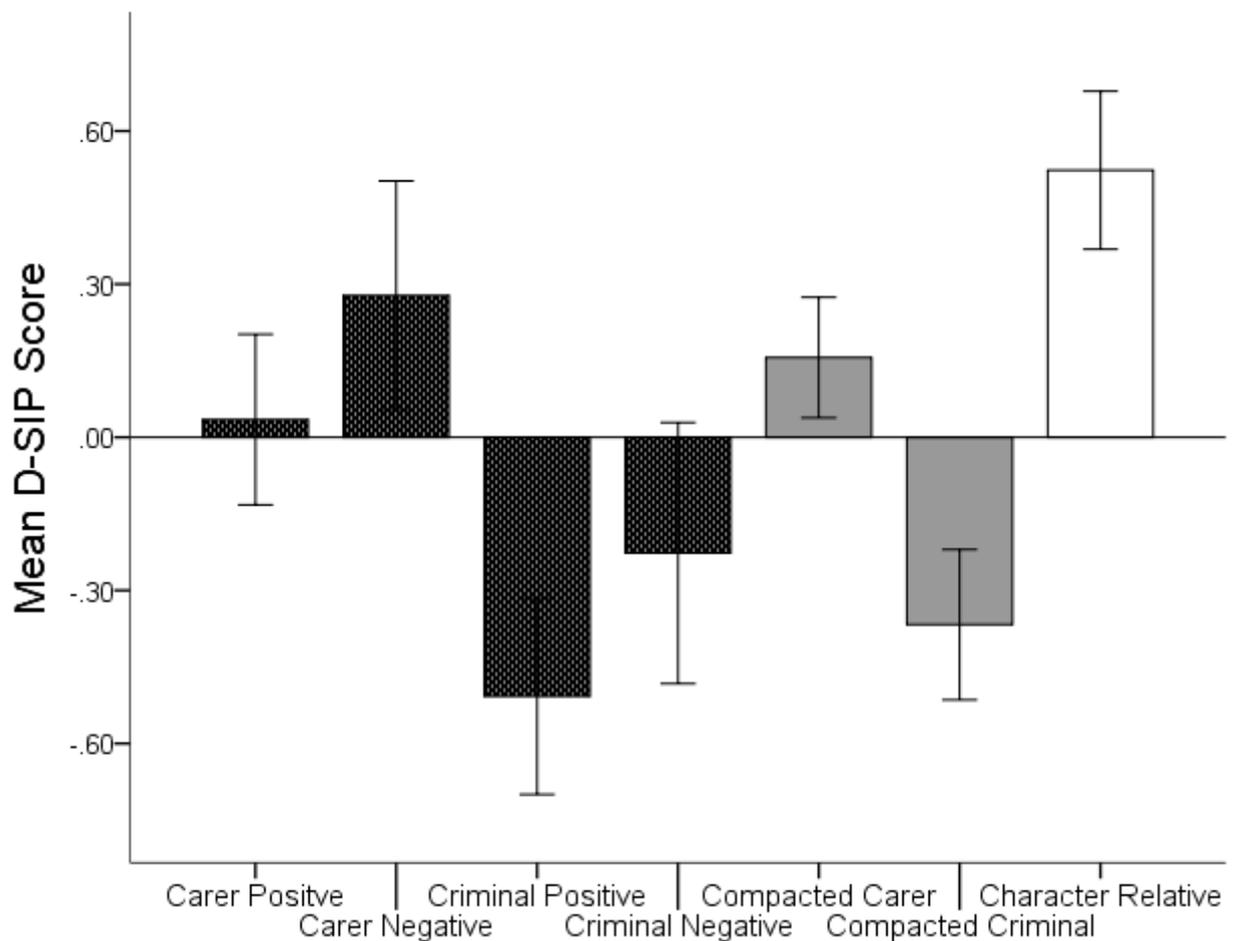


Figure 4.11: The corrected Character SIP results. Error bars with 95% confidence intervals have been included.

Discussion

The aim of the current study was to test the feasibility of measuring affirming response biases towards novel/non-words when using the SIP and controlling for these biases in SIPs that measure beliefs in different domains. It is clear when the response bias transformations are applied to the Nature and Character SIP, findings were in line with expectations (i.e., positive biases towards flowers and carers and negative biases towards insects and criminals were obtained). These findings were especially strong on the compacted trial types, with extremely strong negative biases being shown towards insects and criminals, while positive biases were observed for flowers and carers. The correction may be overcompensating on the Flower-Positive and Carer-Positive separate trial types because neutral attitudes were shown. However, this finding could also be indicating that what drives implicit bias are strong negative biases instead of strong positive biases towards objects.

A limitation of this study is that the Non-Words SIP was carried out after one SIP (either the Nature or Character SIP) had been completed and perhaps weaker or stronger affirming biases would have been shown if the task had been carried out first. A disadvantage of the current method to reduce the response biases is that it took a whole SIP task (10-12 minutes) to determine an individual's affirming biases. An alternative method to overcome the affirming biases in the SIP that might take less time would be to initially measure the speed at which an individual presses "Yes" to both the positive and negative target words, as well as their speed in pressing "No" to the same positive and negative words. The latency difference of a "Yes" vs. "No" response to positive and negative words could then be determined and controlled in a subsequent SIP. Additionally, using this method would reduce the likelihood that participants have a positive or negative implicit bias towards the neutral/novel/non-word category stimuli. Finally, ascertaining the minimum number of trials needed to determine the response biases in

the SIP would be valuable to shorten the amount of time required for participants to complete the task.

Again, this study shows that the scores on the SIP's absolute compacted trial types correlate (albeit non-significantly) with the explicit absolute SDS in the expected directions. These expected correlations were also observed for the relative SIP and explicit measures but the SIP's unipolar relative results showed moderately strong correlations with the unipolar explicit relative results which suggests that the SIP is accurately measuring an individual's implicit beliefs. The internal reliability was again appropriate for an implicit measure. This study also revealed preliminary evidence indicating that having more experience or practice with the SIP does not reduce the person's overall score. For example, this study showed that the participants that completed the Nature SIP and Character SIP first (position 1) had similar implicit bias scores to participants that completed these tasks following two other SIPs (position 3). If the SIP was not limited by practice/experience effects it would give it a major advantage over the IAT, which has strong practice effects (Greenwald et al., 2003). Study 3 addresses practice/experience in the SIP in more detail.

Study 3

Nosek et al., (2002) speculated and provided preliminary evidence that including multiple IAT submissions from a single respondent in an analysis of data collected through Project Implicit over a four year period did not greatly influence IAT results apart from a reduction in extreme scores from respondents who had taken one or more IATs (see also Greenwald & Nosek, 2001). A more detailed analysis by Greenwald et al., in 2003 examined whether prior experience with the IAT reduced IAT scores. When participants complete the IAT online through the Project Implicit website, they can optionally respond to the following question: "How many IATs have you previously performed?" There are five response options: 0, 1, 2, 3-5 and 6 or more. Greenwald et al., (2003) reported that there was a significant

reduction in the of IAT scores after one previous use and little or no further reduction after two or more previous uses.

Unpublished research by O'Shea, De Houwer, Ratliff, Brown and Watson (2017) directly challenges this finding, using a larger sample size, by showing that an individual's IAT score consistently reduces with more prior experience as far as 6+ exposures. In addition, they confirmed experimentally that both the IAT and the SC-IAT have a strong practice effect limitation. The practical limitations of prior experience with the IAT and the SC-IAT mean that scores from novices cannot be compared directly with those of non-novices. Likewise, post-tests cannot be compared directly with pre-tests and therefore, a control condition is needed to compare change (see Lai et al., 2016, for a recent example). Requiring a control condition is clearly problematic in clinical settings because it would be ethically dubious to use one when dealing with vulnerable patients that need urgent treatment.

The current study tests whether the SIP has similar practice effect limitations as the IAT and SC-IAT. Preliminary evidence from Study 2 indicates that D-SIP scores does not reduce after two exposures. Therefore, it is hypothesised that the SIP does not have a practice effect limitation due to its novel procedural set up. To test this hypothesis, participants were required to complete three Flower-Insect (Nature) SIPs in one sitting. As shown in the Character SIP in Study 1, including more stimuli increases the overall D-SIP score. Therefore, it is also possible that changing the stimuli throughout each SIP exposure would prevent a reduction or even increase the overall D-SIP score.

To address this possibility, we included one condition where participants completed the SIP with the same nature and target stimuli throughout, and another where all the stimuli changed on each SIP exposure. We remain agnostic regarding the outcome of the stimuli remaining the same or being different on each exposure. In this study, we did not measure any affirming biases when participants completed the SIP. Consequently, similar findings to the

Nature SIP in Study 1 were expected (i.e., positive attitudes towards flowers and neutral attitudes towards insects on the compacted trial types).

Method

Participants: 84 participants (53 females) took part in the study. The ethnicity of the sample was mainly Asian ($N = 55$) and White ($N = 22$). 32 participants were from Malaysia, 22 were British and the remaining were from a diverse set of countries. The mean age of the sample was 21.50 ($SD = 5.50$). Each participant was paid £2.00 to complete the experiment which took approximately 20 minutes.

Materials

Demographic Information: Each participant's age, gender, nationality and ethnicity were gathered using the same paper and pen questionnaire as Study 2 (see Appendix 4).

SIP: The SIPs used here were designed using the same procedure as in Study 1 and 2. The only difference was that each Nature SIP (A-C) was composed of only 8 test blocks rather than 16. There were three versions of the Nature SIP and each had different words for flowers and insects, and positive and negative target words. See Table 4.2 for the stimuli used in each Nature SIP.

Explicit Questionnaire: Using the same paper questionnaire as Study 2 (see Appendix 4), participants completed the feeling thermometer for how they felt towards insects and flowers. These were the unipolar/ absolute items. A relative/bipolar score was also created from these thermometer scales by subtracting each participant's insect score from their flower score.

Table 4.1: The stimuli used in the three Nature SIPs (A-C).

Nature SIP A		Nature SIP B		Nature SIP C	
Category Labels					
Flower	Insect	Flower	Insect	Flower	Insect
Category Stimuli					
Clover	Fly	Aster	Caterpillar	Jasmine	Locust
Gladiola	Tarantula	Bluebell	Cricket	Shamrock	Tick
Lily	Bee	Iris	Dragonfly	Rose	Bedbug
Carnation	Mosquito	Magnolia	Earwig	Buttercup	Ant
Daffodil	Beetle	Orchid	Lice	Marigold	Cockroach
Sunflower	Grasshopper	Petunia	Maggot	Tulip	Flea
Dandelion	Hornet	Poppy	Bug	Peony	Housefly
Pansy	Wasp	Violet	Termite	Lilac	Moth
Target Labels					
Positive	Negative	Positive	Negative	Positive	Negative
Target Stimuli					
Joy	Agony	Splendid	Nasty	Adore	Abandoned
Wonderful	Terrible	Great	Failure	Delicious	Bleak
Pleasure	Hurt	Friend	Evil	Enjoy	Crushed
Happy	Tragic	Love	Horrible	Pleasing	Rotten
Superb	Disgusting	Peace	Disaster	Fabulous	Miserable
Brilliant	Destroy	Glorious	Painful	Healthy	Injure
Amazing	Awful	Laughter	Sad	Kind	Hostile
Fantastic	Brutal	Excellent	Barbaric	Joyful	Suffer

Design and procedure

The experiment used a 2 (SIP compacted trial type: Compacted Flower, Compacted Insect) \times 3 (repeat: time 1, time 2, time 3) \times 2 (stimuli: same, different) mixed design. SIP compacted trial type and repeat were within-subject factors and stimuli was a between-subject factor. Each participant completed the experiment individually in a small, well-lit room. To ensure the correct counterbalancing was used among the participants who repeated the Nature

SIP with the same stimuli and those who repeated it with different stimuli the following steps were taken. Participant 1 completed SIP A (see Table 4.2) three times in succession. Participant 2 completed SIP A, followed by SIP B and then SIP C. Participant 3 did SIP A three times, and Participant 4 completed SIP A, followed by SIP C and lastly, SIP B. Participants 5-8, completed a similar sequence except with SIP B as the initial SIP. Participants 9-12 begun with SIP C. This counterbalancing sequence (participants 1-12) was repeated 7 times for the remaining 72 participants. When participants had completed the Nature SIP three times, they were instructed on the screen to call the experimenter. Participants then completed demographic information followed by the explicit questionnaire. Finally, they were thanked, debriefed and compensated for their time.

Results

The same analytic procedures as those in Study 1 were used. For the sake of clarity, only the most relevant results are reported. Specifically, only the two compacted trial types rather than the 4 separate trial types are reported. The mixed design ANOVA is reported first. Null hypothesis significance testing (NHST) can only estimate if variables are statistically different from one another, while Bayes factors can estimate the likelihood that variables are statistically similar. Hence Bayes factors is useful to use in the current experiment, because it can test how much evidence there is that the SIP does not have a practice effect problem by estimating evidence in favour of the null hypothesis relative to the alternative hypothesis (Jarosz & Wiley, 2014). Bayes Factor 10 (BF_{10}) will be used. See Table 4.3 for a description of how to interpret Bayes Factor 10.

Table 4.3 adapted from *Wetzels et al., (2011)*.

Bayes Factor 10 (BF10)	Interpretation
> 100	Decisive evidence for HA
30 - 100	Very strong evidence for HA
10 - 30	Strong evidence for HA
3 - 10	Substantial evidence for HA
1 - 3	Anecdotal evidence for HA
1	No evidence
0.33 - 1	Anecdotal evidence for H0
0.10 - 0.33	Substantial evidence for H0
0.03- 0.10	Strong evidence for H0
0.01 - 0.03	Very strong evidence for H0
< 0.01	Decisive evidence for H0

Next, a one-sample t-test is conducted on the averaged compacted Flower and Insect trial types as well as the overall D-SIP across the three time points. The split half (odd and even) reliabilities is reported for each of the three Nature SIPs (A-C), including test-retest reliability results. For the test re-test SIP results, the participants were split into one group that completed the Nature SIP three times with the same stimuli and another group that completed it with different stimuli. This splitting was done because of the expectation that presenting the same stimuli to participants would show higher test re-test reliability due to the consistency of the items, while those presented with different stimuli across SIPs would show lower test-retest reliability. Finally, the correlations between the averaged compacted trial types, the overall D-SIP and the absolute and relative explicit measures are reported. Matching the method used in Study 1, the unipolar relative implicit and explicit correlations are reported, including the

appropriate cutting of the data to ensure each participant had only one implicit and explicit data point.

A 2 (SIP compacted trial type) \times 3 (repeat) \times 2 (stimuli) mixed ANOVA showed a significant main effect of SIP compacted trial type, $F(1, 82) = 46.16, p < .001, \eta^2 = .36$, with the compacted Flower trial type showing more positive attitudes ($M = .48, SD = .04$) than the compacted Insect trial type ($M = .07, SD = .04$). The main effect of repeat, $F(2, 164) = .53, p = .593, \eta^2 = .01$, and stimuli $F(1, 82) = .11, p = .745, \eta^2 = .00$, were not significant. All the interactions were non-significant, $F_s < .84, p_s > .365, \eta^2 < .01$. The estimated Bayes factor of repeat ($BF_{10} = .033$) suggested that the data strongly supported the null hypothesis (i.e., there is no practice effect in the SIP). For stimuli, the estimated Bayes factor ($BF_{10} = .130$) suggested that the data substantially supports the null hypothesis (i.e., there is no difference between using similar or different stimuli throughout the three Nature SIP time points). See Figure 4.12.

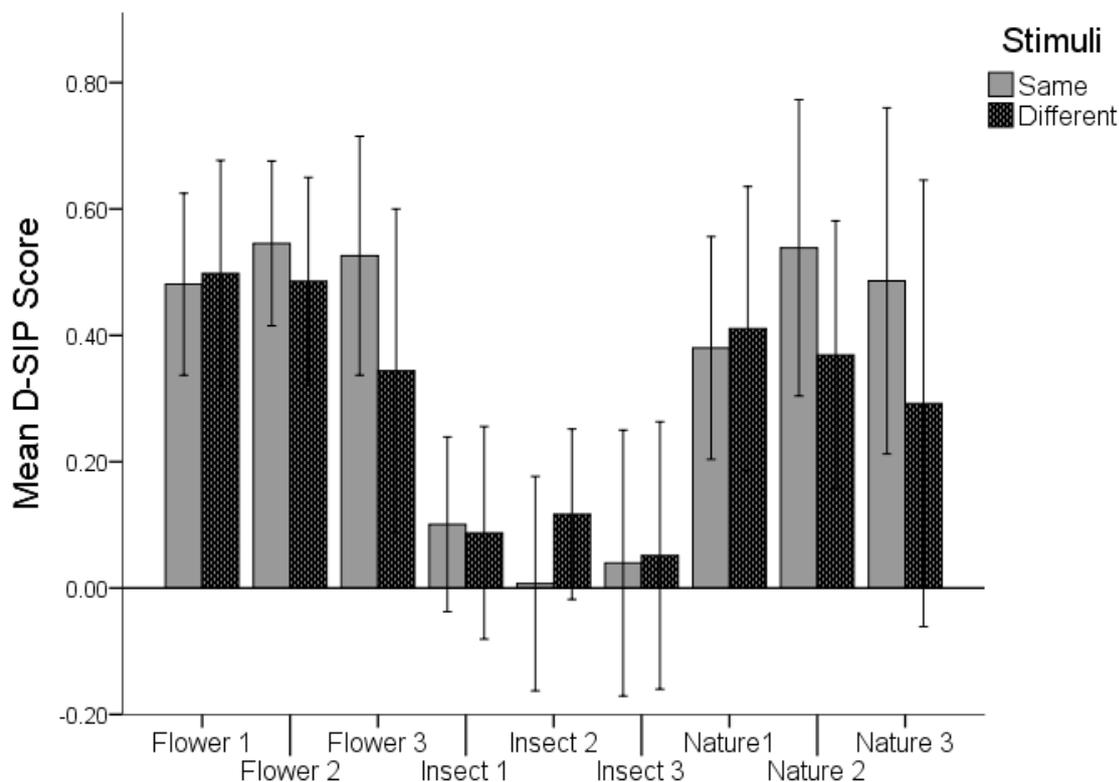


Figure 4.12: The compacted Flower and Insect trial types including the overall D-SIP across the three time points. Error bars with 95% confidence intervals have been included.

The one-sample t-test carried out on the averaged compacted Flower trial type scores was significantly above zero, $t(83) = 11.76, p < .001$, and the averaged compacted Insect trial type scores did not differ significantly from zero, $t(83) = 1.53, p = .129$. The averaged relative overall D-SIP score was significantly above zero, $t(83) = 6.80, p < .001$, consistent with a pro-flower/anti-insect bias. The split half odd and even reliabilities for the three-time points were as follows, Time 1: $r = .358, n = 84, p = .001$, Time 2: $r = .147, n = 84, p = .183$, and Time 3: $r = .096, n = 84, p = .383$. This finding suggests that the more time a participant spends carrying out the SIP, the weaker the internal reliability becomes. This reduction is likely due to fatigue and reduced concentration while completing the task over a relatively long period. As shown in Table 4.4, higher test-retest reliability was found in the group that was presented with the same stimuli throughout. Although Time 1 and Time 2 were not strongly related, the other time points were more in line with IAT test re-test findings (Nosek, et al., 2007), particularly for the group with consistent stimuli.

Table 4.4: Test re-test reliability for participants in the same or different stimuli condition.

Stimuli		Time 1	Time 2	Time 3
Same	Time 1	1	.088	.291†
	Time 2		1	.495**
	Time 3			1
Different	Time 1	1	.016	.152
	Time 2		1	.235
	Time 3			1

† $p < .065$, ** $p < .001$

The correlation between the relative/bipolar implicit and explicit scores was significant, $r = .244, n = 84, p = .024$. For both the (compacted) absolute flower, $r = .206, n = 84, p = .060$, and insect, $r = .067, n = 84, p = .543$, implicit and explicit measure scores did not significantly

correlate, but the correlations were in the predicted direction. The correlation between unipolar relative implicit and explicit scores was significant, $r = .434$, $n = 168$, $p < .001$, as were the correlations for cut 1, $r = .446$, $n = 84$, $p < .001$ and cut 2, $r = .423$, $n = 84$, $p < .001$.

Discussion

The aim of the current study was to test whether the SIP has practice effect limitations like the IAT. On the basis of findings from Study 2 we hypothesised that the SIP would not have a practice/experience effect problem and this prediction was confirmed. Similar implicit biases were shown if participants saw the same stimuli throughout the three SIPs or if the stimuli changed on each new exposure. Bayes Factors were used to confirm that both these effects were statistically comparable to each other, something NHST cannot do. However, the internal reliability of the SIP does reduce over time and the test-retest reliability was higher when the same stimuli were used. This finding suggests that giving participants breaks after each SIP or perhaps getting them to complete the SIP the following day may improve the internal reliability over multiple SIP exposures. Lastly, the expected positive correlations were found in the relative implicit and explicit results for both the bipolar and unipolar measures. The absolute implicit and explicit measures were also in the expected direction. These findings indicate that the SIP is correctly measuring an individual's relative bias towards flowers and insects.

General discussion

The SIP was developed to measure implicit beliefs, to be more user-friendly to complete and to overcome the positive framing bias (PFB) limitation in the IRAP. The SIP utilised the most beneficial aspects from the IRAP, such as the measurement of two separate biases (positive/negative) towards a single attitude object and the ability to measure propositional statements and not just associations. Other associative implicit measures, such as the SC-IAT and the GNAT can measure only one separate bias towards objects, while the most

popular associative implicit measure (IAT) can only compare one attitude object to another (i.e., Flowers vs. Insects). Therefore, absolute implicit measures, particularly the SIP, provide researchers with added specificity in determining the driving mechanism behind an implicit bias (i.e., whether participants have a neutral bias towards flowers and dislike insects, have a neutral bias towards insects and like flowers, or some combination of these).

Response biases inherent to the English language result in participants having an affirming bias (faster to press “Yes” than “No” to both positive and negative words; see O’Shea et al., 2017a), while carrying out the SIP. However, this affirming bias can be remedied easily by measuring response biases towards neutral/non-words and then using these metrics to correct for response biases in subsequent SIPs. Importantly, an individual’s response biases rather than the group’s response bias must be measured and controlled for because some individuals may have more extreme affirming biases and perhaps others will have a negating bias.

Across two different attitude domains with near universal bias, it was found (after the appropriate response biases transformations were made) that participants had strong negative implicit attitudes towards insects and criminals, and positive attitudes towards flowers and carers. Correlations were always in the expected direction and the unipolar implicit and explicit relative results showed moderately strong correlations. When measuring attitudes towards the self, correlations were observed between the implicit and explicit measures, again suggesting that the SIP is measuring constructs appropriately.

Finally, the Gender SIP showed that a strong bias for relating males with careers rather than families is the mechanism driving the males-career/female-family bias in the IAT. Although participants were faster to associate females with family rather than careers implicitly, the effect did not reach conventional levels of significance. In the Gender-SIP, implicit and explicit scores did not correlate and often the correlations showed negative

coefficients. However, the SIP did correlate with an explicit measure assessing gender system justification and it made a unique contribution when predicting these explicit gender biases.

The internal reliability of the SIP was also examined extensively. Except for the IAT and the SC-IAT, most other associative implicit measures of social cognition have poor reliability (Bosson et al., 2000; Karpinski & Steinman, 2006; Nosek, et al., 2007). The odd-even split half reliability in the SIP was acceptable (average $r = .403$, ranging from $r = .289$ to $r = .582$), showing better internal reliability than most other associative measures. Although the IAT and the SC-IAT generally have higher internal reliabilities, participants in the current experiment completed a number of SIPs in one sitting, and fatigue could have contributed to the lower reliability. Other propositional implicit measures such as the IRAP and RRT have reported similar levels of internal reliability (De Houwer et al., 2015; Golijani-Moghaddam et al., 2013). Thus, initial evidence indicates that the SIP has a sufficient level of internal reliability to be used as an individual difference tool that measures implicit social cognition.

A major advantage of using the SIP compared to the most popular implicit measures (IAT, SC-IAT) is that it is not limited by practice/experience effects, with the D-SIP scores remaining stable over time. This stability means that attitudes and stereotypes of participants who used the SIP many times can be compared with those who are completely new to the task. More practically, clinical or health psychologists can use the SIP to measure the success of a therapy at the individual level, without the need of a control condition. This practicality, as well as the added understanding of where changes in attitudes/stereotypes are occurring could initiate a new wave of individualised therapies aimed at tackling a specific bias in patients. Further research should be conducted to test if the methods that are successful at reducing implicit prejudice in the IAT (Lai et al., 2014) are also successful at reducing prejudice when measured with the SIP.

Simple is the word that is used for the “S” in the SIP acronym. As this word suggests, developing a user-friendly implicit measure was a central aim. None of the participants were removed from the analysis based on the D-SIP algorithm. In contrast, in the IRAP attrition rates of 20% or greater are not uncommon (Hughes et al., 2012). Furthermore, the SIP in the studies described above averaged 11-12 minutes to complete, while participants may need 20 minutes to complete the IRAP (De Houwer et al., 2015), because of the difficulty some participants have in meeting the criteria in the practice blocks. Importantly, the time to complete the SIP could be greatly reduced by using fewer test blocks throughout. With the inclusion of blocks measuring the affirming/positivity biases, a SIP that takes under 10 minutes to complete would likely be possible.

Automaticity has been argued to be a central feature for a measure to be described as implicit, and hence, capable of capturing implicit beliefs (De Houwer & Moors, 2012; De Houwer et al., 2009). For example, the average response latency in the Nature SIP (Study 2) was 1,054ms with accuracy averaging 94.8%. The other SIP experiments showed similar results. Response latencies in the SIP are similar if not faster than other implicit measures and the quick responses are likely due to participants having to store only one relational statement in memory. Therefore, the SIP certainly can be classified as an implicit measure on the basis of the speeded responses that are required and easily achieved throughout the task.

Recommendations for using the SIP

Based on the results reported here and other pilot studies conducted, some recommendations are given for researchers interested in using the SIP as a response latency measure for capturing individuals’ beliefs.

First, it would be advised that piloting should be initially conducted, to test if the intended sample can respond quickly and with a high level of accuracy to the stimuli used. In the studies reported here, as each participant completed several SIPs, we purposely left the

median response latency criteria (1,800ms) longer than what pilot studies suggested could easily be met (1,600ms), due to potential fatigue. Barnes-Holmes, et al., (2010) recommended using a 2,000ms latency criterion rather than 3,000ms because it showed more accurate results and improved the internal reliability in the IRAP. If participants are completing only one SIP in a single session, then a 1,600ms or perhaps even a 1,500ms latency criteria would be optimal. Ensuring the faster response latency criteria does not greatly increase the percentage of error rates would also be important.

Second, it is recommended that practice blocks always be included and test blocks should have at least 18 trials if not more. The effects that task switching (i.e., having 20 trial across 16 test blocks vs having 40 trials across 8 test blocks) can have on implicit bias scores and the internal reliability has never been tested systematically with any implicit measure. There is some evidence from other absolute implicit measures that having more trials per block improves the internal reliability. For example, Karpinski and Steinman, (2006) reported lower internal consistencies when 48 trials rather than 72 trials per block were used in the SC-IAT. The Single Target (ST)- IAT which is conceptually similar to the SC-IAT, had low internal consistency when 20 trials were used per block (Wigboldus, Holland, & van Knippenberg, 2005), but when 35 trials were used the reliability was much improved (Bluemke & Friese, 2008). Therefore, researchers should experiment with different block lengths, but with the same overall number of trials, to test the effects these changes have on implicit biases and reliability.

Third, when analysing the SIP data, it is recommend using the D-algorithm adapted from Greenwald et al., (2003). Numerous other implicit tasks have reported that the D-algorithm has an advantage over using raw latency scores (e.g., Nosek, Bar-Anan, Sriram, Axt, & Greenwald, 2014) because it generally increases the reliability, increases the correlation between the implicit and explicit measures and reduces the correlation between the implicit scores and the speed of responding (Greenwald et al.,2003; see Richetin, Costantini, Perugini,

& Schönbrodt, 2015, for potential improvements to D algorithm). When O'Shea et al., (2016) initially introduced the SIP, they advised that "Yes" responses should be compared with "Yes" responses and "No" responses compared with "No" across the test blocks. This method of analysis results in the same findings to the analytic method used above, but is disadvantaged by the fact that only a single rather than two response biases can be measured towards each category. Consequently, it is advised that researchers use the analytic strategy used throughout the three studies above.

Fourth, although "absolute" has been used throughout the manuscript to describe the SIP's ability to measure response biases towards separate categories, this label might be premature. All the studies conducted used a contrast/comparison category which may have influenced the results in a certain way. For example, previous research has suggested that using a different contrast category in the IAT (e.g., Karpinski, 2004; Robinson, Meier, Zetocha, & McCaul, 2005), and the IRAP (Hussey et al., 2016) can increase or decrease an individual's implicit bias. Testing the effects of using the SIP with a strongly negative, a strongly positive and a neutral contrasting category would be illuminating and beneficial. The contrast categories will undoubtedly influence the relative results, but it remains to be seen if the compacted SIP trial types are influenced by a contrasting category. Of note, it is even possible that there is no such thing as an absolute attitude because it has been argued that all attitudes require some type of comparative judgement (Festinger, 1954).

Finally, it is certainly possible to include sentences rather than single words in the SIP, to precisely measure the relationship between stimuli (e.g. "I am good" vs "I want to be good"). This method may allow for more control over the way that participants respond to the relationship between stimuli and therefore, allow for more control over the type of beliefs that scores on the SIP can detect (see De Houwer et al., 2015, for a similar argument). Furthermore, a personalised SIP (see Olson & Fazio, 2004) can be developed using the current methodology,

by simply replacing the category labels “positive” and “negative” with “I love” and “I hate”. A similar adaption would also be possible when creating an autobiographical SIP (see Agosta & Sartori, 2013; Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008). Therefore, the versatility of the SIP allows it to be used in various research domains within and outside psychology.

Conclusion

This paper is the first to validate the SIP as a tool for obtaining an implicit measure of beliefs. Through three studies covering several attitude domains and a stereotype domain, we demonstrated that the SIP correlates with explicit measures, provides increased specificity of where an individual’s implicit biases lie, has acceptable reliability and is not limited by practice/experience effects. Determining whether the SIP can predict real life behaviours and directly comparing it with other implicit measures would be a fruitful next step to test its validity, utility and employability. Through new measures of implicit social cognition, researchers can gain an enhanced understanding of an individual’s implicit and explicit social cognition. The SIP appears to be a valuable tool in this pursuit.

**Chapter 5: Some words are more like
pictures: Word type response biases in
reaction time tasks**

Abstract

Responding quickly to negative stimuli is evolutionarily advantageous. Generally, negative images are recognised and responded to faster than positive images. For words, the reverse is found: Positive words are normally recognised and responded to faster than negative words. Estes and Verges (2008) reported an exception. They showed that negative nouns are responded to more quickly than positive nouns in a valence judgement task (VJT). Crucially, nouns are inevitably the most concrete/imaginable words. Here we examine how the use of more abstract words (i.e., verbs/adjectives) influences reaction times in VJTs. Using secondary data and three experiments, we report that reaction times to negative images and nouns are faster than to positive images and nouns, but the reverse applies for verbs and adjectives. Other influential response biases are also discussed. The findings have substantial implications for the interpretation of results from various psychological tasks and can provide an explanation for previous counterintuitive findings in the literature.

Introduction

Quickly detecting and responding to images of negative/threatening stimuli such as snakes conveys many evolutionary advantages (Öhman, Flykt, & Esteves, 2001). Indeed, faster reaction times (RTs) occur when participants search for negative facial expressions (anger) than when they search for positive expressions (such as happiness: Ohman, Lundqvist, & Esteves, 2001). This difference might be because negative stimuli tend to capture (preferential engagement: Pratto & John, 1991) and hold (delayed disengagement: Fox, Russo, Bowles, & Dutton, 2001) attention more than positive images. However, studies in linguistics often find the opposite — people respond faster to positive words than to negative words (Kuperman, Estes, Brysbaert, & Warriner, 2014). This response bias has been found in lexical decision-making (Estes & Adelman, 2008), word recognition (Algom, Chajut, & Lev, 2004) and emotional Stroop tasks (Larsen, Mercer, & Balota, 2006).

One exception is found in Estes and Verges (2008). Using concrete nouns as stimuli, they found that participants responded faster to negative nouns than to positive nouns in a Valence Judgement Task (VJT), while the reverse occurred in a lexical decision-making task which was similar to previous findings (see also Nasrallah, Carmel, & Lavie, 2009). Thus, it appears at least that concrete words produce the same negative stimulus advantage as (naturally concrete) images when valence judgements are made. However, no research has considered whether comparable or different response biases will occur for verbs and adjectives in a VJT.

Nouns are typically more concrete and more easily imagined than are adjectives (Brysbaert, Warriner, & Kuperman, 2014; see also Carnaghi et al., 2008). We might, therefore, expect that nouns will be processed in a similar way to images, producing faster RTs for negative items in a VJT. In contrast, there is less evolutionary urgency to detect and respond to more abstract stimuli (typically conveyed by verbs and adjectives) because they are less likely to signal the presence of immediate environmental threats. VJTs are used in a host of

popular implicit measures such as the Implicit Association Test (IAT; Greenwald, Nosek, & Banaji, 2003), the Single Category (SC)-IAT (Karpinski & Steinman, 2006) and the Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001). Therefore, if nouns, verbs and adjectives influence valence-related RTs in different ways, the interpretation of the results drawn from such implicit tasks could be markedly different. More generally, the results will have an impact in many areas of research that solely present verbs, adjectives or nouns or their combination.

A secondary aim was to uncover further biases that can influence responding on other psychological tasks. A recent study (O'Shea et al., 2016) reported a Positive Framing Bias (PFB, Dodds et al., 2015; Matthews & Dylman, 2014) in a popular implicit RT task - the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010). This PFB occurs because participants prioritise positive over negative associations. O'Shea et al. (2016) described a new measure for estimating implicit attitudes (the Simple Implicit Procedure; SIP) that aimed to remove the PFB. Regardless, other positivity biases remained. For example, participants appeared biased towards making faster affirming responses ("Yes") compared with negating responses ("No") (see also Deutsch, Gawronski, & Strack, 2006). This affirming bias was especially pronounced when participants were associating any category (positive or negative) with positive words (O'Shea, Brown, & Watson, 2017, see also Mitchell, 2004; Proctor & Cho, 2006). We therefore examined whether participants were faster at sorting/associating affirming and positive words together and negating and negative words than the reverse associations.

Overview

The main aim of the present work was to determine the influence of word type (noun, verb, adjective) on valence judgement RTs. We also examined the effect of making affirming versus negating responses on RTs. In addition, we tested for response biases participants had in associating positive and negative words with affirming and negating words. Based on the

polarity correspondence principle (Proctor & Cho, 2006), we expected participants to more easily associate negating and negative words with each other and affirming and positive words with each other. We tested three specific hypotheses:

H1: RTs in a VJT will be faster for concrete negative stimuli (images and nouns) than for positive stimuli. For abstract stimuli (verbs and adjectives), the reverse will be found (Studies 1-3).

H2: RTs will be faster for affirming words than negating words (Studies 2-3).

H3: Participants' RTs will be faster when associating affirming and positive words, and negating and negative words than the reverse (Study 2).

Study 1

Using secondary data, Study 1 tested whether people judged the valences of negative concrete stimuli (images and nouns) and positive abstract stimuli (adjectives) faster than they judged the valences of their antonyms in the Black-White (Race) IAT (Greenwald et al., 2003) and the Flower-Insect (Nature) IAT. In the Nature IAT, for example, participants are successively presented with various pictures of flowers and insects and positive and negative words. They must sort these stimuli into the correct category label at the top of the screen. In one of the two critical blocks, participants have to press the E key on a computer keyboard if a positive word or a name of a flower appears and press the I key if a negative word or a name of an insect appears (congruent task). In the other critical block, participants have to press E if a positive word or a name of an insect is shown and they must press the I key if a negative word or a name of a flower is shown (incongruent task). The basic idea underlying the IAT is that participants will make faster and more accurate responses when those responses are congruent with their current beliefs than when they are not.

Method

Participants: For the race IAT, the sample consisted of 345 white participants, recruited through Reddit, who previously completed an online experiment (O’Shea, Brown, Watson, & Fincher, 2017). For the nature IAT, 88 participants recruited through Amazon Turk completed the task three times in succession (i.e., 1st, 2nd and 3rd Nature IAT, see Table 5.1 for number of trials in each block in the three Nature IATs) resulting in 264 usable scores in a study that examined practice effects (O’Shea, De Houwer, et al., 2017).

Materials: The race IAT used the same stimuli and procedure as the standard race IAT used by Project Implicit (Greenwald et al., 2003; Nosek, Smyth, et al., 2007). Participants sorted greyscale pictures of black and white individuals as well as positive and negative adjectives into the appropriate categories. The category labels used were “African American”, “European American”, “Good” and “Bad”. The Flower-Insect IAT used only words (no pictures) referring to flowers and insects (all nouns), in addition to positively and negatively valenced adjectives. The category labels in this task were “Flower”, “Insect”, “Pleasant” and “Unpleasant”. Appendix 5 includes the stimuli used in the current experiment.

Designs and Procedure: For the race IAT, a 2×2 repeated measures design was used with stimulus type (image, adjective) and valence (positive/non-threatening, negative/threatening) as within-subject factors. The nature IAT used another 2×2 repeated measures design with stimulus type (noun, adjective) and valence (positive/non-threatening, negative/threatening) as within-subject factors. The procedure was the same for both IATs, with domain-specific stimuli used for each task. After participants gave informed consent and completed some other tasks related to the source experiment, they read the instructions on how to complete the IAT. The instruction read: “You will be presented with a set of words to classify into groups. Classify items as quickly as you can while making as few mistakes as possible.

Going too slow or making too many mistakes will result in an uninterpretable score. Keep your index fingers on the E and I keys to enable rapid response”.

The category labels used in each block remained at the top of the screen throughout a block of trials and the target word or image appeared at the centre. If a correct response was given, the target stimulus disappeared and a new target stimulus appeared after 400ms. If an incorrect response was given, a red X appeared directly below the target stimulus and both remained until the correct response was given. Table 5.1 summarises the number of trials and stimuli in each of the seven blocks of trials. When participants finished the IAT, they continued with the remaining tasks within the study followed by an online debrief form.

Table 5.1: Number of trials and stimuli presented in the Race and Nature IAT blocks (Study 1).

Race & 1st Nature IAT			2nd & 3rd Nature IAT		
Block Number	Number of trials	Stimuli presented	Block Number	Number of trials	Stimuli presented
1	20	Image (Race IAT)/ Noun (Nature IAT)	1	10	Noun
2	20	Adjective	2	10	Adjective
3	20	Mixed (Image/Noun & Adjective)	3	10	Mixed (Noun & Adjective)
4	40	Mixed	4	20	Mixed
5	40	Image/Noun	5	20	Noun
6	20	Mixed	6	10	Mixed
7	40	Mixed	7	20	Mixed

Results

The primary data obtained from IAT trials are RTs measured from the onset of the stimulus to a response. In this and the remaining studies, only correct responses were included in the analysis (correct responses > 93.5%). Mean RTs that were 3 standard deviations above

or below the mean RT in each cell (e.g., black person, flower, negative adjective) were removed from the analysis. If an individual's data were removed from any cell, then that participant was removed from the analysis ($N < 12$). We first analysed the race IAT results to test for the predicted interaction between stimulus (image/adjective) and valence (positive/negative), and then tested for an interaction between word type (noun/adjective) and valence in the nature IAT

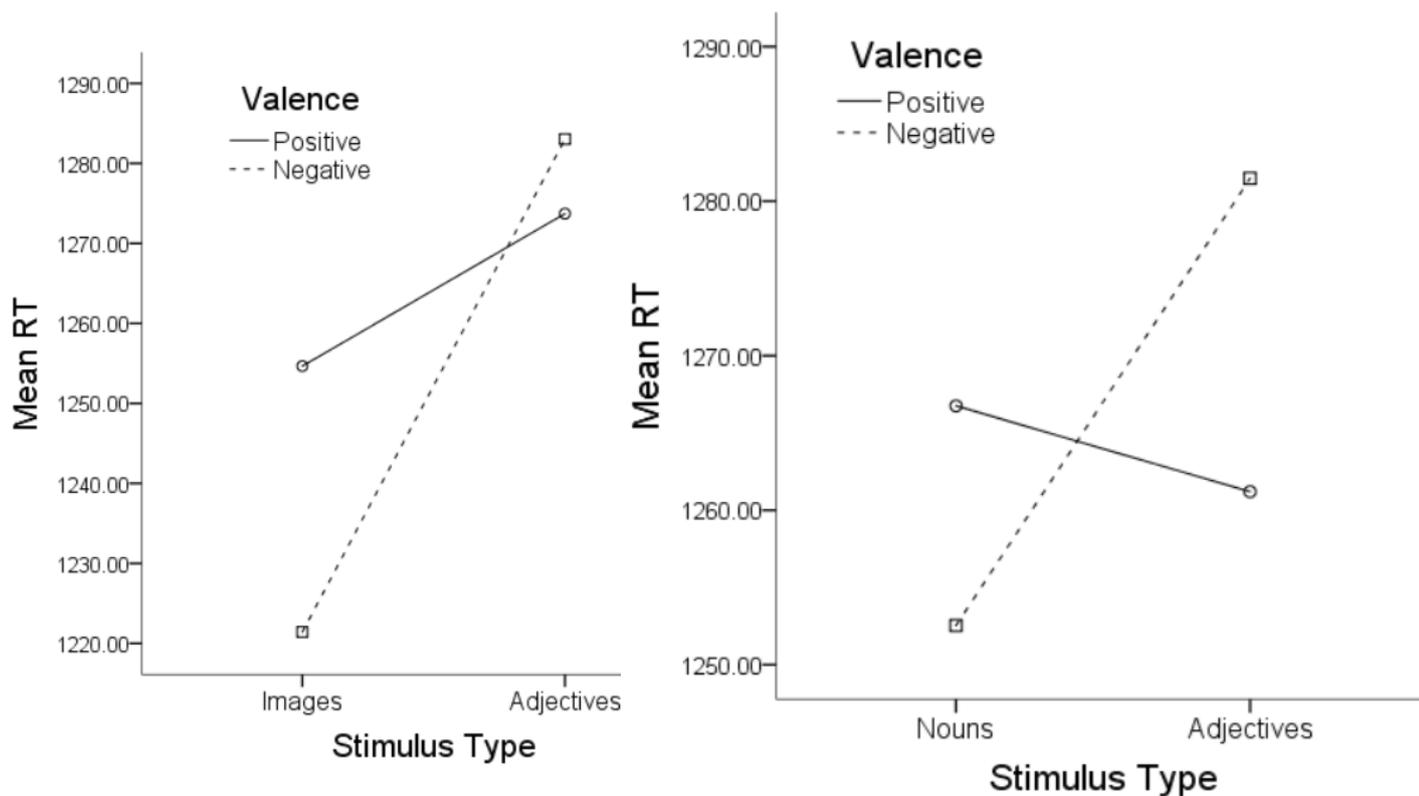


Figure 5.1: (left) Interaction between stimulus type and valence for the Race IAT. (right) Interaction between stimulus type and valence for the Nature IAT

For the race IAT, a 2×2 repeated-measures analysis of variance (ANOVA) showed that RTs were overall faster for images ($M = 1238.05$, $SE = 9.23$) than for words ($M = 1278.38$, $SE = 10.03$), $F(1, 332) = 106.46$, $p < .001$, $\eta^2 = .24$. RTs were also faster for negative/threatening items ($M = 1252.23$, $SE = 9.51$) than for positive/non-threatening items ($M = 1265.19$, $SE = 9.64$), $F(1, 332) = 17.17$, $p < .001$, $\eta^2 = .05$. However, of most interest

was a significant stimulus type \times valence interaction, $F(1, 332) = 52.01, p < .001, \eta^2 = .14$, indicating that participants showed faster RTs to *negative* images and *positive* adjectives than to their antonyms (Figure 5.1)²⁷.

For the nature IAT, RTs to nouns ($M = 1230.74, SE = 9.29$) were faster than RTs to adjectives ($M = 1271.34, SE = 9.46$), $F(1, 253) = 7.56, p = .006, \eta^2 = .03$. There was also a significant word type \times valence interaction, $F(1, 253) = 23.25, p < .001, \eta^2 = .08$; participants showed faster RTs to *negative* nouns and *positive* adjectives than to their antonyms (Figure 5.1). The main effect of valence was not significant, $F(1, 253) = .82, p > .250, \eta^2 = .00$, (Positive: $M = 1263.98, SE = 9.15$; Negative: $M = 1267.01, SE = 9.42$).

Discussion

Overall the findings support H1. People processed (concrete) nouns in a similar way to how they processed images – showing faster RTs on a VJT to negative nouns and negative images than to positive nouns and positive images. In contrast, with (less concrete) adjectives, the reverse held: people showed faster RTs to positive adjectives than to negative adjectives.

Study 2

Study 2 again used the IAT but with completely new stimuli. The study tested whether the interaction between valence and word type (nouns/adjectives) occurred when using stimuli that were specifically matched on word length (Balota, Cortese, Sergent-Marshall, Spieler, &

²⁷ Similar results were found using the raw data of the old-young and disabled-abled IAT made available to the authors by Project Implicit. For example, people were faster to sort pictures of old people and disabled people (presumably due to these groups evoking a fear/avoidance response, although salience could also be a potential explanation, see Rothermund & Wentura, 2004). In both these IATs, positive adjectives were also sorted faster than negative adjectives.

Yap, 2004), word frequency/contextual diversity (Adelman, Brown, & Quesada, 2006) and arousal (Kuperman et al., 2014). Furthermore, we evaluated participants' RTs to positive and negative verbs (which fall between nouns and adjectives in terms of concreteness). Study 2 also tested whether participants showed faster RTs to affirming words than negating words (H2) and, if participants showed faster RT for positive-affirming and negative-negating associations (H3).

Method

Participants: 168 participants (89 Males, 78 Females and 1 Other) took part in the online experiment and were recruited through Reddit (<https://www.reddit.com/r/SampleSize/>) over a three-month period. The majority of the sample were white (87%) and were mainly from the US (61%). 152 participants were in or had completed education above high school level and the mean age of the sample was 24.9 ($SD = 6.86$). Participation was voluntary. In order to increase the sample size and encourage appropriate responses, participants were informed that the “top 4 fastest and most accurate people to sort words will win \$20”.

Materials

Demographic information: Participants' gender, age, race, education and country of residence were obtained through an online questionnaire.

Implicit Association Test (IAT): Three IATs were presented, one using nouns, one using verbs and one using adjectives. The three IATs also required participants to sort affirming (e.g., True, Yes) and negating words (e.g., False, No; see Table 5.2 for stimuli). The positive and negative nouns, verbs and adjectives were matched on word length, arousal and contextual diversity. For each word type, the positive and negative stimuli on the valence dimension were significantly different from one another based on the Warriner, Kuperman and Brysbaert (2013) ratings ($ts > 15.65$, $ps < .001$). Furthermore, the positive and negative valence scores for each of the word types were similarly different from one another.

On concreteness (measured using ratings from Brysbaert et al., 2014), the positive and negative nouns, verbs and adjectives all differed significantly from each other ($F_s > 16.95$, $p_s < .001$). Nouns were the most concrete, followed by verbs and then adjectives. The category labels used in each IAT were “Support”, “Oppose”, “Positive” and “Negative”. “Support” and “Oppose” were used instead of the category labels “Affirm” and “Negate” on the expectation that participants would find it easier to associate “Negative” and “Negate” together simply because of the semantic similarity between these two category labels. Each IAT comprised 160 trials divided into seven blocks (see Table 5.3 for items and number of trials in each IAT block). Equal numbers of positive and negative words as well as supporting and opposing words were used in blocks which used a particular stimulus type.

Table 5.2: Stimuli used in the Noun, Verb and Adjective IAT as well as the VJT (Study 2-3). The support and oppose stimuli were used in all the tasks.

Noun (Positive)	Noun (Negative)	Verb (Positive)	Verb (Negative)	Adjective (Positive)	Adjective (Negative)	Support (Affirm)	Oppose (Negate)
Cash	Shark	Smile	Puke	Cozy	Scary	Yes	No
Puppy	Rat	Hug	Kidnap	Nicer	Horrible	True	False
Gold	Rapist	Achieve	Lie	Glorious	Lazy	Agree	Disagree
Sun	Vomit	Flirt	Cry	Brave	Bored	Confirm	Deny
Cake	Bomb	Kiss	Stab	Amazing	Nasty	Accept	Reject
Volunteer	Terrorist	Enjoy	Irritate	Friendly	Rotten		
Treasure	Disease	Win	Scare	Exciting	Ignorant		
Kitten	Virus	Create	Sicken	Jolly	Failed		
Vehicle	Death	Overcome	Panic	Gentle	Cruel		
Cinema	Garbage	Value	Upset	Wise	Jealous		

Table 5.3: Number of trials and stimuli presented in each IAT block (Study 2) and each VJT block (Study 3).

IAT			VJT		
Block Number	Number of trials	Stimuli presented	Block Number	Number of trials	Stimuli presented
1	20	Positive, Negative	1	30	Support, Oppose
2	20	Support, Oppose	2	20	Noun
3	20	Mixed (Support, Oppose, Positive, Negative)	3	20	Verb
4	20	Mixed	4	20	Adjective
5	40	Positive, Negative	5	20	Noun
6	20	Mixed	6	20	Verb
7	20	Mixed	7	20	Adjective

Design and Procedure: The experiment used a 3 (word type: nouns, verbs, adjectives) \times 2 (valence: positive, negative) mixed design with valence as the within-subject factor and word type as the between-subject factor. After participants had given informed consent, they completed the demographic information section. Participants were then allocated randomly to either the noun or adjective condition. Data for the verb condition were collected after the noun and adjective conditions. Therefore, the word type factor was not fully randomised²⁸.

Participants in each condition completed the standard IAT as described in Study 1 with the domain-specific stimuli used for each of the three IATs. After participants completed the

²⁸ See Appendix 5 where the findings were replicated using a fully randomised design, had a larger sample size (222 participants) and participants were incentivised (payment of \$1 for completing the experiment) through a different recruitment platform (Amazon Mechanical Turk).

IAT, they had the option to input their email address to enter a competition that awarded \$20 to the top four most accurate participants. If scores were tied on accuracy, the award was given to the participants with the smallest mean RTs. Following this, participants read an online debriefing form. The full experiment can be viewed at <https://brianpsychexperiments.warwick.ac.uk/Response/testa.html>

Results

H1 nouns vs. verbs vs. adjectives. A mixed 3 (word type: noun, verb, adjective) \times 2 (valence: positive, negative) ANOVA revealed that RTs to positive words ($M = 1214.42$, $SD = 139.19$) were faster than to negative words ($M = 1231.06$, $SD = 146.62$), $F(1, 164) = 12.55$, $p = .001$, $\eta^2 = .07$. Furthermore there was a significant word type \times valence interaction, $F(2, 164) = 6.14$, $p = .003$, $\eta^2 = .07$. As shown in Figure 5.2, RTs were numerically faster to negative nouns than positive nouns, $t(52) = .87$, $p > .25$, but RTs were significantly slower for negative verbs and adjectives than for their antonyms, $ts > 3.19$, $ps < .003$. The main effect of word type was not significant, $F(2, 164) = .58$, $p > .250$, $\eta^2 = .01$.

H2 affirming words vs. negating words. A mixed 3 (word type: noun, verb, adjective) \times 2 (response type: affirm, negate) revealed a main effect of response type, $F(1,160) = 37.74$, $p < .001$, $\eta^2 = .19$. Participants were faster to sort words into the support category ($M = 1209.63$, $SD = 139.90$) than into the oppose category ($M = 1248.60$, $SD = 157.37$). There was a significant difference between sorting the affirming (support) and negating (oppose) words in the verb, $t(54) = 4.69$, $p < .001$, the noun, $t(53) = 3.99$, $p < .001$ and adjective IATs, $t(53) = 2.104$, $p = .04$ (see Figure 5.2). The main effect of word type and the word type \times response type interaction was not significant, $F_s < 1.73$, $ps > .18$.

H3 affirming-positive, negating-negative association biases. Individual's D-IAT scores were computed using the standard procedure/algorithm used in IAT research (Greenwald et al., 2003). Participants with error rates above 30% or RTs above 10,000ms or below 300ms on

more than 10% of trials were excluded (four participants). Scores above zero indicate participants were faster to respond to affirming and positive words, as well as negating and negative words. In contrast, scores below zero indicate the reverse. We used a one-sample t -test to determine whether participants D-IAT scores were significantly above or below zero. The results showed that participants had significantly faster RTs when associating affirming and positive words and negating and negative words in memory than the opposite, $t(163) = 41.49, p < .001, d = 3.24$. Furthermore, participants were significantly faster to sort affirming and positive words ($M = 1207.19, SD = 126.07$), then to sort negating and negative words ($M = 1233.76, SD = 139.81; t(163) = 6.88, p < .001, d = 0.54$).

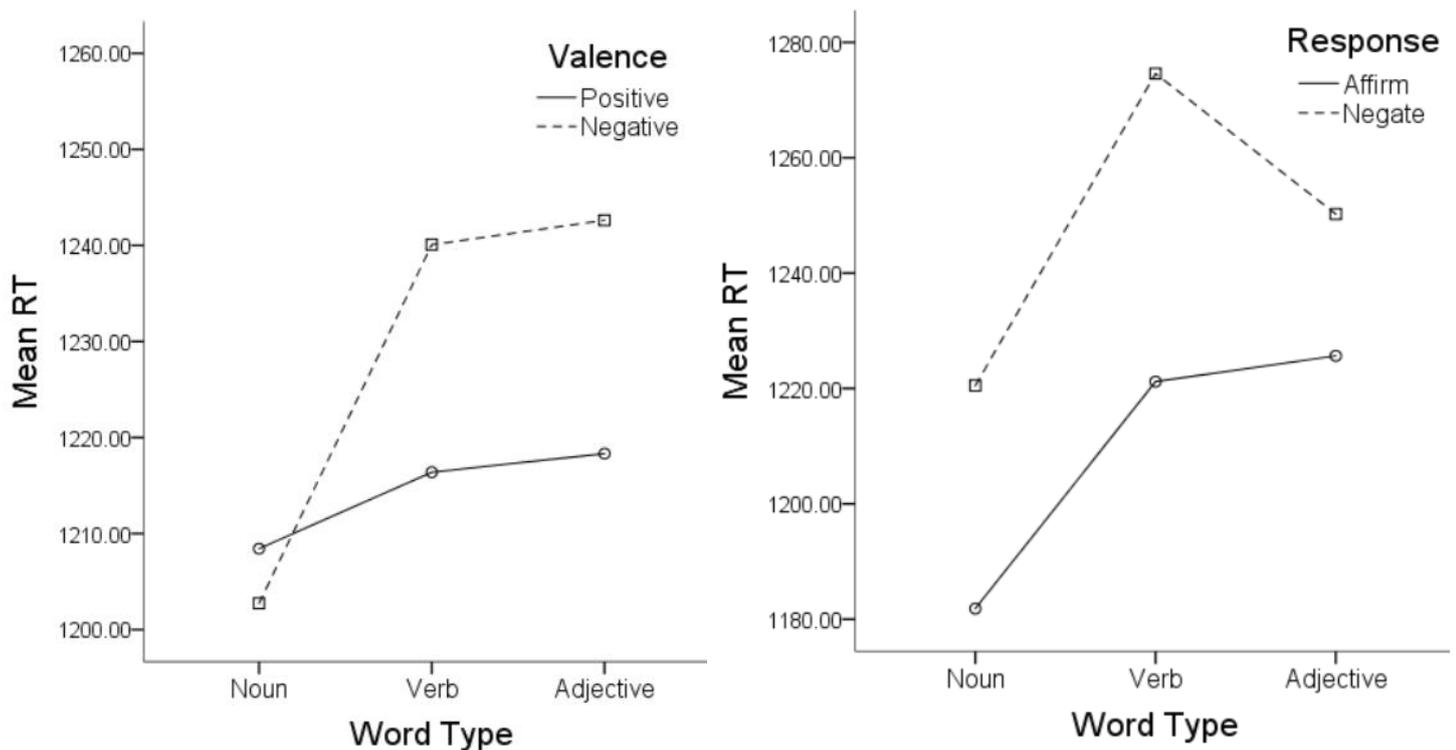


Figure 5.2: (left) Interaction between word type and valence. (right) Main effect of response type.

Discussion

Using stimuli matched on numerous factors, we again found the predicted (H1) interaction between nouns (most concrete), verbs (less concrete) and adjectives (least concrete). In addition, participants had faster RTs when sorting positive adjectives and verbs than their antonyms, but for nouns, there was little difference. Participants were also faster to sort affirming words than negating words (supporting H2). Finally, participants had faster RTs when associating positive and affirming, and negating and negative, compared to the reverse association (supporting H3).

Study 3

Study 3 aimed to replicate previous findings (supporting H1 and H2) using a within-subject design in which participants responded to the positive and negative nouns, verbs and adjectives as well as the affirming and negating words used in Study 2. Rather than using a standard IAT, we used a simpler approach in which only two (rather than four) concepts were presented within a block. This new task will test the generalisability of the effects, in addition to determining whether the same effects occur when participants complete just one sorting task (sort word types or response type) rather than two (sort word types and response type).

Method

Participants: The replication of Study 2 (see Appendix 5) had approximately 70 participants in each cell in a between-subject design. We therefore set the stopping rule at 70 participants for the current experiment which used a within-subject design. Recruitment was via Reddit. 40 males, 29 females and 1 other completed the experiment. The mean age of the sample was 26.4 ($SD = 8.19$), with a majority of white participants (>81%) from the US (70%) taking part. Over 61 participants had college/university degrees or were in education above high school level.

Materials: The stimuli were the same as Study 2. However, instead of the IAT, a simpler RT task (VJT) was used.

Procedure: The experiment used a 3 (word type: nouns, verbs, adjectives) \times 2 (valence: positive, negative) within-subject design. The major difference between the current study and the previous study was that only two categories appeared within each block throughout the task. The word that appeared at the centre of the screen had to be quickly and accurately sorted into one of two categories displayed at the top of the screen. If the word belonged to the category on the top left, pressing the “E” key was the correct response, and if the word belonged to the category on the top right, pressing the “I” key was the correct response.

Overall, each participant completed 150 trials over seven blocks. Block 1 always had 30 trials with the “Support” and “Oppose” category labels as well as the corresponding stimuli. Blocks 2-4 comprised of one block of nouns, one block of verbs and one block of adjectives with block order randomised. For each participant, the order of blocks 5-7 was the same as the order of blocks 2-4. All other aspects of the VJT were the same as the IAT (see <https://brianpsycheperiments.warwick.ac.uk/Response/testc.html> for the full experiment).

Results

A 3 \times 2 repeated measures ANOVA with word type (noun, verb and adjective) and valence (positive and negative) as factors revealed a significant main effect of word type $F(2, 130) = 16.23, p < .001, \eta p^2 = .20$, and a significant word type \times valence interaction, $F(2, 130) = 8.07, p < .001, \eta p^2 = .11$. As shown in Figure 5.3, participants responded faster to both positive adjectives, $t(65) = 2.32, p = .023$ and verbs, $t(65) = 2.78, p = .007$, but numerically

slower to positive nouns, $t(65) = 1.90, p = .062^{29}$ relative to their antonyms. The main effect of valence was not significant, $F(1, 65) = 2.60, p = .112, \eta p^2 = .04$. With respect to response type, a paired samples t-test showed that affirming/support responses ($M = 1081.70, SD = 90.96$) were faster than negating/oppose responses ($M = 1108.79, SD = 113.85$; Figure 5.3), $t(66) = 2.54, p = .013, d = .31$

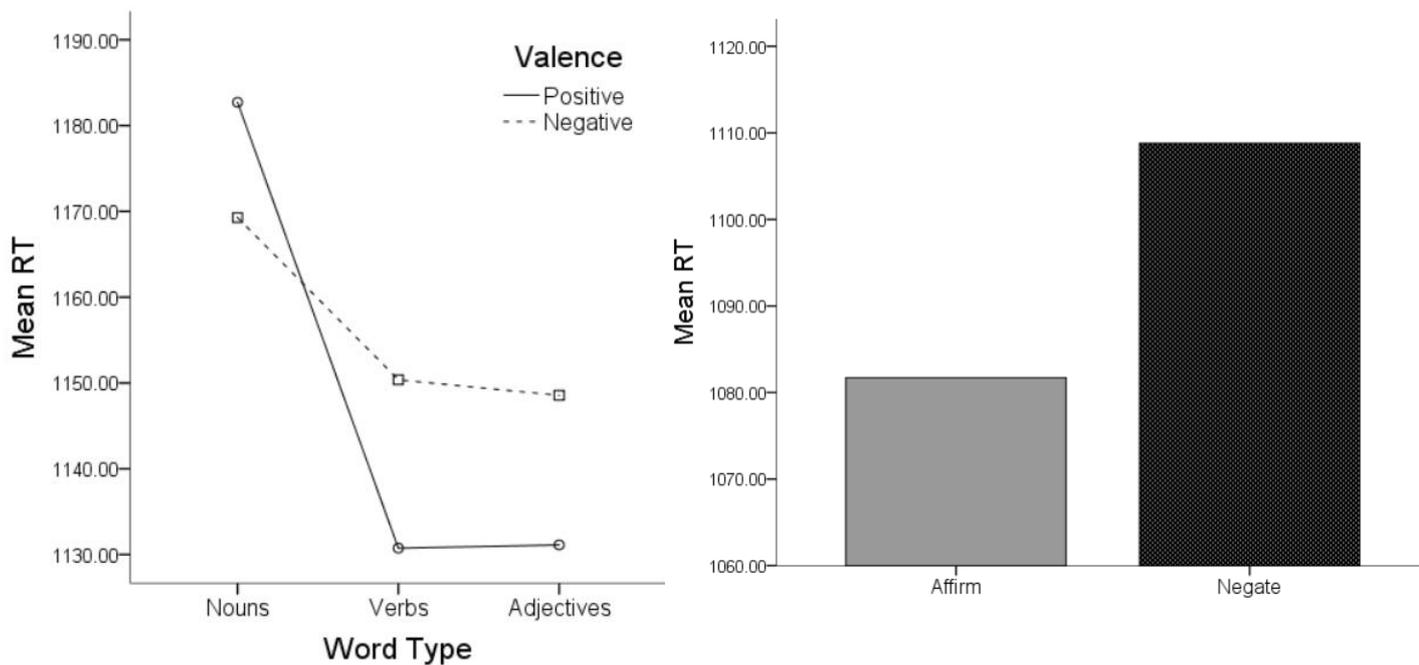


Figure 5.3: (left) Interaction between word type and valence. (right) Main effect of response type.

²⁹ Given our directional prediction and the previous findings, we would expect that negative nouns would be responded to faster than positive nouns and so there is some justification for treating this test as one-tailed which would then be significant at the .05 level.

Discussion

Study 3 replicated the findings from Studies 1-2, showing that people responded faster to negative nouns, positive verbs and positive adjectives than to their antonyms. The expected interaction between word type and valence was also replicated (H1). We again found that participants were faster at sorting affirming words than sorting negating words (H2). This study provides strong supporting evidence for H1 and H2 using a within-subject design.

General discussion

We predicted that highly concrete words such as nouns would be processed like images, resulting in faster RTs to negative than to positive stimuli on a VJT. We expected that this would not occur with less concrete words because they convey more abstract concepts which are less likely to signal an immediate environmental threat. Therefore, H1 predicted the presence of a word type \times valence interaction – which we found in all three studies. This finding has important implications for many psychological tasks in which different types of word stimuli are separated or mixed. Systematic analysis of the influence that word type has in the most popular measures of implicit attitudes (IAT, SC-IAT, GNAT, Brief IAT) is therefore warranted.

An additional explanation for the slower RTs to negative verbs and adjectives than to their antonyms is that people might perceive these negative words as relating to themselves rather than an external threat in the environment. This perception may evoke an ego-threat, lowering an individual's self-esteem and emotional stability (Waller, Watkins, Shuck, & McManus, 1996). This ego-threat response potentially inhibits the processing of negative verbs/adjectives in order to protect an individual's self-esteem (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). A common finding in implicit self-esteem research is that both normative and depressed samples are faster to associate positive rather than negative adjectives with the self (De Raedt, Schacht, Franck, & De Houwer, 2006). The current research suggests

that it might be more appropriate to use nouns instead of adjectives in the self-esteem IAT, as this change may improve the detection of those with negative self-biases.

H2 predicted that participants would show faster RTs to affirming words (True, Yes) than to negating words (False, No). A between-subject analysis (Study 2) and a within-subject analysis (Study 3) found this to be true. This bias has major implications for many tasks that contain affirming/negating components such as the IRAP (Barnes-Holmes, Barnes-Holmes, et al., 2010), the SIP (O'Shea, Brown, & Watson, 2017) and the Relational Responding Task (De Houwer et al., 2015) to name just three. This affirming bias will lead to an over-inflation of estimates of positivity to any category (e.g., flowers or insects) associated with positive words.

Our final hypothesis (H3) predicted that people are more likely to store affirming-positive, and negating-negative word associations in memory than the opposite associations (see Proctor & Cho, 2006, for an explanation of why this occurs). Using the IAT, Study 2 showed this to be the case yielding an extremely large effect size. Various measures of implicit attitudes (ST-IAT, GNAT) are likely to be affected by this bias which could result in inflated positive/negative attitudes towards concepts/categories.

Limitations

One potential limitation is that we used a web-based rather than a tightly controlled experimental-based approach (but see Briones & Benham, 2017). However, even if our data were noisy, this cannot account for the differences found between the images and word types. Another limitation is that we did not determine the specific level of concreteness which results in a switchover from faster to slower responses for negative stimuli. Determining this would be a useful goal for future research, but nonetheless, this issue does not undermine our main findings.

Conclusion

Overall, our study illustrates previously unknown issues in the use of nouns, verbs and adjectives that can have serious consequences for measures of implicit attitudes and likely many other tasks. Pre-testing the positive and negative valenced words used in implicit tasks (especially words used in absolute implicit task) is necessary to ensure RTs to each item are matched as closely as possible. Furthermore, researchers should be aware that faster RTs are likely to occur whenever affirming, rather than negating, responses are used. Greatest caution is needed when positive items must be affirmed and when negative items are to be negated. All these biases will influence participants' RTs as well as the subsequent interpretations made by researchers and therefore, must be taken into account.

Chapter 6: Disgust versus fear: Using the SIP to measure racial prejudice

Abstract

Both high rates of naturally occurring disease rates in the environment and experimental presentation of disease primes lead to heightened prejudice towards out-groups. The SIP has been shown not to be subject to practice/experience effects over time (Chapter 4). Therefore, any changes from baseline implicit biases detected using the SIP, following a terror or disease priming induction, will be likely due to the primes. It was found that white female participants had an implicit preference for whites relative to blacks and that this bias remained stable over time, regardless whether participants were primed with disease or terrorism images. Similar findings were found for participants' explicit biases (i.e., stronger anti-black/pro-white biases) and the relations between the implicit and explicit measures were significant. Unexpectedly, before viewing the disease and terror primes, the white female participants reported that white people are more likely to use faster life history strategies (e.g., more impulsive strategies; strategies of lower investment in education and their children) than black people. However, following the primes, they reported that both white and black people used similar life history strategies (i.e., scores converged). Gender differences, particularly tending and befriending responses, and sexual strategies are invoked to explain the findings. It is also possible that the SIP is not influenced by context effects but instead measures stable, long-term, early socialisation biases.

Introduction

Many studies show that outcome scores on implicit measures may decrease, increase or can even reverse as a function of the context (Gawronski & Sritharan, 2010). For example, a reduction in anti-black/pro-white biases on the IAT occurs after: (1) showing participants images of disliked white individuals (e.g., serial killers) and admired black individuals (e.g., Martin Luther King; Dasgupta & Greenwald, 2001; see also Barden, Maddux, Petty, & Brewer, 2004) (2) showing participants a movie clip of black individuals in a positive situational context compared to a negative one (Wittenbrink, Judd, & Park, 2001) and (3) participants interact with a black experimenter (Lowery, Hardin, & Sinclair, 2001). Lai et al., (2014) repeatedly tested 17 interventions that had previously been effective at reducing implicit racial prejudice across various labs. They found that only eight interventions were consistently effective at reducing implicit racial prejudice. These were typically interventions that invoked high self-involvement or linked white people with negativity and black people with positivity. None of the 17 interventions consistently reduced explicit racial biases.

When implicit measures were initially developed, there was a dominant assumption that the associations they assessed represented a history of learning, influenced by early socialisation, and therefore must be stable over time and difficult to change (Rudman, 2004; Wilson, Lindsey, & Schooler, 2000). The findings described above led researchers to hypothesise that implicit measures capture biases that are constructed on the spot (e.g., Schwarz, 2007, constructionist account), while others argued that implicit representations are stable across time but that contextual information only changes the representation of a specific target object (e.g., Fazio, 2007, stable representation account). For example, an individual can express negative biases towards black people as a group but the same individual can have positive biases towards specific black people, such as Barack Obama or Martin Luther King.

Recently Lai et al. (2016) assessed the effectiveness of the 8 best interventions (see Lai et al., 2014) in reducing racial prejudice both immediately following the intervention and over time (between one and four days). Explicit attitudes remained stable throughout but implicit prejudice was reduced immediately following an intervention. However, reassessment of participants after a day or a few days showed that the interventions were not effective in the long term because participants' racial prejudice increased to similar levels as seen in the control condition. Three explanations were proposed to explain these findings. (1) Effective long-term mechanisms have not yet been developed, (2) interventions need to be longer and more intensive and (3) interventions could be more effective on children than adults. Nevertheless, these findings are more in line with Fazio's (2007) stable representations in memory account because implicit representation rebounded back to initial/original levels of implicit prejudice.

There is also the possibility that the interventions used were not actually changing implicit prejudice per se, but were instead changing non-associative factors that are related to IAT performance (e.g., Calanchini, Sherman, Klauer, & Lai, 2014). For example, changes in IAT scores might have reflected temporary changes in task performance, rather than altering associations in memory (Lai et al., 2016). This possibility is especially likely when participants try to fake or exert strategic control over the outcome of their IAT score (see Blair, 2002; Fiedler & Bluemke, 2005). The current experiment will test the extent to which the SIP is influenced by context effects by using disease primes, which have previously been shown (Chapter 2) to increase anti-black/pro-white biases in the IAT. Other studies (e.g., Duncan & Schaller, 2009; Park, Schaller, & Crandall, 2007) have also shown that disease primes increase prejudice towards out-groups.

One of the advantages of using the SIP is that it allows a researcher to gain an in-depth understanding of what mechanisms are driving an individual's prejudice. This is because the SIP measures attitudes towards separate concepts or categories (i.e., attitudes towards white

and black people can be separately explored). This contrasts with the IAT, which can only measure attitudes relatively (i.e., attitudes towards white people relative to black people). Therefore, the SIP can identify the mechanisms that result in an increase in anti-black/pro-white biases for individuals primed with diseases and subsequently measured with the SIP (i.e., pro-white bias increasing, anti-black bias increasing, or both). Another advantage of the SIP is that it does not have a practice effect problem like the IAT (see Chapter 4, Study 3). This means that the SIP can be used pre- and post-intervention, allowing researchers to measure changes in individuals' attitudes over time without needing a control condition.

The SIP has an affirming bias (see Chapter 4, Study 1 & 2) but this flaw can be overcome through a pre- and post-intervention design. Prior to an intervention, affirming biases towards the target attitude object (e.g., black or white people) can be determined (baseline), and following the intervention when the SIP is repeated, changes in biases can be compared to the baseline SIP with the affirming bias. O'Shea and colleagues (Chapter 2, Study 3) used a between-subject design and showed that disease primes increased implicit anti-black/pro-white biases compared to a terror and control condition. For the explicit attitudes, both the disease and terrorism condition showed increased anti-black/pro-white biases compared to the control. Of note, when the data were split by gender (see Appendix 2), only males showed an increase in their implicit and explicit anti-black/pro-white biases in the disease condition compared to the control condition.

A potential explanation for these findings is that fewer females (170) than males (224) completed the experiment and that no significant differences were observed for females due to power issues arising from the smaller sample size. Additionally, the size of the effect linking disease primes to increased males' prejudice was, while significant, small. Consequently, the between-subject experimental design used might not have been optimal to detect any small, but observable, effects of disease primes increasing prejudice for females. Perhaps a within-

subject design would be more suited to detect the effects of disease primes on females' implicit and explicit biases.

Present study

To examine this idea, the current experiment used a within-subject design across a control (baseline/pre) and intervention³⁰/post time points because the SIP allows such a design to be used. This experiment aims to build on the existing evidence by determining whether environmental threats (e.g. disease & terrorism) increase females' prejudice towards black people. A large body of evidence has found that males express increased prejudice towards out-groups compared to females (e.g., McDonald, Navarrete, & Van Vugt, 2012, Chapter 2, Studies 1-3). Some studies have also reported gender differences in response to disease primes, with males, but not females, often showing increased implicit prejudice towards out-group (Klavina, Buunk, & Pollet, 2011). Females often display tending and befriending behaviours, resulting in more empathetic and caring responses compared to males (Taylor et al., 2000). This tending and befriending response could inhibit a rise in prejudicial attitudes occurring for females following disease primes.

Recent evidence has also shown that females express increased attraction to out-group males as their fertility increases due to the need to create genetic diversity in their offspring to improve evolutionary fitness (Salvatore, Meltzer, March, & Gaertner, 2017). Importantly, females commonly express increased vulnerabilities to disease (Duncan, Schaller, & Park, 2009), which is related to a various of behaviours (i.e., prejudice towards out-group,

³⁰ The term "intervention" often denotes a method aimed at improving a situation (e.g., reducing racial prejudice). However, in the current study the intervention aims to increase racial prejudice by use of disease or terror images to prime disgust or fear. Therefore, instead of using the term "intervention", the term "threat priming" will be used.

conservatism, higher religious belief, see Chapter 2). On the basis of this conflicting evidence, two predictions can be made regarding the effects of the disease primes. (1) Primes will increase white females' anti-black/pro-white biases and this finding would be more in line with PST, (2) The primes will not influence implicit biases (i.e., biases remain stable over time) and this finding would be more in line with Taylor's tending and befriending hypothesis.

In line with previous research on implicit and explicit attitude change using within-subject designs, explicit attitudes are expected to remain stable over time even when changes are observed for implicit attitudes (Lai et al., 2014, 2016). The lack of explicit attitude change are expected because only one or two of the same items are assessed in close succession, pre- and post-threat primes (i.e., measuring explicit attitudes towards black people and ten minutes later using the same explicit self-report measure to assess attitudes towards black people again). Therefore, it is easy for participants to remember how they had previously responded on the explicit measure. To appear consistent and not easily manipulated by the experimenter, participants might be matching how they had previously responded. Therefore, the current experiment used the standard explicit bipolar and feeling thermometer scales that have been used in previous research; people's responses on these scales normally remain stable over time. In addition, a much longer questionnaire relating to life history strategies (e.g., Kaplan & Gangestad, 2015), which uses different but similar questions, was used pre- and post-threat primes. Using a within-subject design, it is expected that the longer life history scale would have a better chance of detecting changes in participants' explicit attitudes over time. Based on prior research (Williams, Sng, & Neuberg, 2016), it was predicted that the sample would report that black people use faster life history strategies than white people.

Method

Participants: 90 participants took part in the study. However, 20 were discarded because they did not select white as their race. Five remaining white males were also excluded

because this experiment aimed to measure biases only in white females³¹. Therefore, the final sample was composed of 65 female participants. Of these, 52 described themselves as English/British, with the others being from European countries and one individual from Russia. The mean age of the sample was 18.82 ($SD = .66$). Participants were politically moderate ($M = 3.58$, $SD = 1.60$) and mainly non-religious/slightly religious ($M = 1.52$, $SD = .64$). 31 participants were assigned to the disease condition and 34 to the terrorism condition. All participants were first-year psychology students and received course credit for taking part.

Materials

Demographic information was obtained via an online questionnaire relating to participants' gender, age, race, nationality, political ideology (1 = Strongly Liberal to 7 = Strongly Conservative) and religious belief (1 = not at all religious to 4 = strongly religious).

Simple Implicit Procedure (SIP): The SIP was programmed using Blitzmax (www.blitzbasic.com/Products/blitzmax.php), running on an Intel Windows 7 desktop, attached to a 22" LCD screen with a resolution of 1440 x 900 pixels. A single trial consisted of three elements:

(1) on the top of the screen one of 16 positive or negative target stimuli/words appeared (see Table 6.1 for stimuli)

(2) a category label appeared at the centre of the screen which was one of 18 stimuli composed of 3 words referring to black people, 3 words referring to white people and 12 images of a single black or white individual's face. The majority of these items were taken from the standard IAT used by Project Implicit (see Nosek et al., 2007)

³¹ In order to receive course credit, non-white and male psychology students also had to be accepted to take part in the experiment. No differences apart from slightly increased pro-white/anti-black biases were shown if the white males were included in the analysis.

(3) “Yes”, “Space” and “No” response options labels appeared simultaneously at the bottom of the screen. The word “Space” always appeared at the bottom centre of the screen, with the word “Yes” at the bottom left and “No” at the bottom right of the screen for half of the participants. The other half had “Yes” on the right and “No” on the left of the screen throughout the task (counterbalanced across participants, see Figure 6.1). The response labels and keypresses were spatially congruent (i.e., an “E” keypress corresponded to the response label on the left and an “O” keypress corresponded to the response label on the right of the screen, and “Space” corresponded to the space bar key).

Table 6.1: Stimuli used in the SIP.

Positive Target Stimuli	Negative Target Stimuli
Joy	Agony
Love	Terrible
Peace	Horrible
Wonderful	Nasty
Pleasure	Evil
Glorious	Awful
Laughter	Failure
Happy	Hurt
European/White Category Label	African/Black Category Label
Six images of white people (3 males)	Six images of black people (3 males)
White Person	Black Person
European	African
British White	British Black
Response Options: True, False, Space.	

Questionnaires: Explicit attitudes were obtained using two separate questionnaires (QA and QB) that measured comparable constructs, but used different terminology to measure these constructs. One was used following the control primes, while the other was used following the threat primes (counterbalanced across participants). The Cronbach's alphas for QA and QB were moderate ($\alpha = .68$ and $.72$ respectively). Each questionnaire included 22 questions relating to life history strategies used by both black and white people. These questions were taken from Williams, Sng and Neuberg (2016) and included items related to impulsiveness, opportunistic behaviours, investments in education, investment in children and sexual unrestrictedness. A Likert scale ranging from 1 to 6 was used for each question, with higher scores indicating a faster life history strategy (i.e., more impulsive, higher risk taking, lower investment in education and their children and less sexual restrictiveness). One questionnaire (QA) also measured how warm or cold participants felt towards black and white people using a feeling thermometers (0 = very cold, 5 = neutral and 10 = very warm). The other questionnaire (QB) measured attitudes towards black people relative to white people using a bipolar Likert scale (response options ranged from 1 to 10) (see Appendix 6 for items).

Control, disgust and terror images: A total of 90 images were used as primes in the experiment. 30 of the images acted as controls; these included 15 images of individual buildings and 15 of furniture items. 30 images showed people with infections and diseases which aimed to evoke disgust. The final 30 images included scenes of destruction and gun wielding individuals and these aimed to prime terrorism. All these images were taken from study 3 reported in Chapter 2. The same number of black and white individuals were shown in the disease prime and terrorism prime condition.

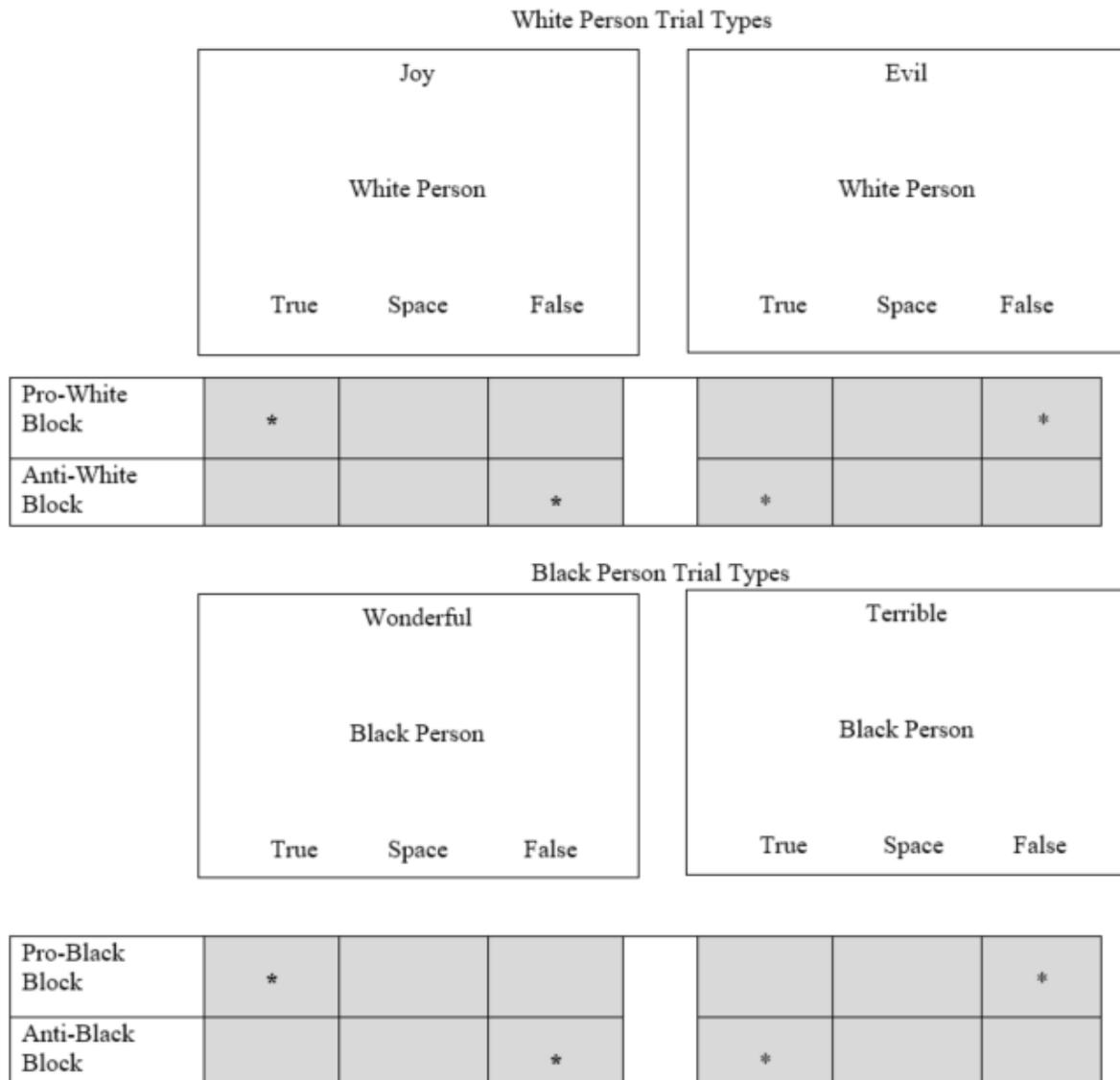


Figure 6.1: Screen shot examples of the four SIP trial types. Asterisks indicate the correct response option to press on each block. For each trial type, latency differences are compared between the correct “Yes” and “No” responses across the pro-white and the anti-white blocks, and the pro-black and anti-black blocks.

Design and procedure

The experiment used a 4 (SIP trial type: White-Positive, White-Negative, Black-Positive, Black-Negative) \times 2 (time: pre-threat primes, post-threat primes) \times 2 (prime: disease, terrorism) mixed design. The SIP trial type and time were within-subject factors and prime was a between-subject factor. Participants completed the experiment individually in a small well-

lit room. After giving informed consent, participants were directed towards an online survey which collected demographic information. Following these questions, participants viewed the 30 control images (order randomised for each participant) for as long as they needed but a minimum of 30 seconds had to elapse before they could continue to their assigned explicit questionnaire (i.e., QA or QB that was counterbalanced across participants). Before completing their assigned explicit questionnaire, the experimenter informed each participant to “try avoid constantly selecting the neutral option”. This statement was included because pilot testing revealed that the neutral option was usually selected. To comply with ethics a neutral option had to be included.

Following the baseline questionnaire, participants viewed the same control images for at least 30 seconds but in a different randomised order. Each participant then completed the baseline SIP. On-screen instructions informed participants that: “In this task, you will have to respond to statements in the appropriate fashion as fast and as accurately as possible” and the experimenter verbally explained how the task was to be performed. Each participant was required to successfully complete the practice blocks which included a minimum of four blocks of trials. For each practice block, participants had to respond in accordance with one of four rules across a block of trials.

These response rules were (1) “On this block respond as if European/White Person is POSITIVE” (pro-white), (2) “On this block respond as if European/White Person is NEGATIVE” (anti-white), (3) “On this block respond as if African/Black Person is POSITIVE” (pro-black) and (4) “On this block respond as if African/Black Person is NEGATIVE” (anti-black). Each participant viewed these four response rules in random order. For example, if a participant was completing the block “respond as if European/White Person is Negative”, when a positive word and a picture/word of a white person appeared, “No” was the correct response. If a negative word appeared with a picture/word of a white person, “Yes”

was the correct option. If a picture/word of a black person was shown and regardless if the target word was positive or negative, “Space” was the correct response option.

Each practice block had 10 trials, which included 4 “Space”, 3 “Yes” and 3 “No” response trials. Trials were presented quasi-randomly, such that the same trial type could not be repeated across two successive trials. If a correct response was made, the screen was cleared for 400ms before the next trial’s stimuli appeared. If an incorrect response was made, a red “X” appeared below the category label and remained there until the correct response was selected. Following each block, participants’ median accuracy and response latency were presented on screen and if the combined scores across the 4 practice blocks averaged $\geq 80\%$ accuracy and $\leq 1,800\text{ms}$ response latency, participants continued to the test blocks. These criteria were used to ensure that participants understood and complied with each block’s response rule. If participants did not meet the criteria, the 4 practice blocks were repeated. If they failed to meet the accuracy and response latency criteria on the second attempt, the study ended and they were thanked and debriefed.

After completing the practice blocks successfully, participants were informed on screen to “press space bar to begin the test blocks”. For the test blocks, participants completed 8 blocks (i.e., a random order of the 4 response rule blocks and when they were completed, the same random order of the 4 response rule blocks had to be completed again). Each test block contained 22 trials which consisted of 6 “Space” bar, 8 “Yes” and 8 “No” response trials. When all the 8 blocks were completed, participants were informed on screen that the task was over and instructed to tell the experimenter. Participants were then shown either the disgust or terrorism primes, using a similar procedure to the baseline/control primes (image set was counterbalanced across participants). The explicit questionnaire that participants had not previously completed (either QA or QB) was presented and then the same images (disgust/terror) were observed again but in a different random order for at least another 30

seconds. The same SIP as the baseline SIP was completed following the primes but it did not include the practice blocks because participants had already completed the task with the exact same stimuli. When the SIP was completed, participants were thanked and debriefed. Figure 6.2 shows a diagram illustrating the procedure.

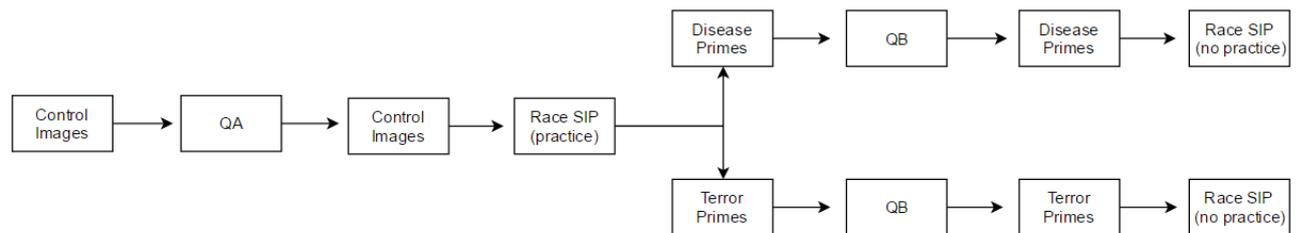


Figure 6.2: Diagram of the procedure used in the experiment. The QA and QB switched location for each participant (counterbalanced). Allocation to the disease or terrorism priming condition was also counterbalanced across participants.

Results

The primary data obtained from the SIP are raw latency scores, defined as the time in milliseconds that elapsed between the onset of the stimulus and the correct response being made by the participant. The dependent variable was the mean “No” minus “Yes” RT, for each of the four trial types across the pro-white and anti-white blocks as well as the pro-black and anti-black blocks. For the White Person-Positive and Black Person-Positive, values above zero indicate a positive bias and negative values indicate a negative bias. In contrast, both the White Person-Negative and Black Person-Negative trial types were multiplied by -1. The purpose of this was to match the valenced output of the white and black person trial types, such that scores above zero indicate a positive bias and scores below zero indicate a negative bias.

The steps involved in calculating both the absolute and relative D-SIP scores are as follows: (1) only response latency data from the 16 test blocks were used; (2) latencies above 10,000ms were discarded; (3) if latencies from more than 10% of a participant’s trials throughout the 16 test blocks were less than 300ms, that participant was removed from the

analysis; (4) for each SIP task, 16 individual standard deviations were calculated for each trial type across the 16 test blocks (White - Positive, White - Negative, Black - Positive, Black - Negative $\times 4 =$ repeating blocks); (5) 16 mean latencies were calculated for both the “Yes” and “No” responses for each trial type across the 16 test blocks; (6) difference scores were calculated for each trial type by subtracting mean latencies of “Yes” responses from mean latencies of “No” responses in each test block pair (i.e., pro-white block and anti-white block; pro-black block and anti-black block); (7) each difference score was then divided by its corresponding standard deviation from step 4, yielding 16 D-SIP scores, one score for each trial type across the 16 test blocks; (8) four overall trial type D-SIP scores were calculated by averaging the four scores for each of the four trial types across the blocks (these calculations revealed the absolute/non-relative separate trial type results); (9) averaging the positive and negative trial type for each category showed the absolute/non-relative compacted trial type results); (10) To compute the relative comparison, equivalent to that of the IAT, an overall relative D-SIP score was calculated by subtracting the compacted Black score from the compacted White score.

The baseline D-SIP absolute trial type data were subjected to a 4×2 mixed analysis of variance (ANOVA) with trial-type (white positive, white negative, black positive and black negative) as the within-subject variable, and condition (disease and terror)³² as the between-subject variable. The ANOVA yielded a significant main effect of trial type, $F(3, 63) = 42.87$, $p < .001$, $\eta^2 = .41$, but all other results were non-significant (F s < 1.49 , p s $> .23$). Figure 6.3 clearly shows why the main effect of trial type occurs by presenting the mean D-SIP scores for

³² For the baseline SIP, the conditions (disease & terror) should not influence the results because all participants saw the same furniture and building (control condition) at baseline.

the four individual trial types. The compacted White and compacted Black trial types as well as the relative/overall D-SIP score are also included in Figure 6.3.

Several one-sample t-tests were conducted to verify which of the absolute trial types scores, as well as the relative/overall D-SIP score, differed significantly from zero. The White-Positive, Black-Positive, compacted White and compacted Black scores were significantly above zero, $t_s > 4.75$, $p < .001$. The White-Negative and Black-Negative trial types were a non-significant distance away from zero, $t_s < 1.43$, $p > .16$. Finally, the overall D-SIP score was a significant distance above zero, $t(64) = 2.07$, $p = .043$, indicating that participants overall had a stronger anti-black/pro-white bias.

Like the baseline SIP scores, the post threat primes D-SIP's individual trial type data were subjected to a 4×2 mixed repeated measures ANOVA. The analysis again showed a significant main effect of trial type $F(3, 63) = 43.99$, $p < .001$, $\eta p^2 = .41$, but no effect of condition or interaction ($F_s < 1.90$, $p_s > .17$). Figure 6.4 shows the mean D-SIP scores for the four individual trial types, the compacted White and Black trial types and the relative/overall D-SIP score following the disease and terrorism primes. A $4 \times 2 \times 2$ mixed ANOVA with trial type and time as within-subject factors and condition as the between subject factor showed all the main effects and interactions to be non-significant, $F_s < 3.54$, $p_s > .07$.

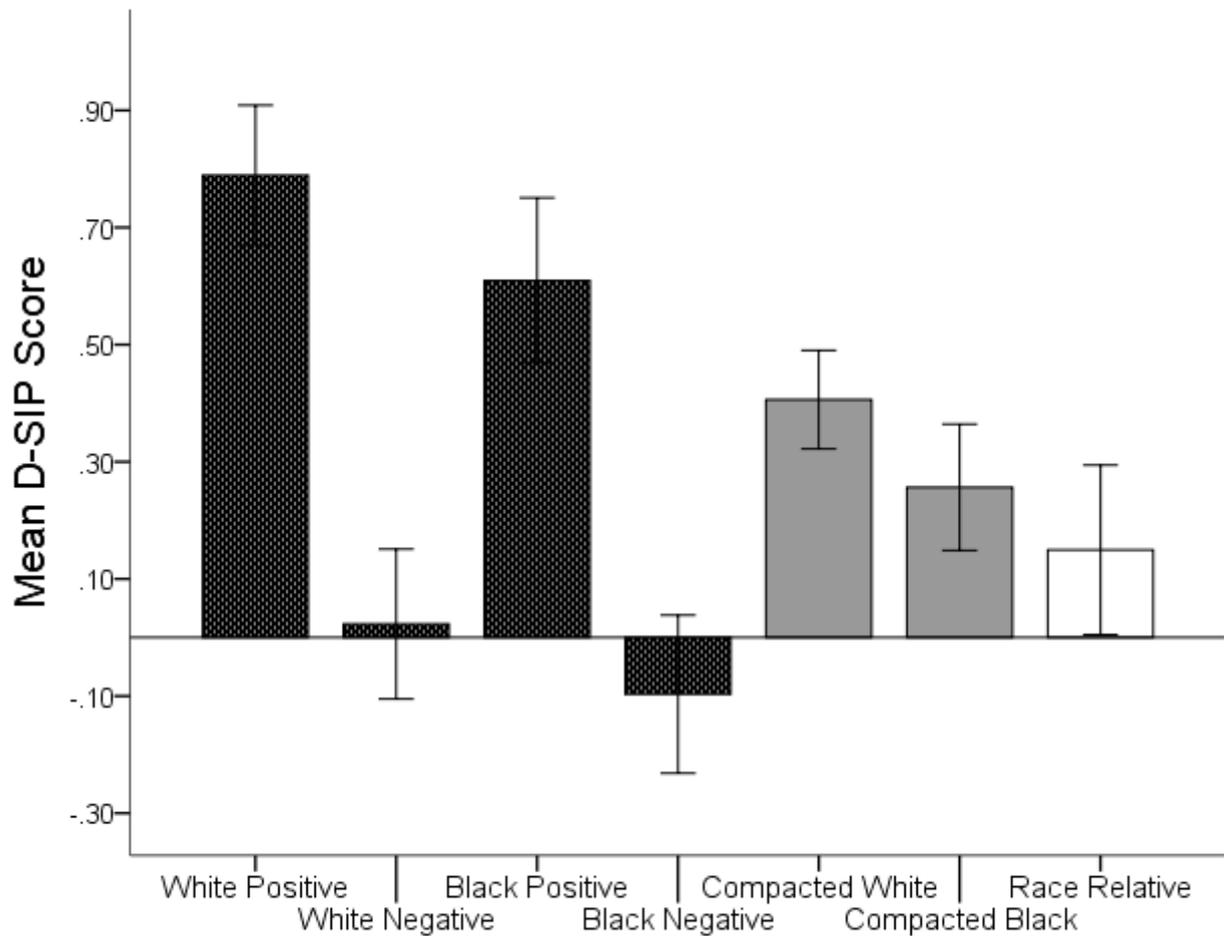


Figure 6.3: Baseline Race SIP: Mean D-SIP scores for the four individual trial types, the compacted trial types and the relative/overall D-SIP results are shown. For the positive trial types (i.e., relating white and black people with positive words) positive D-SIP scores reflect faster “Yes” than “No” responses, while negative scores reflect faster “No” than “Yes” responses. For negative trial types (i.e., relating black and white people with negative words) positive D-SIP scores indicate a faster “No” than “Yes” response and negative D-SIP scores are showing a faster “Yes” rather than “No” response. 95% confidence interval error bars have been included. Error bars that cross the zero mark indicate a statistically neutral attitude; those that do not cross zero indicate a significant positive or negative attitude.

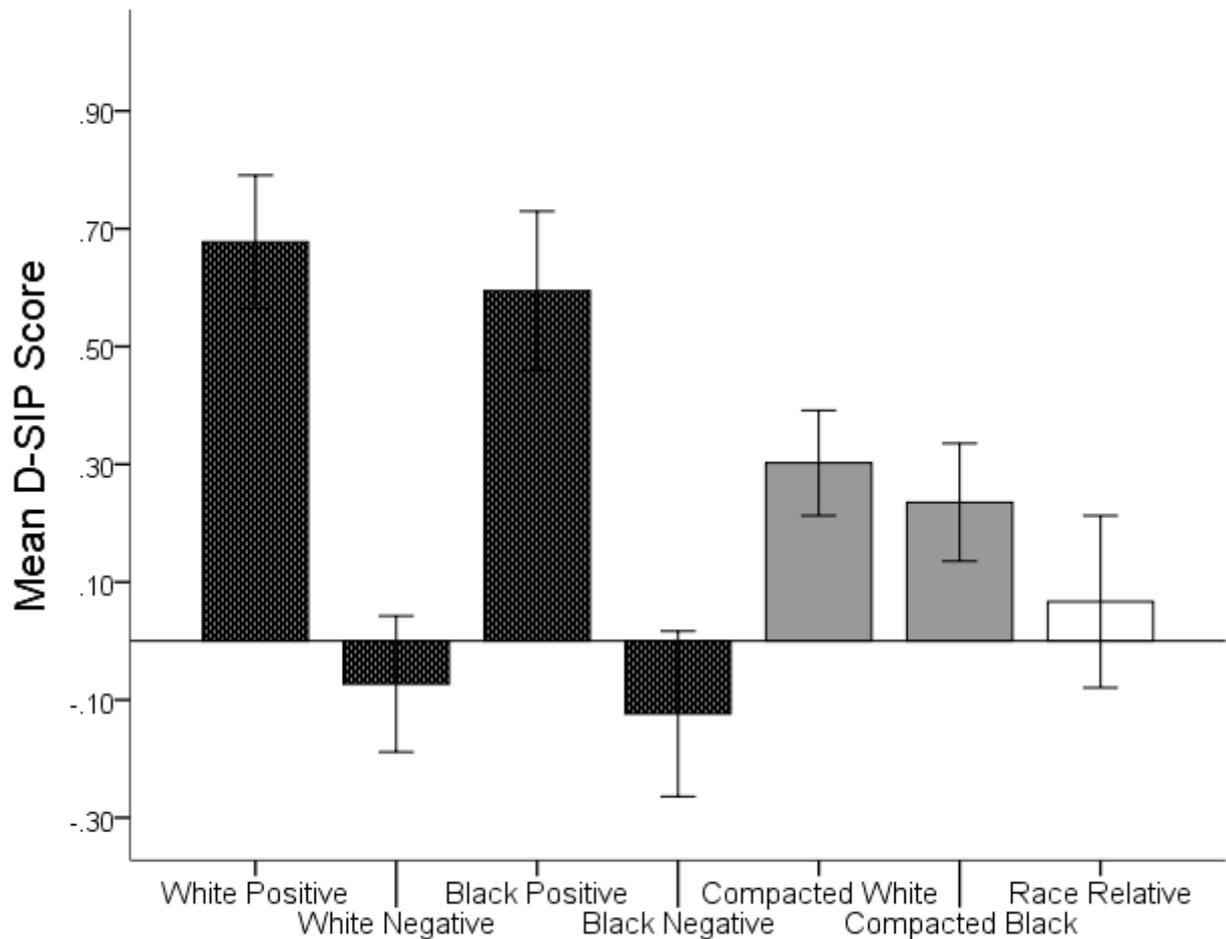


Figure 6.4: Post threat primes Race SIP: Mean D-SIP scores for the four individual trial types, the compacted trial types and the relative/overall D-SIP results are shown. 95% confidence interval error bars have been included.

A number of one-sample t-tests were conducted on the SIP following the threat primes to verify which of the absolute trial types scores, including the relative/overall D-SIP score, differed significantly from zero. As in the baseline findings, the White-Positive, Black-Positive, compacted White and compacted Black scores were significantly above zero, $t_s > 4.75$, $p < .001$. The White-Negative, $t(64) = 1.26$, $p > .250$, and Black-Negative, $t(64) = 1.76$, $p = .083$ trial types did not differ from zero. In contrast to the baseline findings, the relative D-SIP score did not differ from zero, $t(64) = .91$, $p > .365$, which could be taken to indicate a neutral attitude towards black people relative to white people.

To test the internal reliability in the SIP, separate D-SIP scores were calculated for odd and even trials. The correlation between scores on the odd and even trials was significant, $r = .276$, $n = 65$, $p = .026$ for the baseline SIP and, $r = .273$, $n = 65$, $p = .028$, for the SIP after the threat primes. The baseline overall D-SIP score and the priming D-SIP score were correlated to assess the test-retest reliability across the two SIPs. The correlation was significant, $r(65) = .319$, $p = .01$, indicating that the SIP appears to be stable and reliable over time.

QA and QB results: Participants completed both the bipolar explicit measure and the unipolar feeling thermometers within the experiment. Using the two unipolar feeling thermometers, a relative score was calculated for each participant by subtracting their black feeling thermometer score from their white feeling thermometer score. This relative unipolar score was calculated to match the bipolar explicit (relative) score. The relative baseline explicit scores and the relative post-threat primes explicit scores were entered into a mixed 2×2 ANOVA, with time as the within-subject factor (baseline explicit vs. post-primes explicit) and condition (disease & terror) as the between-subject factor. The threat primes did not have any effect on relative explicit attitudes ($F_s < .58$, $p_s > .45$). A one-sample t-test on the relative explicit scores found that participants' scores were significantly higher than neutral indicating a stronger anti-black/pro-white bias, at the baseline, $t(62) = 4.71$, $p < .001$, and following the threat primes, $t(64) = 4.51$, $p < .001$.

The pre- and post-threat primes life history questionnaire assessed life strategies used by both white and black individuals. A $2 \times 2 \times 2$ mixed ANOVA was used to test the effects of race (white and black), time (pre- and post-priming threat) and priming threat (terrorism and disease). Both race and time were within-subject factors and priming threat was a between-subject factor. There was a main effect of race, $F(1, 63) = 9.11$, $p < .01$, $\eta^2 = .13$, with white people being unexpectedly viewed as having a faster life history (i.e., more impulsive, more sexually promiscuous, etc.; $M = 4.14$) than black people ($M = 3.88$). There was also a

significant interaction between race and time, $F(1, 63) = 36.76, p < .001, \eta p^2 = .37$; such that white people were viewed as having a faster life history than black people at baseline, $t(64) = 5.35, p < .001$, while after the threat primes there was no difference between the two races, $t(64) = .31, p > .05$, (see Figure 6.5). No other significant effects were found, $F_s < .2.84, p_s > .10$.

Implicit - Explicit Correlations: Table 6.2 shows a correlation matrix of data obtained from the explicit White and Black feeling thermometers, the relative feeling thermometer and the single item bipolar explicit measure. All these scores were correlated with the scores from the compacted White trial type, compacted Black trial types and relative overall D-SIP score.

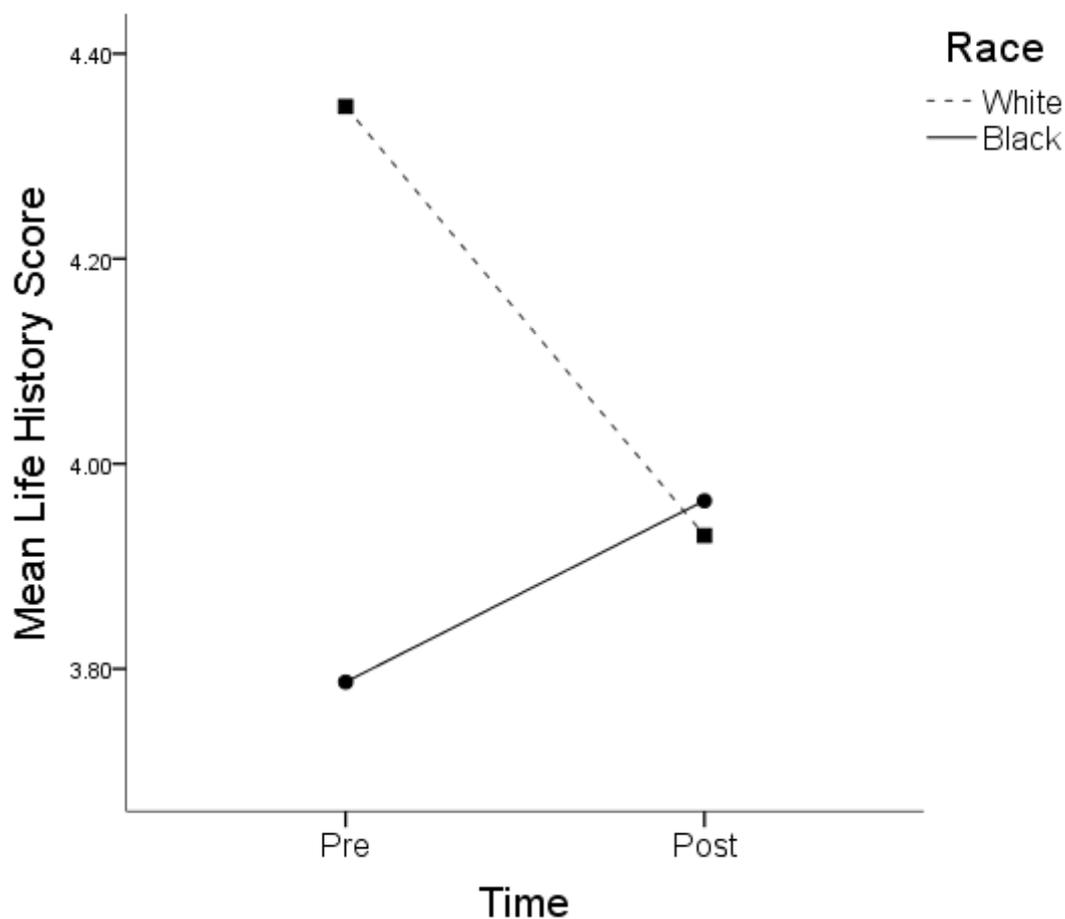


Figure 6.5: Interaction between race and time for participants' life history scores.

Table 6.2: Matrix of correlations between implicit and explicit measures

	SIP_White	SIP_Black	SIP_Therm	SIP_Bipolar	Therm_White	Therm_Black	Explicit_Therm	Explicit_Bipolar
SIP_White	—	0.032	0.661***	0.272*	0.129	-0.108	0.338**	0.174
SIP_Black		—	-0.729***	-0.220†	0.127	0.234†	-0.208	-0.260*
SIP_Therm			—	0.352**	-0.008	-0.245†	0.378**	0.314*
SIP_Bipolar				—	-0.124	-0.179	0.125	0.229†
Therm_White					—	0.777***	0.052	-0.114
Therm_Black						—	-0.588**	-0.534***
Explicit_Therm							—	0.701***
Explicit_Bipolar								—

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Note: Scores in bold indicate the most important correlations measuring the same construct. SIP_White and SIP_Black are the respective compacted SIP scores. SIP_Therm (Therm = Feeling Thermometer) are the relative SIP scores carried out during the same phase as the feeling thermometers. SIP_Bipolar are the relative SIP scores carried out during the same phase as the bipolar scale. Therm_White & Therm_Black are the respective feeling thermometer scores. Explicit_Therm are relative scores calculated using the Therm_White and Therm_Black scores. Explicit_Bipolar are the bipolar scale scores.

Of most relevance, the scores from relative feeling thermometer significantly correlated with the scores from the SIP (SIP_Therm). Also, the scores from explicit bipolar measure marginally correlated with the scores from the SIP (SIP_Bipolar). When the scores from the compacted White (SIP_White) and compacted Black (SIP_Black) trial types were correlated with the White and Black feeling thermometers, no significant correlations were found. However, both the compacted SIP scores showed a positive correlation with the feeling thermometers, consistent with the suggestion that the implicit and explicit measures were measuring racial biases similarly. To further address the usefulness of the SIP's absolute measurement, another correlation was performed (see Chapter 4 for more details). Each individual's White SIP score and explicit White feeling thermometer score and the same individual's Black SIP score and Black feeling thermometer score were included in the same correlation. For this analysis, we found a significant correlation, $r = .211$, $n = 125$, $p = .02$.

Based on the previous correlation, two additional correlation scores are reported. Each of these correlations report scores that alternate between using White implicit and explicit scores to Black implicit and explicit scores across each participant. For example, analysis 1 (cut 1) will include: participant 1 = White, participant 2 = Black, participant 3 = White, etc. While analysis 2 (cut 2) will include: participant 1 = Black, participant 2 = White, participant 3 = Black, etc. For cut 1, the correlation was in the same direction but not significant, $r = .108$, $n = 63$, $p = .298$, whereas for cut 2 a significant correlation was shown, $r = .320$, $n = 63$, $p = .011$

Discussion

The current experiment aimed to build on the preliminary evidence described in Chapter 2 (see Appendix 2), that females' implicit and explicit anti-black/pro-white scores do not increase following disease or terrorism primes. A new implicit measure called the Simple Implicit

Procedure (SIP) was used and allowed a within-subjects experimental design to be used because it does not have a practice effect limitation like the IAT (see Chapter 4, Study 3). The most important finding from this experiment is that exposure to threatening primes (diseases or terrorism) did not increase females' anti-black/pro-white bias on either implicit (using the SIP) or explicit (using the bipolar measure and the feeling thermometers) measures. These findings provide further preliminary evidence that females do not respond negatively towards out-groups when experiencing disease and terror threats.

Although the scores from the overall D-SIP technically showed neutral implicit attitudes towards black and white people following the threatening primes, it should be noted that at baseline, females had an implicit anti-black/pro-white bias. Using the explicit bipolar and feeling thermometer, anti-black/pro-white biases remained stable, pre- and post- the threat primes. These findings indicate that females initially had implicit and explicit pro-in-group/anti-outgroup biases. These biases essentially remained stable for explicit attitudes but slightly reduced (although not significantly) for implicit attitudes, following the terrorism and disease threats. How can we explain the current results, particularly at the implicit level?

To help answer this, Taylor et al.'s (2000) tending and befriending hypothesis but particularly the befriending aspect could be used. For example, following a stressful situation, females' affiliative behaviours increase (in order to form new alliances), ameliorating their stress response during threatening situations. In contrast, males are more likely to exhibit a fight or flight response when they experience stressful situations and to appraise out-groups more negatively compared to females (see Klavina et al., 2011; McDonald et al., 2012). Therefore, it is likely that threatening primes are especially likely to heighten males' prejudice towards out-groups.

The current experiment did not determine when females' anti-black/pro-white biases will remain stable, increase or decrease over time. A factor explaining the stability in females' implicit bias could be due to the disease and terror threats being very general (i.e., showing images of males, females, children and adults with black, brown and white skin tone). If the threats were described and shown as coming from a specific group, increased prejudice in females towards the targeted group is more likely. For example, heightened prejudice against black people is likely to be observed if females are primed with exaggerated threats that specifically black people harbour deadly diseases or are extremely likely to kill or rape women (e.g., Navarrete, Fessler, Fleischman, & Geyer, 2009)

Correlations were shown between the implicit and explicit measures which indicates that the SIP is capable of accurately measuring meaningful constructs. Previous research has highlighted the fact that correlations are often not found between implicit and explicit measures, particularly for socially sensitive topics like racial prejudice (Fazio & Olson, 2003). Therefore, perhaps simply asking participants to avoid constantly selecting neutral, as was done in this experiment, could be sufficient to reduce socially desirable responding and increase the correlations between the implicit and explicit measures. In addition, there was appropriate internal reliability in the SIP, pre- and post-the threat primes, and the internal reliability in the current experiment was similar to the internal reliability for SIPs that measured several different constructs in one sitting (see Chapter 4). It should be noted that the SIPs used in the current experiment had less trials than the SIPs used in Chapter 4 (176 trials vs. 288 trials) which could have weakened the internal reliability. Other methods such as increasing the length of each block of trials should be tested to address whether this factor increases the internal reliability of the SIP.

Unexpectedly (see Williams et al., 2016), females initially viewed white people as having a faster life history than black people. This is arguably a more negative bias towards white people relative to black because having a faster life history implies more impulsivity, sexual risk taking and less investment in children and education. Since Williams et al.'s (2016) research was carried out in the U.S and the current experiment was conducted in the UK, variations in exposure to different races from deprived areas might explain the current findings. For example, perhaps students from the UK have more encounters with white rather than black people from “desperate” environments compared to students in the US. More research is needed to determine whether other races or cultural minorities (e.g., travellers/gypsies) in other countries are stereotyped as using fast life history strategies.

We observed that following the threatening primes, views about the life history of white and black people converged such that the participants saw each race as having a similar life history strategy. This finding could again be explained by Taylor et al.'s (2000) tending and befriending hypothesis, according to which stressful situations may induce females to become equally accepting/critical towards in and out-groups. This strategy could increase their potential of gaining new affiliations. The explicit bipolar and thermometer scores remained stable (anti-black/pro-white) following the threatening primes but this could be due to the participants aiming to appear consistent across the two time points, and is consistent with previous research (Lai et al., 2014, 2016).

Limitations and conclusion

A similar experiment to the current one (i.e., using the SIP and a within-subject design) should be conducted with male participants to provide confirming evidence that disease primes increase males' prejudice towards out-groups. Measuring life history biases among males would

also be revealing, as a divergence rather than a convergence in preferential biases towards their in-group and increased prejudice towards their out-group would be expected, based on the male warrior hypothesis (McDonald et al., 2012). In addition, collecting data from black males and females is important to determine whether increased prejudice towards white people occurs for only black males experiencing disease threats.

A further limitation of the study is that as the SIP is relatively new, potential flaws or biases due to the procedural set-up cannot be ruled out. Chapter 4 (Study 2) explains a method to remove the affirming bias in the SIP (i.e., participants express positive biases when connecting any attitude object with positive words). The affirming bias likely resulted in an inflated pro-white and pro-black bias on the positive target words trial types. However, this potential problem only affects the absolute results while the relative overall D-SIP results remain accurate, because the affirming bias essentially gets cancelled out. Nevertheless, the current experiment aimed to measure changes in biases following threatening primes when the extent of an individual's positivity bias had already been determined (baseline SIP). Therefore, changes in the absolute SIP score following the primes would also be accurate at measuring increases or decrease in racial biases towards black and white people. Unfortunately, no changes occurred and therefore on this occasion the unique benefit of the SIP cannot be emphasised.

Another potential explanation for the stability of females' implicit bias over time is that the SIP, unlike the IAT and most other implicit measures, is not influenced by context or situational factors. Of note, when an intervention/prime does not impact/change the attribute being measured, the absence of an effect of the intervention/prime on the SIP says nothing about the validity of the SIP (De Houwer et al., 2009). If the SIP is impervious to situational influences, then it would not be suited for the experimental design used here. This factor could also explain females' implicit

biases remaining stable over time. More optimistically, if the SIP is not influenced by context effects then it could provide a context independent assessment of a person's "true" implicit social representations. Therefore, the SIP may be addressing more long term early socialisation biases which were initially what implicit measures were believed to be assessing (e.g., Rudman, 2004; Wilson et al., 2000). More basic research is needed to test whether scores on the SIP are influenced by contextual effects.

Overall, this experiment builds on the preliminary evidence that females do not express increased prejudice both implicitly and explicitly towards racial out-groups after experiencing disease or terrorism primes. Although stronger explicit and implicit biases in favour of white people relative to black people were shown at the baseline, neutral attitudes were observed implicitly following the threatening primes, possibly due to an unconscious/natural befriending response in females that is evoked under stressful situations. Perhaps, when a threatening situation or event is encountered, females respond in a more composed and fair manner towards racial out-groups. This finding could have implications for those making important decisions (e.g., political leaders, police, army personnel) that can greatly impact groups/individuals associated with terrorism or disease outbreaks. Developing a better understanding of how males and females respond towards out-groups when encountering stressful and threatening situations could assist in reducing conflicts and discrimination.

Chapter 7: Concluding remarks and future directions

As reported in five empirical chapters (Chapter 2-6), I used measures of implicit attitudes to advance knowledge in implicit social cognition research and provide evidence supporting parasite-stress theory. Chapter 1 introduced the historical origins of implicit social cognition and described how the Implicit Association Test (IAT: Greenwald, Nosek, & Banaji, 2003) greatly accelerated and expanded the type of research carried out by social psychologists relating specifically to the areas of attitudes, stereotypes and the self. Chapter 2 introduced parasite-stress theory (PST; Thornhill & Fincher, 2014) which was heavily influenced by Behavioural Immune System (BIS) research (Murray & Schaller, 2016) and which aimed to explain the influence that infectious diseases can have on intergroup relations and societal values or belief systems.

In Chapter 2, secondary data made available by Project Implicit were used to test a hypothesis based on PST - that residents of states across the US and countries across the world with higher disease rates will show increased prejudice towards racial out-groups, relative to US states and countries with lower disease rates. The Project Implicit dataset provided thousands of responses from individuals measuring implicit (using the IAT) and explicit (using questionnaires) attitudes including important demographic information of the participants (e.g., their age, gender, education, political ideology and religious belief). Even when controlling for these individual characteristics and several other state or country level factors often used in prejudice research, both black and white respondents across the US and white respondents across the world showed increased anti-out-group/pro-in-group biases in regions that had higher disease rates.

The experimental results from Study 3 in Chapter 2 were also used to test a hypothesis based on PST- that being primed with images of diseases will increase prejudice towards racial out-groups compared to a control condition (i.e., being primed with images of furniture and buildings). A terror threat prime condition was also included to determine whether priming by any

threat caused an increase in prejudice towards racial out-groups or whether any increase was selective for disease-related threat primes.

Considering the explicit attitudes of the white participants, both the disease and terrorism primes increased anti-black/pro-white prejudice compared to the control primes. However, for the implicit attitudes of the same participants, only the disease primes increased anti-black/pro-white prejudice compared with the control and terror threat conditions. This finding is in line with other studies in which disease primes had a unique effect of increasing implicit prejudice compared to primes depicting accident-related images (for review see Murray & Schaller, 2016).

Of note, males consistently expressed more prejudice towards out-groups than females. When the experimental data were analysed separately for males and females, only males showed increased anti-black/pro-white prejudice both on implicit and explicit measures. Although an interaction did not occur between the priming conditions and gender, further research was needed to determine whether disease primes increase white females anti-black/pro-white prejudice. Answering this question was one of the aims of the work presented in Chapter 6.

A major limitation of the findings from Chapter 2 was that the IAT can only measure attitudes in a relative way, (i.e., the IAT measures implicit attitudes towards black people *relative* to white people). Consequently, use of an absolute implicit measure (i.e., one that can measure implicit attitudes towards black and white people separately) such as the Implicit Relational Assessment Procedure, (IRAP; Barnes-Holmes, et al., 2010) would enable a deeper understanding of the mechanisms driving implicit prejudice to be gained (e.g., do participants have a strong pro-in-group preference, a strong prejudice towards racial out-groups, or some combination of these two effects?).

The IRAP is a relatively new measure of absolute implicit attitudes and although it has been used a number of times with clinical or vulnerable samples (Vahey, et al., 2015), there is a dearth of basic research scrutinising its validity. The work presented in Chapter 3 aimed to test whether some of the typical findings when using the IRAP (e.g., a normative sample had a positive bias towards death and dying; Hussey, et al., 2015) might be explained not by implicit attitudes/biases, but instead by other cognitive biases not related to implicit attitudes. Specifically, it was hypothesised that when participants have to store a positive and negative statement in memory (as they do in the IRAP), they will use a cognitive heuristic to simplify the task and store only one of the statements in memory (i.e., almost always the positive statement)

Positivity biases are inherent to the English language (Dodds et al., 2015; Matthews & Dylman, 2014) and the human propensity toward optimism (e.g., Peterson, 2000) were used to explain this positive framing bias (PFB) when participants used the IRAP. The PFB results in scores in the IRAP overestimating the positivity of implicit attitudes towards any object (e.g., non-words, insect, fat people). Importantly, the PFB could easily be manipulated/reversed by instructing participants to focus and store the negative statement in memory. This finding further questions the validity of the IRAP's estimates of absolute implicit attitudes, because the estimates can be influenced by how the instructions are presented. Moreover, completely controlling for instructional presentations at the individual level is not possible with the IRAP.

These flaws with the IRAP would have made the interpretations of future results more challenging. Therefore, the IRAP was not used again to examine absolute racial biases when individuals had been primed with diseases. However, I did not want to abandon the core features of how the IRAP measures implicit biases, particularly the affirming and negating aspect of the task. Compared with implicit measures that use a sorting task (e.g., IAT, GNAT and SC-IAT), the

affirming and negating response options allow for a more in-depth understanding of the mechanisms driving implicit biases. In addition, implicit measures that use a sorting task can only measure the strength of the association between items, while tasks with affirming and negating response options allow researchers to better understand how two concepts (i.e., “I” and “good”) are related together (i.e., “I am good” vs “I want to be good”).

In Chapter 4, I introduced a new implicit measure named the Simple Implicit Procedure (SIP) which also uses affirming and negating response options. The SIP was developed to remove the PFB apparent in the IRAP by requiring participants to store only one statement (either positive or negative) in memory during each block of trials. This new procedure was successful at removing the PFB. However, another response bias became apparent that influences the SIP’s absolute results and would be likely to influence any other task that uses affirming and negating response options. The bias can be described as an affirming bias and results from participants being faster to press “Yes” than “No”, particularly on positive word trials. The PFB in the IRAP is also influenced by this affirming bias but determining the magnitude of each influence (i.e., PFB and affirming bias) on estimates of implicit attitudes is difficult.

The affirming bias in the SIP results in an over-inflation of the estimates of implicit biases towards any object. I developed a method to eliminate the affirming bias in the SIP by measuring participants’ response biases towards non-words. If neutral scores are not shown when using the non-words SIP then these scores would be indicating a response bias (i.e., normally participants showed an affirming bias, rather than neutral scores or a negating bias). Then each participant’s response bias on the non-word SIP was used to correct/transform their scores on other SIP tasks they completed. When this method was used, the scores on the absolute trial types (i.e., separate score for flowers and insects) in the SIP were in line with expectations (e.g., strong negative biases

towards criminals and insects and positive biases towards carers and flowers). The relative results in the SIP were always in line with expectations and previous IAT results because the affirming biases get cancelled out when the scores from the relevant conditions are combined.

The measured internal reliability of the SIP suggests it is suitable to measure individual differences in implicit bias scores. The implicit and explicit relative/bipolar scores positively correlated and the directionality of the absolute scores were always positively related for non-socially sensitive topics. In a socially sensitive domain (i.e., relating male and female names with career and family words) it was found that the scores from the implicit and explicit measures negatively correlated and the implicit measure was better than the explicit measure at predicting responses on a scale addressing those most/least likely to uphold and maintain gender inequalities.

In Chapter 5, I confirmed that participants are strongly biased towards responding faster to affirming words than negating words and that they are also faster to associate positive with affirming and negative with negating words than the reverse associations. Furthermore, the work presented in this chapter showed that participants are faster overall to sort positive and affirming than negative and negating words. These findings clearly show why the affirming bias occurs in the SIP.

Chapter 5 also examined the influence that word type (i.e., nouns, verbs and adjectives) can have on processing RTs in implicit measures. Using secondary data as well as three experiments that used words that were controlled for on a number of important dimensions, it was shown that participants responded faster to negative nouns than positive nouns, while for verbs and adjectives the reverse occurred. Participants also showed faster RTs when responding to negative images than positive images. Therefore, it appears that nouns are processed in a similar fashion to images in valence judgement RT tasks. Although it was not the scope of this thesis to

determine how exactly these word type biases could influence outcomes on implicit measures, it suggests that the choice of words used will have potentially important influences in any RT task that uses words, especially for absolute implicit measures.

As described and shown empirically in Chapter 4, another major advantage that the SIP has over sorting tasks like the IAT is that it is not limited by practice/experience effects. This benefit allows researchers to use a within-subject design to test implicit biases pre-and post an intervention. Preliminary evidence presented in Chapter 2 (see Appendix 2) indicated that, compared to females, white males expressed stronger prejudice towards black people when primed with diseases. The experiment in Chapter 2 used a between-subject design and was also limited by the IAT being only able to show relative rather than absolute results. Consequently, in Chapter 6, the SIP was used which allowed for absolute attitudes to be measured. It also enabled a within-subject design to be used, pre-and post-threatening primes, to measure the influence they had on implicit anti-black and pro-white attitudes.

Using white female participants, the results showed using the SIP and explicit measures, that neither the disease nor the terror priming condition increased anti-black/pro-white prejudice. These results were in line with previous findings (Chapter 2, Appendix 2). However, not enough white males completed the experiment to enable testing of previous findings that disease primes increase males' implicit and explicit anti-black/pro-white prejudice. Further research will need to use the SIP to explore the influence of disease primes on males' prejudice. If an increase in implicit prejudice is not observed for males when using the SIP, then more basic research addressing whether the SIP is immune to context effects (i.e., disease primes) will be required. If this basic research shows that the SIP measures stable, long-term, early socialisation implicit biases that are

not influenced by contextual influences, then the SIP would not be suitable for measuring immediate changes in implicit biases.

Prior to this thesis, the only implicit measure that has been scrutinised for practice/experience was the IAT but it is likely that other implicit measures also suffer from this limitation. I am currently working on a project testing practice effects in other implicit measures. Preliminary evidence indicates that the SC-IAT is also limited by practice effects but no practice effects have yet been seen with the GNAT. The major difference between the IAT/SC-IAT and the GNAT is that the GNAT has only one response option (space bar), while in IATs there are always two. In the GNAT, the space bar response is pressed when the stimulus presented at the centre of the screen matches one of the two categories at the top of the screen. If the stimulus does not match either of the two categories then participants make no response. Perhaps participants are encoding the space bar response as a “Yes” and the no space bar response as “No”. If participants are encoding the space bar and no space bar response in this way, then the GNAT might have more similarities with the SIP’s affirming and negating task than the sorting task in the IAT. Therefore, having both affirming and negating response options within a task might be the reason why practice/experience effects do not occur in the SIP and the GNAT.

More research is needed to test the influence that the biases described in this thesis (e.g., PFB, affirming bias and word type biases) might have on other implicit measures, particularly absolute implicit measures like the SC-IAT and the GNAT. For example, other preliminary evidence I gathered showed that participants had a positive bias towards neutral furniture items in the SC-IAT and the GNAT, while the same participants had a negative bias towards non-words. The negative bias toward non-words was explained by the polarity correspondence (i.e., “non” and “negative”; see Proctor & Cho, 2006) or semantic similarities between the category label “Non-

Word” and “Negative”. If participants’ implicit biases towards non-words can be influenced by simply changing the category labels to “Special-Words” (a positive bias is expected) or “Stupid-Words” (a negative bias is expected), estimates of implicit biases can be easily manipulated by changing the category labels. Of note, there was always positivity bias towards furniture and non-words when the SIP was used. This finding suggests that the category labels used in the SIP are less likely to shift implicit biases from positive to negative estimates.

Although much more basic research is needed before the SIP can be used with vulnerable populations, initial findings look promising. If the SIP is capable of measuring changes in participants’ implicit biases, then it could be used to test the effectiveness of intensive therapies (e.g., testing the long-term success of drug rehabilitation and determining which patients are most likely to relapse). The SIP could initiate a new wave of individualised clinical therapies aimed at tackling specific biases directly because of its ability to measure absolute implicit biases.

The next step in the development of the SIP would be to develop it to run online. Developing the IAT to run online through Project Implicit was one of the major factors for its success. The ease with which people can complete the SIP would make transitioning the SIP from the laboratory to the internet more manageable. An online version of the SIP would make data collection much more efficient and cost effective, albeit with the potential cost of noisier data due to an inability to control for distraction in the participants’ environment and variations in individual’s computers.

Another path worth exploring is the potential of using only images in measures of implicit attitudes to estimate individuals’ implicit biases. This procedure would remove the problems that result from using language/verbal stimuli (i.e., positivity bias, affirming bias and the influence of nouns, verbs and adjectives). We have tested the influence of responding to positive and negative

images which are presented at the top and bottom of the screen. Initial evidence indicates that people are faster to respond to negative/threatening images at the top of the screen than to positive or non-threatening images, while at the bottom of the screen, this difference is less pronounced and sometimes reverses. However, as shown in Chapter 5 there consistently appears to be a negativity bias with faster processing RTs for negative images compared to positive ones. Rather than measuring implicit biases, using only images in valence judgement RT tasks are likely measuring how salient or socially relevant the image is.

Relating to PST, there are still many research questions that can be answered using Project Implicit datasets. This thesis has only described data from the Race IAT, but I have also obtained similar findings (i.e., increased prejudice in regions with higher disease rates) on the Old-Young IAT, Fat-Thin IAT, Straight-Homosexual IAT, Abled-Disables IAT, White-Native American IAT, and the White-Asian IAT. A caveat is that the pattern of increased prejudice in regions with higher disease rates towards out-groups did not always survive the inclusion of control factors. I have also shown that across the US, even when controlling for several important factors, respondents in regions with higher disease rates, have higher religious, conservative, right-wing and socially dominant worldviews. All these findings are in line with PST. There are also many personality factors (e.g., Big 5, social desirability, need for structure/closure) collected through Project Implicit that could be used to test many predictions from PST.

To conclude, the work presented throughout this thesis has identified various biases (i.e., PFB, affirming biases, biases towards positive and negative, nouns, verbs and adjectives) that can greatly influence the results obtained from measures of implicit attitudes. I developed and empirically investigated a new method of measuring implicit attitudes (the SIP) which did not suffer from the PFB. Although an affirming bias occurs in the SIP, the bias can easily be measured

and removed for each participant, resulting in an accurate estimate of an individual's implicit biases. The findings in this thesis also indicate the strong influence that disease rates can have on increasing prejudice towards racial out-groups, particularly for males. Finally, the work presented in this thesis provides the foundations for devising strategies aimed at understanding and ameliorating the precise mechanisms driving prejudice by using the SIP. With this improved understanding, the SIP's results could be used to impact policy aiming to provide achievable methods of reducing or solving intergroup tensions.

References

- Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology, 69*(2), 334–345. <https://doi.org/10.1037/0021-9010.69.2.334>
- Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814–823. <https://doi.org/10.1111/j.1467-9280.2006.01787.x>
- Agosta, S., & Sartori, G. (2013). The autobiographical IAT: a review. *Frontiers in Psychology, 4*. <https://doi.org/10.3389/fpsyg.2013.00519>
- Alba, R., Rumbaut, R. G., & Marotz, K. (2005). A Distorted Nation: Perceptions of Racial/Ethnic Group Sizes and Attitudes toward Immigrants and Other Minorities. *Social Forces, 84*(2), 901–919.
- Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., & Wacziarg, R. (2003). *Fractionalization* (Working Paper No. 9411). National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w9411>
- Algom, D., Chajut, E., & Lev, S. (2004). A rational look at the emotional stroop phenomenon: a generic slowdown, not a stroop effect. *Journal of Experimental Psychology. General, 133*(3), 323–338. <https://doi.org/10.1037/0096-3445.133.3.323>
- Allport, G. W. (1935). Attitudes. In C. Murchison, *A handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press. Retrieved from <http://noneedto.read.blogsport.de/images/allportattitudes.pdf>

- Back, M. D., Schmukle, S. C., & Egloff, B. (2005). Measuring task-switching ability in the Implicit Association Test. *Experimental Psychology*, *52*(3), 167–179.
<https://doi.org/10.1027/1618-3169.52.3.167>
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology. General*, *133*(2), 283–316. <https://doi.org/10.1037/0096-3445.133.2.283>
- Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger & I. N. Nairne, *The nature of remembering: Essays in honor of Robert G. Crowder*. (pp. 117–149). Washington, D.C.: APA. Retrieved from
http://www.people.fas.harvard.edu/~banaji/research/publications/articles/2001_Banaji_HLRoediger.pdf
- Banaji, Mahzarin R., & Greenwald, A. G. (2016). *Blindspot: Hidden Biases of Good People*. New York, NY: Bantam.
- Banaji, Mahzarin R., Nosek, B. A., & Greenwald, A. G. (2004). No place for nostalgia in science: A response to Arkes and Tetlock. *Psychological Inquiry*, *15*(4), 279–289.
- Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668–688. <https://doi.org/10.3758/s13428-013-0410-6>
- Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The Sorting Paired Features Task: A Measure of Association Strengths. *Experimental Psychology*, *56*(5), 329–343.
<https://doi.org/10.1027/1618-3169.56.5.329>
- Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual Moderation of Racial Bias: The Impact of Social Roles on Controlled and Automatically Activated

Attitudes. *Journal of Personality and Social Psychology*, 87(1), 5–22.

<https://doi.org/10.1037/0022-3514.87.1.5>

Bargh, J. A. (1994). The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control in Social Cognition. In R. Wyer & T. Srull (Eds.), *Handbook of Social Cognition*. Mahwah, NJ, US: Lawrence Erlbaum.

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist*, 32(7), 169–177.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, 60(3), 527–542.

Barnes-Holmes, D., Murphy, A., Barnes-Holmes, Y., & Stewart, I. (2010). The Implicit Relational Assessment Procedure: Exploring the impact of private versus public contexts and the response latency criterion on pro-white and anti-black stereotyping among white Irish individuals. *The Psychological Record*, 60(1), 57–66.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), 323–370. <https://doi.org/10.1037//1089-2680.5.4.323>

Baumeister, R. F., & Bushman, B. J. (2010). *Social psychology and human nature*. Boston, MA, US: Cengage Learning.

- Benson, P. L., Karabenick, S. A., & Lerner, R. M. (1976). Pretty pleases: The effects of physical attractiveness, race, and sex on receiving help. *Journal of Experimental Social Psychology, 12*(5), 409–415. [https://doi.org/10.1016/0022-1031\(76\)90073-1](https://doi.org/10.1016/0022-1031(76)90073-1)
- Bessenoff, G. R., & Sherman, J. W. (2000). Automatic and Controlled Components of Prejudice Toward Fat People: Evaluation Versus Stereotype Activation. *Social Cognition, 18*(4), 329–353. <https://doi.org/10.1521/soco.2000.18.4.329>
- Blackstone, A. (2017). *Principles of Sociological Inquiry: Qualitative and Quantitative Methods*. Retrieved from http://catalog.flatworldknowledge.com/bookhub/reader/3585?e=blackstone_1.0-ch11_s02
- Blair, I. V. (2002). The Malleability of Automatic Stereotypes and Prejudice. *Personality and Social Psychology Review, 6*(3), 242–261. https://doi.org/10.1207/S15327957PSPR0603_8
- Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*(1), 27–41. <https://doi.org/10.1037/0003-066X.61.1.27>
- Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology, 38*(6), 977–997. <https://doi.org/10.1002/ejsp.487>
- Boroditsky, L. (2011). How language shapes thought. *Scientific American, 304*(2), 62–65.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>

- Bosson, J. K., Swann Jr., W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, *79*(4), 631–643. <https://doi.org/10.1037/0022-3514.79.4.631>
- Briones, E. M., & Benham, G. (2017). An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples. *Behavior Research Methods*, *49*(1), 320–334. <https://doi.org/10.3758/s13428-016-0710-8>
- Brochu, P. M., Gawronski, B., & Esses, V. M. (2011). The integrative prejudice framework and different forms of weight prejudice: An analysis and expansion. *Group Processes & Intergroup Relations*, *14*(3), 429–444. <https://doi.org/10.1177/1368430210396520>
- Brown, G. D. A., Fincher, C. L., & Walasek, L. (2016). Personality, parasites, political attitudes, and cooperation: A model of how infection prevalence influences openness and social group formation. *Topics in Cognitive Science*, *8*(1), 98–117. <https://doi.org/10.1111/tops.12175>
- Brown, G. D. A., Walasek, L., & Mullett, T. L. (2016). *Parasites, politics, and social groups: Changing environmental origins of ideology*. Manuscript submitted for publication.
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. <https://doi.org/10.3758/BRM.41.4.977>
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. <https://doi.org/10.3758/s13428-013-0403-5>

- Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review, 1*(1), 3–25. https://doi.org/10.1207/s15327957pspr0101_2
- Calanchini, J., Sherman, J. W., Klauer, K. C., & Lai, C. K. (2014). Attitudinal and non-attitudinal components of IAT performance. *Personality and Social Psychology Bulletin, 40*(10). <https://doi.org/10.1177/0146167214540723>
- Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review, 16*(4), 330–350. <https://doi.org/10.1177/1088868312440047>
- Carnaghi, A., Maass, A., Gresta, S., Bianchi, M., Cadinu, M., & Arcuri, L. (2008). Nomina sunt omina: On the inductive potential of nouns and adjectives in person perception. *Journal of Personality and Social Psychology, 94*(5), 839–859. <https://doi.org/10.1037/0022-3514.94.5.839>
- Carnevale, J. J., Fujita, K., Han, H. A., & Amit, E. (2015). Immersion versus transcendence: How pictures and words impact evaluative associations assessed by the Implicit Association Test. *Social Psychological and Personality Science, 6*(1), 92–100. <https://doi.org/10.1177/1948550614546050>
- Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology* (Vol. xiii). New York, NY: Guilford Press.
- Chaplin, G. (2004). Geographic distribution of environmental factors influencing human skin coloration. *American Journal of Physical Anthropology, 125*(3), 292–302. <https://doi.org/10.1002/ajpa.10263>

- Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review*, 76(4), 387–404.
- Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, 89(4), 469–487. <https://doi.org/10.1037/0022-3514.89.4.469>
- Crandall, C. S., & Moriarty, D. (1995). Physical illness stigma and social rejection. *The British Journal of Social Psychology*, 34, 67–83.
- Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review. *Psychological Bulletin*, 87(3), 546–563. <https://doi.org/10.1037/0033-2909.87.3.546>
- 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.
- De Houwer, J. (2003a). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219–244). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- De Houwer, J. (2003b). The extrinsic affective Simon task. *Experimental Psychology*, 50(2), 77–85.
- De Houwer, J. (2006). What Are Implicit Measures and Why Are We Using Them? In *Handbook of implicit cognition and addiction* (pp. 11–28). Thousand Oaks, CA, US: Sage Publications, Inc. <https://doi.org/10.4135/9781412976237.n2>

- De Houwer, J. (2009). Comparing measures of attitudes at the functional and procedural level: Analysis and implications. In R. H. Petty, R. H. Fazio, & P. Brinol, *Attitudes: Insights from the new implicit measures* (pp. 361–390). New York, NY: Psychology Press.
Retrieved from <http://users.ugent.be/~jdhouwer/bookpetty.pdf>
- De Houwer, J., & De Bruycker, E. (2007). The implicit association test outperforms the extrinsic affective Simon task as an implicit measure of inter-individual differences in attitudes. *British Journal of Social Psychology*, *46*(2), 401–421.
<https://doi.org/10.1348/014466606X130346>
- De Houwer, J., Geldof, T., & De Bruycker, E. (2005). The implicit association test as a general measure of similarity. *Canadian Journal of Experimental Psychology*, *59*(4), 228–239.
- De Houwer, J., Heider, N., Spruyt, A., Roets, A., & Hughes, S. (2015). The relational responding task: toward a new implicit measure of beliefs. *Frontiers in Psychology*, *6*.
<https://doi.org/10.3389/fpsyg.2015.00319>
- De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*, 176–193.
- De Houwer, J., & Moors, A. (2012). How to define and examine implicit processes. *Implicit and Explicit Processes in the Psychology of Science*, 183–198.
- De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368.
<https://doi.org/10.1037/a0014211>
- De Raedt, R., Schacht, R., Franck, E., & De Houwer, J. (2006). Self-esteem and depression revisited: Implicit positive self-esteem in depressed patients? *Behaviour Research and Therapy*, *44*(7), 1017–1028. <https://doi.org/10.1016/j.brat.2005.08.003>

- Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology, 91*(3), 385–405.
<https://doi.org/10.1037/0022-3514.91.3.385>
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5.
- 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, J. M. (1999). *Guns, germs, and steel: The fates of human societies*. New York, NY: Norton.
- Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., ... Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences, 112*(8), 2389–2394.
<https://doi.org/10.1073/pnas.1411678112>
- Doob, L. W. (1947). The behavior of attitudes. *Psychological Review, 54*(3), 135–156.
<https://doi.org/http://0-dx.doi.org.pugwash.lib.warwick.ac.uk/10.1037/h0058371>
- 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.
[https://doi.org/10.1016/0022-1031\(86\)90039-9](https://doi.org/10.1016/0022-1031(86)90039-9)
- Duncan, L. A., & Schaller, M. (2009). Prejudicial Attitudes toward older adults may be exaggerated when people feel vulnerable to infectious disease: Evidence and implications. *Analyses of Social Issues and Public Policy, 9*(1), 97–115.
<https://doi.org/10.1111/j.1530-2415.2009.01188.x>

- Duncan, L. A., Schaller, M., & Park, J. H. (2009). Perceived vulnerability to disease: Development and validation of a 15-item self-report instrument. *Personality and Individual Differences, 47*(6), 541–546. <https://doi.org/10.1016/j.paid.2009.05.001>
- 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. <https://doi.org/10.1177/0146167297233009>
- Eppig, C., Fincher, C. L., & Thornhill, R. (2010). Parasite prevalence and the worldwide distribution of cognitive ability. *Proceedings. Biological Sciences, 277*(1701), 3801–3808. <https://doi.org/10.1098/rspb.2010.0973>
- Eppig, C., Fincher, C. L., & Thornhill, R. (2011). Parasite prevalence and the distribution of intelligence among the states of the USA. *Intelligence, 39*(2), 155–160. <https://doi.org/10.1016/j.intell.2011.02.008>
- Estes, Z., & Adelman, J. S. (2008). Automatic vigilance for negative words in lexical decision and naming: comment on Larsen, Mercer, and Balota (2006). *Emotion, 8*(4), 441–444. <https://doi.org/10.1037/1528-3542.8.4.441>
- Faulkner, J., Schaller, M., Park, J. H., & Duncan, L. A. (2004). Evolved disease-avoidance mechanisms and contemporary xenophobic attitudes. *Group Processes & Intergroup Relations, 7*(4), 333–353. <https://doi.org/10.1177/1368430204046142>
- Fazio, R. H., & Olson, M. A. (2014). The MODE model: Attitude-behavior process as a function of motivation and opportunities. In J. W. Sherman, B. Gawronski, & Y. Trope, *Dual-Process Theories of the Social Mind*. New York, NY: Guilford Publications.

- Fazio, Russell H. (1990). Multiple processes by which attitudes guide behavior: The mode model as an integrative framework. *Advances in Experimental Social Psychology*, 23, 75–109.
[https://doi.org/10.1016/S0065-2601\(08\)60318-4](https://doi.org/10.1016/S0065-2601(08)60318-4)
- Fazio, Russell H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, 25(5), 603–637. <https://doi.org/10.1521/soco.2007.25.5.603>
- Fazio, Russell 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: Attitudes and Social Cognition*, 69(6), 1013–1027. <https://doi.org/http://dx.doi.org/10.1037/0022-3514.69.6.1013>
- Fazio, Russell H., & Olson, M. A. (2003). Implicit measures in social cognition. research: their meaning and use. *Annual Review of Psychology*, 54, 297–327.
<https://doi.org/10.1146/annurev.psych.54.101601.145225>
- Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2), 117–140.
<https://doi.org/10.1177/001872675400700202>
- Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology*, 27(4), 307–316.
https://doi.org/10.1207/s15324834basp2704_3
- Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the ‘I’, the ‘A’, and the ‘T’: A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, 17(1), 74–147.
<https://doi.org/10.1080/10463280600681248>

- Fincher, C. L., & Thornhill, R. (2008a). A parasite-driven wedge: infectious diseases may explain language and other biodiversity. *Oikos*, *117*(9), 1289–1297.
<https://doi.org/10.1111/j.0030-1299.2008.16684.x>
- Fincher, C. L., & Thornhill, R. (2008b). Assortative sociality, limited dispersal, infectious disease and the genesis of the global pattern of religion diversity. *Proceedings of the Royal Society of London B: Biological Sciences*, *275*(1651), 2587–2594.
<https://doi.org/10.1098/rspb.2008.0688>
- Fincher, C. L., & Thornhill, R. (2012). Parasite-stress promotes in-group assortative sociality: The cases of strong family ties and heightened religiosity. *Behavioral and Brain Sciences*, *35*(02), 61–79. <https://doi.org/10.1017/S0140525X11000021>
- Fincher, C. L., Thornhill, R., Murray, D. R., & Schaller, M. (2008). Pathogen prevalence predicts human cross-cultural variability in individualism/collectivism. *Proceedings of the Royal Society of London B: Biological Sciences*, *275*(1640), 1279–1285.
<https://doi.org/10.1098/rspb.2008.0094>
- Fox, E., Russo, R., Bowles, R., & Dutton, K. (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology. General*, *130*(4), 681–700.
- Freud, S. (1915). The unconscious. *Standard Edition*, *14*(1957), 159–215.
- Freud, S. (2005). *The Unconscious*. London: Penguin Classics.
- Frey, W. H., & Myers, D. (2005). Racial segregation in US metropolitan areas and cities, 1990–2000: Patterns, trends, and explanations. *Population Studies Center Research Report*, (05-573). Retrieved from <http://www.psc.isr.umich.edu/pubs/pdf/rr05-573.pdf>

- Friese, M., & Fiedler, K. (2009). Being on the Lookout for Validity. *Experimental Psychology*, 57(3), 228–232. <https://doi.org/10.1027/1618-3169/a000051>
- Friese, M., Hofmann, W., & Schmitt, M. (2009). When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology*, 19(1), 285–338. <https://doi.org/10.1080/10463280802556958>
- Friese, M., Hofmann, W., & Wänke, M. (2008). When impulses take over: moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour. *The British Journal of Social Psychology*, 47(Pt 3), 397–419. <https://doi.org/10.1348/014466607X241540>
- Friese, M., Smith, C. T., Koeber, M., & Bluemke, M. (2016). Implicit measures of attitudes and political voting behavior: Implicit measures and political voting behavior. *Social and Personality Psychology Compass*, 10(4), 188–201. <https://doi.org/10.1111/spc3.12246>
- Friese, M., Wänke, M., & Plessner, H. (2006). Implicit consumer preferences and their influence on product choice. *Psychology and Marketing*, 23(9), 727–740. <https://doi.org/10.1002/mar.20126>
- Gaertner, S., & Bickman, L. (1971). Effects of race on the elicitation of helping behavior: The wrong number technique. *Journal of Personality and Social Psychology*, 20(2), 218–222. <https://doi.org/10.1037/h0031681>
- Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, 46(1), 23–30. <https://doi.org/10.2307/3033657>

- Galdi, S., Gawronski, B., Arcuri, L., & Friese, M. (2012). Selective exposure in decided and undecided individuals: Differential relations to automatic associations and conscious beliefs. *Personality and Social Psychology Bulletin*, *38*(5), 559–569.
<https://doi.org/10.1177/0146167211435981>
- Gallup. (2017). In U.S., 64% support death penalty in cases of murder. Retrieved 18 February 2017, from <http://www.gallup.com/poll/144284/Support-Death-Penalty-Cases-Murder.aspx>
- Gawronski, B. (2009). Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology*, *50*(3), 141–150. <https://doi.org/10.1037/a0013848>
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. <https://doi.org/10.1037/0033-2909.132.5.692>
- Gawronski, B., & Bodenhausen, G. V. (2011). The associative–propositional evaluation model. In *Advances in Experimental Social Psychology* (Vol. 44, pp. 59–127). Elsevier.
<https://doi.org/10.1016/B978-0-12-385522-0.00002-0>
- Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd, *Handbook of research methods in social and personality psychology* (Vol. 2, pp. 283–310). New York, NY: Cambridge University Press.
- Gawronski, B., De Houwer, J., Reis, H. T., & Judd, C. M. (2014). Implicit measures in social and personality psychology. *Handbook of Research Methods in Social and Personality Psychology*, *2*, 283–310.

- Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are 'implicit' attitudes unconscious? *Consciousness and Cognition, 15*(3), 485–499.
<https://doi.org/10.1016/j.concog.2005.11.007>
- Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology, 44*(5), 1355–1361. <https://doi.org/10.1016/j.jesp.2008.04.005>
- Gawronski, B., Peters, K. R., Brochu, P. M., & Strack, F. (2008). Understanding the relations between different forms of racial prejudice: a cognitive consistency perspective. *Personality & Social Psychology Bulletin, 34*(5), 648–665.
<https://doi.org/10.1177/0146167207313729>
- Gawronski, B., & Sritharan, R. (2010). Formation, Change, and Contextualization of Mental Associations. In B. Gawronski & B. K. Payne, *Handbook of implicit social cognition: Measurement, theory, and applications*. (pp. 216–240). New York, NY: Guilford Press.
- Gervais, W. M., Shariff, A. F., & Norenzayan, A. (2011). Do you believe in atheists? Distrust is central to anti-atheist prejudice. *Journal of Personality and Social Psychology, 101*(6), 1189–1206. <https://doi.org/10.1037/a0025882>
- Goel, V., & Wingfield, N. (2015, December 1). Mark Zuckerberg vows to donate 99% of his facebook shares for charity. *The New York Times*. Retrieved from <https://www.nytimes.com/2015/12/02/technology/mark-zuckerberg-facebook-charity.html>
- Golijani-Moghaddam, N., Hart, A., & Dawson, D. L. (2013). The Implicit Relational Assessment Procedure: Emerging reliability and validity data. *Journal of Contextual Behavioral Science, 2*(3–4), 105–119. <https://doi.org/10.1016/j.jcbs.2013.05.002>

- Greenwald, A. G. (1990). What cognitive representations underlie social attitudes? *Bulletin of the Psychonomic Society*, 28(3), 254–260.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1), 4–27.
- Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, 108(4), 553–561. <https://doi.org/10.1037/pspa0000016>
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109(1), 3–25. <https://doi.org/10.1037//0033-295X.109.1.3>
- 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), 1464–80.
- Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift Für Experimentelle Psychologie*, 48(2), 85–93.
- Greenwald, A. G., & Nosek, B. A. (2008). Attitudinal dissociation: What does it mean. *Attitudes: Insights from the New Implicit Measures*, 6582. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.158.8361&rep=rep1&type=pdf>
- 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. <https://doi.org/10.1037/0022-3514.85.2.197>

- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. <https://doi.org/10.1037/a0015575>
- Greenwald, A. G., & Sriram, N. (2010). No measure is perfect, but some measures can be quite useful. *Experimental Psychology*, *57*(3), 238–242. <https://doi.org/10.1027/1618-3169/a000075>
- 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. <https://doi.org/10.1037/0022-3514.90.1.1>
- Hahn, A., & Gawronski, B. (2014). Do implicit evaluations reflect unconscious attitudes? *Behavioral and Brain Sciences*, *37*(1), 28–29. <https://doi.org/10.1017/S0140525X13000721>
- Hahn, A., & Gawronski, B. (2015). Implicit social cognition. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed.) (pp. 714–720). Oxford, UK: Elsevier. Retrieved from http://www.bertramgawronski.com/documents/HG_StevensHandbook.pdf
- Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology. General*, *143*(3), 1369–1392. <https://doi.org/10.1037/a0035028>
- Harrison, D. P., Stritzke, W. G. K., Fay, N., Ellison, T. M., & Hudaib, A.-R. (2014). Probing the implicit suicidal mind: Does the Death/Suicide Implicit Association Test reveal a desire to die, or a diminished desire to live? *Psychological Assessment*, *26*(3), 831–840. <https://doi.org/10.1037/pas0000001>

- Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, *10*(1), 47–66.
- Haspelmath, M. (2006). Against markedness (and what to replace it with). *Journal of Linguistics*, *42*(1), 25–70. <https://doi.org/10.1017/S0022226705003683>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.
<https://doi.org/10.1017/S0140525X0999152X>
- Hodson, G., & Dhont, K. (2015). The person-based nature of prejudice: Individual difference predictors of intergroup negativity. *European Review of Social Psychology*, *26*(1), 1–42.
<https://doi.org/10.1080/10463283.2015.1070018>
- Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421.
<https://doi.org/10.1037/a0018916>
- Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385.
<https://doi.org/10.1177/0146167205275613>
- Hofmann, W., Gschwendner, T., Nosek, B. A., & Schmitt, M. (2005). What moderates implicit—explicit consistency? *European Review of Social Psychology*, *16*(1), 335–390.
<https://doi.org/10.1080/10463280500443228>
- Hofmann, W., Rauch, W., & Gawronski, B. (2007). And deplete us not into temptation: Automatic attitudes, dietary restraint, and self-regulatory resources as determinants of

- eating behavior. *Journal of Experimental Social Psychology*, 43(3), 497–504.
<https://doi.org/10.1016/j.jesp.2006.05.004>
- Holtgraves, T. (2004). Social desirability and self-reports: testing models of socially desirable responding. *Personality & Social Psychology Bulletin*, 30(2), 161–172.
<https://doi.org/10.1177/0146167203259930>
- Houben, K., & Wiers, R. W. (2006). A test of the salience asymmetry interpretation of the alcohol-IAT. *Experimental Psychology*, 53(4), 292–300. <https://doi.org/10.1027/1618-3169.53.4.292>
- Houwer, J. D. (2014). A Propositional Model of Implicit Evaluation. *Social and Personality Psychology Compass*, 8(7), 342–353. <https://doi.org/10.1111/spc3.12111>
- Hruschka, D. J., & Henrich, J. (2013). Economic and evolutionary hypotheses for cross-population variation in parochialism. *Frontiers in Human Neuroscience*, 7.
<https://doi.org/10.3389/fnhum.2013.00559>
- Hughes, S., Barnes-Holmes, D., & Vahey, N. (2012). Holding on to our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. *Journal of Contextual Behavioral Science*, 1(1–2), 17–38.
<https://doi.org/10.1016/j.jcbs.2012.09.003>
- Hughes, S., Hussey, I., Corrigan, B., Jolie, K., Murphy, C., & Barnes-Holmes, D. (2016). Faking revisited: Exerting strategic control over performance on the Implicit Relational Assessment Procedure. *European Journal of Social Psychology*, 46(5), 632–648.
<https://doi.org/10.1002/ejsp.2207>
- Hughes, S. J., & Barnes-Holmes, D. (2013). A functional approach to the study of implicit cognition: the IRAP and the REC model. In S. Dymond & B. Roche, *Advances in*

- relational frame theory & contextual behavioural science: research & applications* (pp. 97–126). Oakland, CA, US.: Context Press.
- Hussey, I., Daly, T., & Barnes-Holmes, D. (2015). Life is Good, But Death Ain't Bad Either: Counter-Intuitive Implicit Biases to Death in a Normative Population. *The Psychological Record*, 65(4), 731–742. <https://doi.org/10.1007/s40732-015-0142-3>
- Hussey, I., Mhaoileoin, D. N., Barnes-Holmes, D., Ohtsuki, T., Kishita, N., Hughes, S., & Murphy, C. (2016). The IRAP Is nonrelative but not acontextual: Changes to the contrast category influence men's dehumanization of women. *The Psychological Record*, 66(2), 291–299.
- Jacoby, L. L., Toth, J. P., Lindsay, D. S., & Debnar, J. A. (1992). Lectures for a layperson: Methods for revealing unconscious processes. In R. F. Bornstein & B. Pittman (Eds.), *Perception without awareness: Cognitive, clinical, and social perspectives*. New York, NY: Guilford Press.
- Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes Factors. *The Journal of Problem Solving*, 7(1). <https://doi.org/10.7771/1932-6246.1167>
- Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, 76(5), 349–364. <https://doi.org/10.1037/h0031617>
- Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, 451(7181), 990–993. <https://doi.org/10.1038/nature06536>

- Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality and Social Psychology*, *85*(5), 969–978. <https://doi.org/10.1037/0022-3514.85.5.969>
- Jost, J. T., & Kay, A. C. (2005). Exposure to benevolent sexism and complementary gender stereotypes: consequences for specific and diffuse forms of system justification. *Journal of Personality and Social Psychology*, *88*(3), 498–509. <https://doi.org/10.1037/0022-3514.88.3.498>
- Kaplan, H. S., & Gangestad, S. W. (2015). Life history theory and evolutionary psychology. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology* (pp. 68–95). John Wiley & Sons, Inc. <https://doi.org/10.1002/9780470939376.ch2>
- Karpinski, A. (2004). Measuring self-esteem using the Implicit Association Test: The role of the other. *Personality and Social Psychology Bulletin*, *30*(1), 22–34.
- Karpinski, A., & Steinman, R. B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, *91*(1), 16–32. <https://doi.org/10.1037/0022-3514.91.1.16>
- Klauer, K. C., & Mierke, J. (2005). Task-set inertia, attitude accessibility, and compatibility-order effects: new evidence for a task-set switching account of the implicit association test effect. *Personality & Social Psychology Bulletin*, *31*(2), 208–217. <https://doi.org/10.1177/0146167204271416>
- Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology*, *93*(3), 353–368. <https://doi.org/10.1037/0022-3514.93.3.353>

- Klavina, L. Ms., Buunk, A. P., & Pollet, T. V. (2011). Out-group mating threat and disease threat increase implicit negative attitudes toward the out-group among men. *Frontiers in Psychology*, 2. <https://doi.org/10.3389/fpsyg.2011.00076>
- Kosnes, L., Whelan, R., O'Donovan, A., & McHugh, L. A. (2013). Implicit measurement of positive and negative future thinking as a predictor of depressive symptoms and hopelessness. *Consciousness and Cognition*, 22(3), 898–912. <https://doi.org/10.1016/j.concog.2013.06.001>
- Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 75(1), 23–55. <https://doi.org/10.1006/obhd.1998.2781>
- Kunstman, J. W., & Plant, E. A. (2008). Racing to help: racial bias in high emergency helping situations. *Journal of Personality and Social Psychology*, 95(6), 1499–1510. <https://doi.org/10.1037/a0012822>
- Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, 143(3), 1065–1081. <https://doi.org/10.1037/a0035669>
- Kurzban, R., & Leary, M. R. (2001). Evolutionary origins of stigmatization: The functions of social exclusion. *Psychological Bulletin*, 127(2), 187–208. <https://doi.org/10.1037//0033-2909.127.2.187>
- Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., ... Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, 143(4), 1765–1785. <https://doi.org/10.1037/a0036260>

- Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., ... Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016.
<https://doi.org/10.1037/xge0000179>
- LaPiere, R. T. (1934). Attitudes vs actions. *Social Forces*, *39*(1), 230–237.
- Larsen, R. J., Mercer, K. A., & Balota, D. A. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion*, *6*(1), 62–72. <https://doi.org/10.1037/1528-3542.6.1.62>
- Leech, G. N. (2006). *A glossary of English grammar*. New York, NY: Columbia University Press.
- Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*(5), 842–855.
- MacInnis, C. C., & Page-Gould, E. (2015). How can intergroup interaction be bad if intergroup contact is good? Exploring and reconciling an apparent paradox in the science of intergroup relations. *Perspectives on Psychological Science*, *10*(3), 307–327.
<https://doi.org/10.1177/1745691614568482>
- Matlin, M. W. (2016). Pollyanna Principle. In Rüdiger F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in judgement, thinking and memory* (pp. 315–333). New York, NY: Routledge.
- Matthews, W. J., & Dylman, A. S. (2014). The language of magnitude comparison. *Journal of Experimental Psychology: General*, *143*(2), 510–520. <https://doi.org/10.1037/a0034143>
- McDonald, M. M., Navarrete, C. D., & Van Vugt, M. (2012). Evolution and the psychology of intergroup conflict: the male warrior hypothesis. *Philosophical Transactions of the Royal*

Society B: Biological Sciences, 367(1589), 670–679.

<https://doi.org/10.1098/rstb.2011.0301>

McFarland, S. G., & Crouch, Z. (2002). A cognitive skill confound on the Implicit Association Test. *Social Cognition*, 20(6), 483–510.

McKenna, I. M., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2007). Testing the fake-ability of the Implicit Relational Assessment Procedure (IRAP): The first study. *International Journal of Psychology and Psychological Therapy*, 7(2), 253–268.

McKenzie, C. R. M., & Nelson, J. D. (2003). What a speaker's choice of frame reveals: reference points, frame selection, and framing effects. *Psychonomic Bulletin & Review*, 10(3), 596–602.

Meissner, F., & Rothermund, K. (2015). A thousand words are worth more than a picture? The effects of stimulus modality on the Implicit Association Test. *Social Psychological and Personality Science*, 6(7), 740–748. <https://doi.org/10.1177/1948550615580381>

Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences*, 7(3), 141–144. [https://doi.org/10.1016/S1364-6613\(03\)00029-9](https://doi.org/10.1016/S1364-6613(03)00029-9)

Miller, L. C., Murphy, R., & Buss, A. H. (1981). Consciousness of body: Private and public. *Journal of Personality and Social Psychology*, 41(2), 397–406.

<https://doi.org/10.1037/0022-3514.41.2.397>

Mitchell, C. J. (2004). Mere acceptance produces apparent attitude in the Implicit Association Test. *Journal of Experimental Social Psychology*, 40(3), 366–373.

<https://doi.org/10.1016/j.jesp.2003.07.003>

- Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition, 19*(4), 395–417. <https://doi.org/10.1521/soco.19.4.395.20759>
- Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology, 11*(1), 56–60. <https://doi.org/10.1080/17470215908416289>
- Murray, D. R. (2014). Direct and indirect implications of pathogen prevalence for scientific and technological innovation. *Journal of Cross-Cultural Psychology, 45*(6), 971–985. <https://doi.org/10.1177/0022022114532356>
- Murray, D. R., & Schaller, M. (2012). Threat(s) and conformity deconstructed: Perceived threat of infectious disease and its implications for conformist attitudes and behavior. *European Journal of Social Psychology, 42*(2), 180–188. <https://doi.org/10.1002/ejsp.863>
- Murray, D. R., & Schaller, M. (2016). The behavioral immune system: Implications for social cognition, social interaction, and social influence. In J. M. O. and M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 53, pp. 75–129). Academic Press. Retrieved from [//www.sciencedirect.com/science/article/pii/S0065260115000246](http://www.sciencedirect.com/science/article/pii/S0065260115000246)
- Murray, D. R., Schaller, M., & Suedfeld, P. (2013). Pathogens and politics: Further evidence that parasite prevalence predicts authoritarianism. *PLoS ONE, 8*(5), e62275. <https://doi.org/10.1371/journal.pone.0062275>
- Nasrallah, M., Carmel, D., & Lavie, N. (2009). Murder, she wrote: Enhanced sensitivity to negative word valence. *Emotion, 9*(5), 609–618. <https://doi.org/10.1037/a0016305>

- Navarrete, C. D., & Fessler, D. M. T. (2006). Disease avoidance and ethnocentrism: the effects of disease vulnerability and disgust sensitivity on intergroup attitudes. *Evolution and Human Behavior*, 27(4), 270–282. <https://doi.org/10.1016/j.evolhumbehav.2005.12.001>
- Navarrete, C. D., Fessler, D. M. T., Fleischman, D. S., & Geyer, J. (2009). Race Bias Tracks Conception Risk Across the Menstrual Cycle. *Psychological Science*, 20(6), 661–665. <https://doi.org/10.1111/j.1467-9280.2009.02352.x>
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231.
- Nock, M. K., & Banaji, M. R. (2007). Prediction of Suicide Ideation and Attempts Among Adolescents Using a Brief Performance-Based Test. *Journal of Consulting and Clinical Psychology*, 75(5), 707–715. <https://doi.org/10.1037/0022-006X.75.5.707>
- Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. *Psychological Science*, 21(4), 511–517. <https://doi.org/10.1177/0956797610364762>
- Nolan, J., Murphy, C., & Barnes-Holmes, D. (2013). Implicit Relational Assessment Procedure and body-weight bias: influence of gender of participants and targets. *The Psychological Record*, 63(3), 467–489.
- Nosek, B. A. (2007). Implicit–Explicit Relations. *Current Directions in Psychological Science*, 16(2), 65–69. <https://doi.org/10.1111/j.1467-8721.2007.00477.x>
- Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1), 101–115. <https://doi.org/10.1037/1089-2699.6.1.101>

- Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition, 19*(6), 625–666. <https://doi.org/10.1521/soco.19.6.625.20886>
- Nosek, B. A., Bar-Anan, Y., Sriram, N., Axt, J., & Greenwald, A. G. (2014). Understanding and using the Brief Implicit Association Test: Recommended scoring procedures. *PLOS ONE, 9*(12), e110938. <https://doi.org/10.1371/journal.pone.0110938>
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*(2), 166–180. <https://doi.org/10.1177/0146167204271418>
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. *Automatic Processes in Social Thinking and Behavior, 265–292*.
- Nosek, B. A., & Hansen, J. J. (2008a). Personalizing the Implicit Association Test increases explicit evaluation of target concepts. *European Journal of Psychological Assessment, 24*(4), 226–236.
- Nosek, B. A., & Hansen, J. J. (2008b). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition & Emotion, 22*(4), 553–594. <https://doi.org/10.1080/02699930701438186>
- Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences, 15*(4), 152–159. <https://doi.org/10.1016/j.tics.2011.01.005>
- Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test: implicit and explicit attitudes are related but distinct constructs. *Experimental Psychology, 54*(1), 14–29. <https://doi.org/10.1027/1618-3169.54.1.14>

- Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*(1), 36–88.
<https://doi.org/10.1080/10463280701489053>
- Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., ... Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences, 106*(26), 10593–10597. <https://doi.org/10.1073/pnas.0809921106>
- O'Brien, K. S., Hunter, J. A., & Banks, M. (2007). Implicit anti-fat bias in physical educators: physical attributes, ideology and socialization. *International Journal of Obesity (2005), 31*(2), 308–314. <https://doi.org/10.1038/sj.ijo.0803398>
- Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General, 130*(3), 466–478.
<https://doi.org/10.1037//0096-3445.130.3.466>
- Ohman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: a threat advantage with schematic stimuli. *Journal of Personality and Social Psychology, 80*(3), 381–396.
- Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12*(5), 413–417. <https://doi.org/10.1111/1467-9280.00376>
- Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: personalizing the IAT. *Journal of Personality and Social Psychology, 86*(5), 653–667. <https://doi.org/10.1037/0022-3514.86.5.653>

- Olson, M. A., Fazio, R. H., & Han, H. A. (2009). Conceptualizing personal and extrapersonal associations. *Social and Personality Psychology Compass*, 3(2), 152–170.
- O’Shea, B. (2015). Capitalism versus a new economic model: Implicit and explicit attitudes of protesters and bankers. *Social Movement Studies*, 14(3), 311–330.
<https://doi.org/10.1080/14742837.2014.938732>
- O’Shea, B. (2017a). Attitudes towards atheism and religious belief: Limitations with the Implicit Relational Assessment Procedure (IRAP). *Unpublished Manuscript*.
- O’Shea, B. (2017b). [The IRAP fails to measure abstract concepts absolutely]. *Unpublished Raw Data*.
- O’Shea, B., Brown, G. D. A., & Watson, D. G. (2017a). *A noun paints a concrete picture: Fundamental response biases in reaction time tasks*. Unpublished Manuscript.
- O’Shea, B., Brown, G. D. A., & Watson, D. G. (2017b). *The Simple Implicit Procedure (SIP): A new method of measuring implicit cognition*. Unpublished manuscript.
- O’Shea, B., Brown, G. D. A., Watson, D. G., & Fincher, C. L. (2017). *Disease rates are associated with racial prejudice across the US and the world*. Unpublished manuscript.
- O’Shea, B., De Houwer, J., Ratliff, K. A., Brown, G. D. A., & Watson, D. G. (2017). *Practice effects in implicit measures*. Unpublished Manuscript.
- O’Shea, B., Watson, D. G., & Brown, G. D. A. (2016). Measuring implicit attitudes: A positive framing bias flaw in the Implicit Relational Assessment Procedure (IRAP). *Psychological Assessment*, 28(2), 158–170. <https://doi.org/10.1037/pas0000172>
- Oskamp, S., & Schultz, P. W. (2005). *Attitudes and Opinions (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105*(2), 171–192. <https://doi.org/10.1037/a0032734>
- Park, J. H., Faulkner, J., & Schaller, M. (2003). Evolved disease-avoidance processes and contemporary anti-social behavior: Prejudicial attitudes and avoidance of people with physical disabilities. *Journal of Nonverbal Behavior, 27*(2), 65–87.
- Park, J. H., Schaller, M., & Crandall, C. S. (2007). Pathogen-avoidance mechanisms and the stigmatization of obese people. *Evolution and Human Behavior, 28*(6), 410–414. <https://doi.org/10.1016/j.evolhumbehav.2007.05.008>
- Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94*(1), 16–31. <https://doi.org/10.1037/0022-3514.94.1.16>
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277–293. <https://doi.org/10.1037/0022-3514.89.3.277>
- Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going. In B. Gawronski & K. B. Payne, *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 1–15). New York, NY: Guilford Press.
- Perugini, M., Richetin, J., & Zogmaister, C. (2010). Prediction of behavior. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 255–277). New York, NY, US: Guilford Press.

- Peters, K. R., & Gawronski, B. (2011). Mutual influences between the implicit and explicit self-concepts: The role of memory activation and motivated reasoning. *Journal of Experimental Social Psychology, 47*(2), 436–442.
<https://doi.org/10.1016/j.jesp.2010.11.015>
- Peterson, C. (2000). The future of optimism. *The American Psychologist, 55*(1), 44–55.
- Petrescu, D. C., & Parkinson, B. (2014). Incidental disgust increases adherence to left-wing economic attitudes. *Social Justice Research, 27*(4), 464–486.
<https://doi.org/10.1007/s11211-014-0221-7>
- Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*(5), 751–783.
<https://doi.org/10.1037/0022-3514.90.5.751>
- Pollet, T. V., Tybur, J. M., Frankenhuys, W. E., & Rickard, I. J. (2014). What can cross-cultural correlations teach us about human nature? *Human Nature, 25*(3), 410–429.
<https://doi.org/10.1007/s12110-014-9206-3>
- Pratto, F., & John, O. P. (1991). Automatic vigilance: the attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*(3), 380–391.
- Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin, 132*(3), 416–442. <https://doi.org/10.1037/0033-2909.132.3.416>
- Putnam, R. D. (2007). E pluribus unum: Diversity and community in the twenty-first century the 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies, 30*(2), 137–174.

- Rae, J. R., Newheiser, A.-K., & Olson, K. R. (2015). Exposure to racial out-groups and implicit race bias in the United States. *Social Psychological and Personality Science*, 6(5), 535–543. <https://doi.org/10.1177/1948550614567357>
- Randall, J. R., Rowe, B. H., Dong, K. A., Nock, M. K., & Colman, I. (2013). Assessment of self-harm risk using implicit thoughts. *Psychological Assessment*, 25(3), 714–721. <https://doi.org/10.1037/a0032391>
- Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, 44(2), 386–396. <https://doi.org/10.1016/j.jesp.2006.12.008>
- Reicher, S. D., Templeton, A., Neville, F., Ferrari, L., & Drury, J. (2016). Core disgust is attenuated by ingroup relations. *Proceedings of the National Academy of Sciences*, 113(10), 2631–2635. <https://doi.org/10.1073/pnas.1517027113>
- Reid, S. A., Zhang, J., Anderson, G. L., Gasiorek, J., Bonilla, D., & Peinado, S. (2012). Parasite primes make foreign-accented English sound more distant to people who are disgusted by pathogens (but not by sex or morality). *Evolution and Human Behavior*, 33(5), 471–478. <https://doi.org/10.1016/j.evolhumbehav.2011.12.009>
- Remue, J., De Houwer, J., Barnes-Holmes, D., Vanderhasselt, M.-A., & De Raedt, R. (2013). Self-esteem revisited: Performance on the implicit relational assessment procedure as a measure of self- versus ideal self-related cognitions in dysphoria. *Cognition & Emotion*, 27(8), 1441–1449. <https://doi.org/10.1080/02699931.2013.786681>

- Remue, J., Hughes, S., Houwer, J. D., & Raedt, R. D. (2014). To be or want to be: Disentangling the role of actual versus ideal self in implicit self-esteem. *PLOS ONE*, *9*(9), e108837. <https://doi.org/10.1371/journal.pone.0108837>
- Richetin, J., Costantini, G., Perugini, M., & Schönbrodt, F. (2015). Should We Stop Looking for a Better Scoring Algorithm for Handling Implicit Association Test Data? Test of the Role of Errors, Extreme Latencies Treatment, Scoring Formula, and Practice Trials on Reliability and Validity. *PLOS ONE*, *10*(6), e0129601. <https://doi.org/10.1371/journal.pone.0129601>
- Richetin, J., Perugini, M., Adjali, I., & Hurling, R. (2007). The moderator role of intuitive versus deliberative decision making for the predictive validity of implicit and explicit measures. *European Journal of Personality*, *21*(4), 529–546. <https://doi.org/10.1002/per.625>
- Robinson, M. D., Meier, B. P., Zetocha, K. J., & McCaul, K. D. (2005). Smoking and the Implicit Association Test: When the Contrast Category Determines the Theoretical Conclusions. *Basic and Applied Social Psychology*, *27*(3), 201–212. https://doi.org/10.1207/s15324834basp2703_2
- Roddy, S., Stewart, I., & Barnes-Holmes, D. (2010). Anti-fat, pro-slim, or both? Using two reaction-time based measures to assess implicit attitudes to the slim and overweight. *Journal of Health Psychology*, *15*(3), 416–425. <https://doi.org/10.1177/1359105309350232>
- Roddy, S., Stewart, I., & Barnes-Holmes, D. (2011). Facial reactions reveal that slim is good but fat is not bad: Implicit and explicit measures of body-size bias. *European Journal of Social Psychology*, *41*(6), 688–694. <https://doi.org/10.1002/ejsp.839>

- Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523–534. <https://doi.org/10.1016/j.labeco.2009.04.005>
- Rosenberg, M. (2015). *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press.
- Rothermund, K., & Wentura, D. (2001). Figure-ground asymmetries in the Implicit Association Test (IAT). *Zeitschrift Fur Experimentelle Psychologie*, *48*(2), 94–106.
- Rothermund, K., & Wentura, D. (2010). It's Brief But Is It Better? An Evaluation of the Brief Implicit Association Test. *Experimental Psychology*, *57*(3), 233–237. <https://doi.org/10.1027/1618-3169/a000060>
- Rothermund, Klaus, Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the Implicit Association Test: the Recoding-Free Implicit Association Test (IAT-RF). *Quarterly Journal of Experimental Psychology*, *62*(1), 84–98. <https://doi.org/10.1080/17470210701822975>
- Rothermund, Klaus, & Wentura, D. (2004). Underlying Processes in the implicit association test: Dissociating salience from associations. *Journal of Experimental Psychology: General*, *133*(2), 139–165. <https://doi.org/10.1037/0096-3445.133.2.139>
- Rothermund, Klaus, Wentura, D., & De Houwer, J. (2005). Validity of the salience asymmetry account of the Implicit Association Test: Reply to Greenwald, Nosek, Banaji, and Klauer (2005). *Journal of Experimental Psychology: General*, *134*(3), 426–430. <https://doi.org/10.1037/0096-3445.134.3.426>
- Rudman, L. A. (2004). Sources of Implicit Attitudes. *Current Directions in Psychological Science*, *13*(2), 79–82. <https://doi.org/10.1111/j.0963-7214.2004.00279.x>

- Rudman, L. A., & Phelan, J. E. (2010). The effect of priming gender roles on women's implicit gender beliefs and career aspirations. *Social Psychology, 41*(3), 192–202.
<https://doi.org/10.1027/1864-9335/a000027>
- Ryan, S., Oaten, M., Stevenson, R. J., & Case, T. I. (2012). Facial disfigurement is treated like an infectious disease. *Evolution and Human Behavior, 33*(6), 639–646.
<https://doi.org/10.1016/j.evolhumbehav.2012.04.001>
- Salvatore, J. F., Meltzer, A. L., March, D. S., & Gaertner, L. (2017). Strangers with benefits: Attraction to outgroup men increases as fertility increases across the menstrual cycle. *Personality and Social Psychology Bulletin, 43*(2), 204–217.
<https://doi.org/10.1177/0146167216678860>
- Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science, 19*(8), 772–780.
<https://doi.org/10.1111/j.1467-9280.2008.02156.x>
- Schacter, D. L., Chiu, C.-Y. P., & Ochsner, K. N. (1993). Implicit memory: A selective review. *Annual Review of Neuroscience, 16*(1), 159–182.
- Schaller, M., Miller, G. E., Gervais, W. M., Yager, S., & Chen, E. (2010). Mere visual perception of other people's disease symptoms facilitates a more aggressive immune response. *Psychological Science, 21*(5), 649–652.
<https://doi.org/10.1177/0956797610368064>
- Schaller, M., & Murray, D. R. (2008). Pathogens, personality, and culture: disease prevalence predicts worldwide variability in sociosexuality, extraversion, and openness to experience. *Journal of Personality and Social Psychology, 95*(1), 212–221.
<https://doi.org/10.1037/0022-3514.95.1.212>

- Schaller, M., & Park, J. H. (2011). The Behavioral Immune System (and Why It Matters). *Current Directions in Psychological Science*, 20(2), 99–103.
<https://doi.org/10.1177/0963721411402596>
- Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, 34(8), 1096–1109.
<https://doi.org/10.1177/0146167208317771>
- Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition*, 25(5), 638–656.
- Sheeran, P. (2002). Intention—Behavior Relations: A Conceptual and Empirical Review. *European Review of Social Psychology*, 12(1), 1–36.
<https://doi.org/10.1080/14792772143000003>
- Sherman, M. A. (1973). Bound to be easier? The negative prefix and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, 12(1), 76–84.
[https://doi.org/10.1016/S0022-5371\(73\)80062-3](https://doi.org/10.1016/S0022-5371(73)80062-3)
- 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. <https://doi.org/10.1037/0033-295X.84.2.127>
- Smith, C. T., & Nosek, B. A. (2011). Affective focus increases the concordance between implicit and explicit attitudes. *Social Psychology*, 42(4), 300–313. <https://doi.org/10.1027/1864-9335/a000072>
- Spence, A., & Townsend, E. (2008). Spontaneous evaluations: Similarities and differences between the affect heuristic and implicit attitudes. *Cognition and Emotion*, 22(1), 83–93.
<https://doi.org/10.1080/02699930701298432>

- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8(3), 220–247.
- Taylor, M. C. (1998). How white attitudes vary with the racial composition of local populations: Numbers count. *American Sociological Review*, 63(4), 512–535.
<https://doi.org/10.2307/2657265>
- Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin*, 116(1), 21–27. <https://doi.org/10.1037/0033-2909.116.1.21>
- Taylor, S. E., Klein, L. C., Lewis, B. P., Gruenewald, T. L., R, A., & Updegraff, J. A. (2000). Biobehavioral responses to stress in females: Tend-and-befriend, not fight-or-flight. *Psychological Review*, 107(3), 411–429. <https://doi.org/10.1037/0033-295X.107.3.411>
- Teachman, B. A., & Allen, J. P. (2007). Development of social anxiety: Social interaction predictors of implicit and explicit fear of negative evaluation. *Journal of Abnormal Child Psychology*, 35(1), 63–78. <https://doi.org/10.1007/s10802-006-9084-1>
- Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2010). A practical guide to the Implicit Association Test and related tasks. In B. Gawronski & B. K. Payne, *Handbook of implicit social cognition: Measurement, theory and applications*. New York, NY: Guilford Press.
- Teige-Mocigemba, Sarah, Klauer, K. C., & Rothermund, K. (2008). Minimizing method-specific variance in the IAT. *European Journal of Psychological Assessment*, 24(4), 237–245.
<https://doi.org/10.1027/1015-5759.24.4.237>
- Terrizzi, J. A., Clay, R., & Shook, N. J. (2014). Does the behavioral immune system prepare females to be religiously conservative and collectivistic? *Personality and Social Psychology Bulletin*, 40(2), 189–202. <https://doi.org/10.1177/0146167213508792>

- Thornhill, R., & Fincher, C. L. (2014). *The Parasite-Stress Theory of values and sociality*. Cham: Springer International Publishing.
- Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry*, *48*, 59–65.
<https://doi.org/10.1016/j.jbtep.2015.01.004>
- Van de Vyver, J., Houston, D. M., Abrams, D., & Vasiljevic, M. (2016). Boosting belligerence: How the July 7, 2005, London bombings affected liberals' moral foundations and prejudice. *Psychological Science*, *27*(2), 169–177.
<https://doi.org/10.1177/0956797615615584>
- Waller, G., Watkins, H., Shuck, V., & McManus, F. (1996). Bulimic psychopathology and attentional biases to ego threats among non-eating-disordered women. *The International Journal of Eating Disorders*, *20*(2), 169–176.
- Wentura, D., & Degner, J. (2010). A practical guide to sequential priming and related tasks. In B. Gawronski & B. K. Payne, *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 95–116). New York, NY: Guilford Press. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.570.629&rep=rep1&type=pdf>
- Wentura, D., & Rothermund, K. (2007). Paradigms we live by. A plea for more basic research on the IAT. In B. Wittenbrink & N. Schwarz, *Implicit measures of attitudes* (pp. 195–215). New York: Guilford Press.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 T-

- Tests. *Perspectives on Psychological Science*, 6(3), 291–298.
<https://doi.org/10.1177/1745691611406923>
- Wigboldus, W. H. J., Holland, R. W., & van Knippenberg, A. (2005). *Single Target Implicit Association*. Unpublished Manuscript.
- Williams, K. E. G., Sng, O., & Neuberg, S. L. (2016). Ecology-driven stereotypes override race stereotypes. *Proceedings of the National Academy of Sciences*, 113(2), 310–315.
<https://doi.org/10.1073/pnas.1519401113>
- Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, 107(1), 101–126.
- Wittenbrink, Bernd, 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. <https://doi.org/10.1037/0022-3514.81.5.815>
- Wu, B.-P., & Chang, L. (2012). The social impact of pathogen threat: How disease salience influences conformity. *Personality and Individual Differences*, 53(1), 50–54.
<https://doi.org/10.1016/j.paid.2012.02.023>
- Xu, K., Nosek, B., & Greenwald, A. (2014). Psychology data from the Race Implicit Association Test on the Project Implicit Demo website. *Journal of Open Psychology Data*, 2(1).
<https://doi.org/10.5334/jopd.ac>

Appendix 1: Chapter 2

At the individual level for white participants across the US, conservative, less religious, and older males displayed higher implicit and explicit anti-black/pro-white biases ($ts > 4.55$, $ps < .001$). All these findings are consistent with previous literature (Hodson & Dhont, 2015), except that research normally finds that amongst white participants people who are *more* religious express higher prejudice. The finding that less religious white individuals express more prejudice warrants further investigation. For black participants, older, more religious females displayed stronger implicit and explicit anti-white/pro-black biases ($ts > 6.15$, $ps < .001$). At the implicit level, black conservatives showed stronger anti-white/pro-black biases ($t = 2.43$, $p < .05$), while at the explicit level liberals showed stronger anti-black/pro-white biases ($t = 20.30$, $ps < .001$).

Focusing on implicit attitudes, less educated white respondents showed the highest prejudice ($t = -4.61$, $p < .001$), while for explicit attitudes, less educated respondents showed the lowest prejudice towards black people ($t = 30.46$, $p < .001$). For black participants, a similar change in directionality was observed for education. For example, less educated respondents showed stronger implicit anti-white/pro-black biases ($t = 2.43$, $p < .05$), while for their explicit attitudes they expressed lower anti-white/pro-black biases ($t = 33.18$, $p < .001$). Therefore, for both black and white respondents, those with less education explicitly express more egalitarian views, while with implicit measures opposite findings were obtained.

A post-hoc explanation for this finding is that those with less education are more likely to live in deprived areas, where they are exposed to various out-groups of a similar social status. This increased contact might reduce their explicit prejudice towards these out-groups to allow for peaceful co-existence. However, evolved survival mechanisms (e.g., disease avoidance) may be amplified in deprived environments to protect the individual and their in-group from potential threats (e.g., pathogens).

For the cross-country analysis, the individual-level factors of being conservative, less religious, older, and male were again associated with increased anti-black/pro-white prejudice implicit and explicit attitudes ($ts > 4.46$, $ps < .001$). Similar to the US findings, we again find that those who were less educated were associated with increased implicit prejudice ($ts > -2.25$, $ps < .001$) while explicitly those less educated were associated with reduced prejudice towards black people ($ts > 33.18$, $ps < .001$) suggesting that the influence that education has on implicit and explicit prejudice warrants further investigation.

Table S2.1: Demographic characteristics of the participants from Project Implicit's Race IAT (Years: 2006-2013)

Characteristic	White World Respondents (N = 1,347,295)		White US Respondents (N= 1,213,085)		Black US Respondents (N=225,556)	
	N	%	N	%	N	%
Political Identification (M±SD)	3.67 ± 1.67		3.72 ± 1.68		3.53 ± 1.48	
Religiosity						
Not at all Religious	141,153	10.20	124,212	10.20	13,826	6.10
Slightly Religious	332,921	24.70	308,447	25.40	45,629	20.20
Moderately Religious	329,188	24.40	313,282	25.80	85,686	38.00
Strongly Religious	140,979	10.50	134,659	11.20	50,756	22.60
Missing	403,054	29.90	332,485	27.40	29,659	13.10
Gender						
Female	756,246	56.10	693,685	57.20	147,573	65.40
Male	584,347	43.40	513,665	42.30	76,913	34.10
Missing/Other	6,702	0.50	5,735	0.50	1,070	0.50
Age in Years (M±SD)	27.60 ± 12.13		27.58 ± 12.20		29.41 ± 11.63	
Education						
High School Graduate or below	782,644	58.10	723,054	59.60	141,080	62.50
Anything above High School	553,781	41.10	480,413	39.60	82,437	36.50
Missing	10,870	0.80	9,618	0.80	2,039	0.90
Reason for Visiting Project Implicit						
Assignment for School	421,536	31.30	409,645	33.80	85,354	37.80
Recommendation of Teacher	59,944	4.40	54,737	4.50	12,057	5.30
Recommendation of Friend	40,147	3.00	33,816	2.80	5,283	2.30
Other	392,311	24.50	280,530	23.20	47,079	21.00
Missing	433,357	36.80	433,357	35.70	75,783	33.60

Table S2.2: Disease Rates (higher numbers indicate a greater anti-black/pro-white bias) and Ratio of Whites to Blacks (logged) Accounting for Segregation (lower scores indicate more black exposure) Predicting US State Level Implicit/Explicit Scores.

Predictor	White Implicit Attitudes			White Explicit Attitudes		
	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>
Disease Rates	0.018	0.004	5.02***	0.082	0.011	7.72***
Race Exposure	-0.031	0.012	-2.53*	-0.051	0.037	-1.36
-2 Log Likelihood	1063071.879			2676569.922		

† $p < .10$, * $p < .05$. ** $p < .01$. *** $p < .001$.

Table S2.3: Disease Rates (higher numbers indicate a greater anti-black/pro-white bias) and Ratio of Whites to Blacks (not logged) Accounting for Segregation (lower scores indicate more black exposure) Predicting US State Level Implicit/Explicit Scores.

Predictor	White Implicit Attitudes			White Explicit Attitudes		
	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>
Disease Rates	0.016	0.004	4.44***	0.078	0.011	6.98***
Race Exposure	-0.001	0.000	-2.64*	-0.001	0.000	-1.76†
-2 Log Likelihood	.063079.716			2676577.039		

† $p < .10$, * $p < .05$. ** $p < .01$. *** $p < .001$.

Table S2.4: Disease Rates (higher numbers indicate a greater anti-black/pro-white bias) and Proportion of Blacks to the Total State Population Accounting for Segregation (lower scores indicate more black exposure) Predicting US State Level Implicit/Explicit Scores.

Predictor	White Implicit Attitudes			White Explicit Attitudes		
	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>
Disease Rates	0.005	0.005	0.93	0.043	0.016	2.65*
Black Exposure	-0.406	0.109	-3.74**	-1.114	0.322	-3.47**
-2 Log Likelihood	.063061.002			2676556.432		

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Note: For Table S2.2-S2.4, scores in bold highlight the model with the best fit based on -2 Log Likelihood.

Table S2.5: Disease Rates (higher numbers indicate a greater anti-black/pro-white bias) and Proportion of Blacks to the Total State Population Accounting for Segregation (lower scores indicate more black exposure), including the Interaction Predicting US State Level Scores.

Predictor	White Implicit Attitudes			White Explicit Attitudes		
	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>
Disease Rates	0.007	0.005	1.31	0.045	0.016	2.84**
Black Exposure	-0.661	0.137	-4.82***	-1.154	0.427	-3.60**
Disease*Exposure	0.156	0.056	2.80**	0.256	0.173	1.48
-2 Log Likelihood	1063057.508			2676555.925		

† $p < .10$, * $p < .05$. ** $p < .01$. *** $p < .001$.

Note: Comparing Supplementary Table 2.4 (no interaction) with Supplementary Table 2.5 (interaction), we find that including the interaction between disease rates and black exposure improves the model. Using the conservative Schwarz's Bayesian Criterion (BIC), we show that there is strong evidence of an improved fit for the model including the interaction at the implicit level (Table 2.5 BIC = 1063085.22 versus Table 2.4 BIC = 1063088.71) but not at the explicit level (Table 2.5 BIC = 2676583.55 versus Table 2.4 BIC = 2676584.06). Therefore, the significant interaction term at the implicit level highlights the fact that white respondents living in states with higher disease rates and lower exposure to black people exhibit higher anti-black/pro-white prejudice. This finding is in line with contact hypothesis. Including the interaction term for black respondents across the US and white respondents across the world did not improve the fit of their respective models.

Table S2.6: Disease Rates (higher numbers indicate a greater anti-black/pro-white bias) and Ratio of Whites to Blacks (logged) Accounting for Segregation (higher scores indicate more white exposure) Predicting US State Level Implicit/Explicit Scores.

Predictor	Black Implicit Attitudes			Black Explicit Attitudes		
	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>
Parasite Stress	-0.025	0.004	-6.71***	-0.076	0.011	-6.98***
Race Exposure	0.008	0.014	0.61	0.086	0.039	2.20*
-2 Log Likelihood	222945.764			603433.464		

† $p < .10$, * $p < .05$. ** $p < .01$. *** $p < .001$.

Table S2.7: Disease Rates (higher numbers indicate a greater anti-black/pro-white bias) and Ratio of Whites to Blacks (not logged) Accounting for Segregation (higher scores indicate more white exposure) Predicting US State Level Implicit/Explicit Scores.

Predictor	Black Implicit Attitudes			Black Explicit Attitudes		
	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>
Parasite Stress	-0.021	0.004	-5.35***	-0.054	0.010	-5.45***
Race Exposure	0.001	0.00	2.79**	0.007	0.001	5.88***
-2 Log Likelihood	222945.439			603411.302		

† $p < .10$, * $p < .05$. ** $p < .01$. *** $p < .001$.

Table S2.8: Disease Rates (higher numbers indicate a greater anti-black/pro-white bias) and Proportion of Whites to the Total State Population Accounting for Segregation (higher scores indicate more white exposure) Predicting US State Level Implicit/Explicit Scores.

Predictor	Black Implicit Attitudes			Black Explicit Attitudes		
	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>	<i>B (est.)</i>	<i>SE B</i>	<i>t</i>
Parasite Stress	-0.026	0.003	-7.76***	-0.083	0.010	-8.65***
White Exposure	0.000	0.000	-0.04	0.220	0.117	1.88†
-2 Log Likelihood	222943.883			603432.543		

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Note: For Table 2.6-2.8, scores in bold highlight the model with the best fit based on -2 Log Likelihood.

Table S2.9: Correlational matrix of the variables in study 3 (Experiment)

	Implicit	Explicit	Gender	Education	Age	Political	Religious	Illness	PBC
Implicit	—	0.196***	0.094	-0.051	-0.041	0.139**	-0.001	0.055	0.041
Explicit		—	0.173***	-0.061	-0.003	0.287***	-0.069	-0.045	0.041
Gender			—	-0.119*	-0.058	0.126*	0.011	-0.019	-0.118
Education				—	0.271***	-0.036	0.098	-0.042	-0.014
Age					—	-0.145**	0.034	-0.056	-0.013
Political						—	0.289***	-0.010	-0.077
Religious							—	0.030	0.002
Illness								—	0.091
PBC									—

* $p < .05$, ** $p < .01$, *** $p < .001$

Study 3 Gender differences

A separate analysis was carried out on male and female implicit and explicit scores across the three priming conditions (control, terror, diseases: see Table S2.10 for means and standard deviations). First focusing on female scores, a one-way between-subject ANOVA showed a non-significant result of priming condition for both the implicit $F(2, 169) = 1.77, p = .173$, and explicit scores $F(2, 169) = .30, p > .250$. Males on the other hand showed a significant difference across the three conditions for both their IAT scores $F(2, 223) = 3.20, p = .043$, and their explicit scores $F(2, 221) = 4.23, p = .016$.

Post-hoc LSD tests indicated that implicit attitudes scores in the disease condition were significantly greater than the control, $t(158) = 2.30, p = .023$. Implicit scores in the disease condition were also higher than scores in the terror threat condition, $t(149) = 1.99, p = .048$, and no significant difference was seen between the control and terror threat conditions, $t(135) = .76, p > .250$. For explicit attitudes, scores in both the terror, $t(135) = 2.37, p = .019$, and the disease condition, $t(156) = 2.79, p = .007$, were significantly higher than the control condition. No difference was seen between explicit scores in the terror and disease condition, $t(147) = .42, p > .250$. Throughout, the greatest increase in anti-black/pro-white bias was seen in the disease condition and this was especially true for males.

Table S2.10: Means and standard deviations for males and females implicit and explicit scores

Variable	Control		Terror		Disease	
<u>Male</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>
Implicit	.31	.40	.33	.36	.44	.36
Explicit	.47	1.20	1.00	1.38	1.11	1.60
<u>Female</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>
Implicit	.26	.42	.26	.34	.38	.36
Explicit	.33	1.07	.48	1.30	.54	2.02

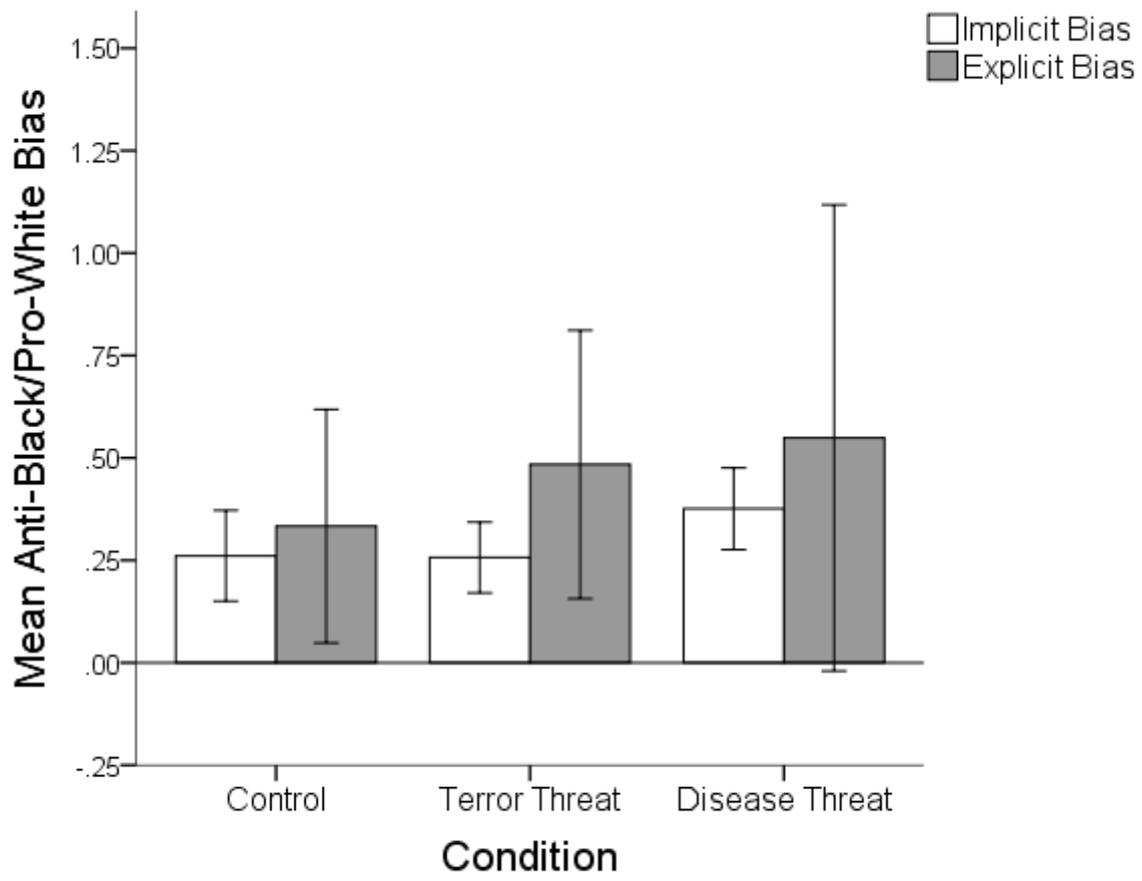


Figure S2.1: Females' implicit and explicit scores in the control, terrorism and disease conditions. 95% confidence intervals error bars have been included.

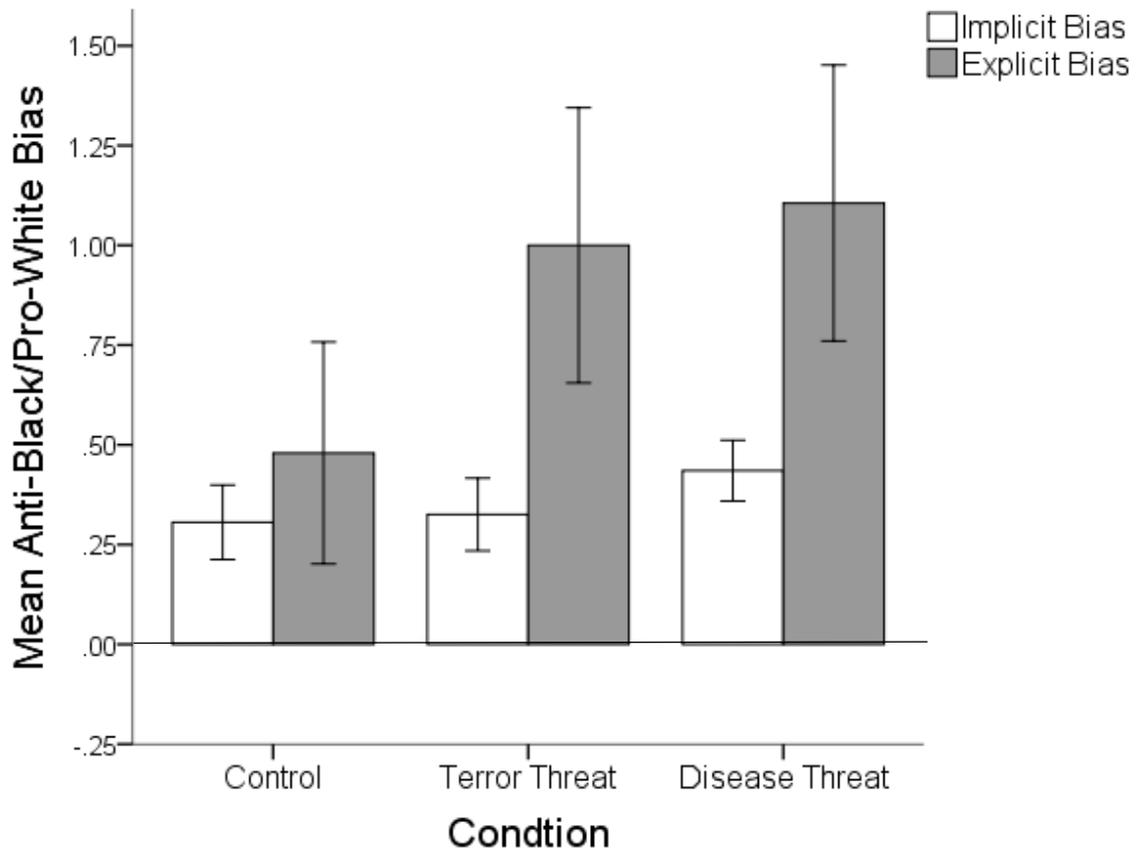


Figure S2.2: Males' implicit and explicit scores in the control, terrorism and disease conditions. 95% confidence intervals error bars have been included.

Appendix 2: Chapter 3

Table S3.1: The Non-word and Social System IRAP Stimuli (absent/weak associations)

Non-word IRAP				Social System IRAP			
Category 1:		Category 2:		Category 1:		Category 2:	
Cug or Tal		Vek or Pid		Capitalism		Socialism	
Target	stimuli	Target	stimuli	Target	stimuli	Target	stimuli
congruent with		congruent with		congruent with		congruent with	
category 1:		category 2 :		category 1:		category 2:	
Happy		Problems		Equality		Poverty	
Alive		Died		Fair		Inequality	
Positive		Negative		Community		Greed	
Freedom		Hated		Wealth		Dictator	
Care		Bad		Sharing		Awful	
Lucky		Sick		Freedom		Corrupt	
Response Option 1:		Response Option 2:		Response Option 1:		Response Option 2:	
True		False		True		False	

Table S3.2: The Nature and Weight IRAP Stimuli (strong associations)

Nature IRAP				Weight IRAP			
(Category 1: Flower)		(Category 2: Insect)		(Category 1: Thin Person)		(Category 2: Fat Person)	
Bluebell		Mosquito		6 different images of thin people		6 different images of fat people	
Daffodil		Spider					
Tulip		Wasp					
Target	stimuli	Target	stimuli	Target	stimuli	Target	stimuli
congruent with category 1:		congruent with category 2:		congruent with category 1:		congruent with category 2:	
Enjoy		Unpleasant		Good		Bad	
Cheer		Poison		Active		Sloppy	
Happy		Evil		Attractive		Ugly	
Lovely		Damage		Healthy		Mean	
Friend		Ugly		Popular		Lazy	
Pleasing		Hurt		Nice		Unhealthy	
Response Option 1:		Response Option 2:		Response Option 1:		Response Option 2:	
True		False		1: True		False	

The steps involved in calculating the D-IRAP scores were as follows (example given for the Weight IRAP version):

(1) Only response latency data from the six test blocks were used; (2) latencies above 10,000ms were discarded; (3) if latencies from more than 10% of a participant's trials throughout the 6 test blocks were less than 300ms, that participant was removed from the analysis; (4) for each IRAP task, 12 standard deviations for the four trial type latencies (e.g., **Thin Person**-Congruent responses, **Thin Person**-Incongruent responses, **Fat Person**-Incongruent responses, and **Fat Person**-Congruent responses) were calculated: four for the responses latencies from Test Blocks 1 and 2, four from the latencies from Test Blocks 3 and 4, and a further four from Test Blocks 5 and 6; (5) 24 mean latencies were then calculated for the four trial types in each of the six test blocks; (6) difference scores were calculated for each of the four trial types by subtracting mean latencies of the pro-**Thin** trials from mean latencies of the pro-**Fat** trials for each test block pair; (7) each difference score was then divided by its corresponding standard deviation from step 4, yielding 12 D-IRAP scores, one score for each trial type for each pair of test blocks; (8) four overall trial type D-IRAP scores were calculated by averaging the three scores for each of the four trial types across the three pairs of test blocks. These calculations revealed the absolute/non-relative results. To compute the relative comparison, equivalent to that of the IAT, an overall D-IRAP score was calculated by averaging all the 12 trial type D-IRAP scores obtained in step 7 above.

Assessing attitudes obtained for the individual IRAPs

Each of the four IRAPs (Non-word, Social System, Nature, and Weight) was analysed separately using the estimates of implicit attitude scores to address more directly whether participants had particularly positive attitudes to categories presented with positive words in the positive and standard framing conditions compared to the negative framing condition.

Non-word IRAP

The absolute D-IRAP (estimate) data (see Figure S3.1; Top Left) for the Non-word IRAP were analysed with a 2 (stimulus category: category 1, category 2) \times 2 (word valence: positive, negative) \times 3 (framing condition: positive, standard, negative) mixed ANOVA. Stimulus category 1 consisted of the Non-words Cug and Tal, category 2 consisted of the Non-words Vek and Pid. Framing condition was a between-subjects factor, stimulus category and word valence were within-subjects factors.

This analysis revealed a significant main effect of IRAP word valence, $F(1, 57) = 59.82$, $p < .001$, $\eta^2 = .52$; scores were higher for positive IRAP words than for negative IRAP words. There was also a significant main effect of framing condition, $F(2, 57) = 42.86$, $p < .001$, $\eta^2 = .61$. The standard framing had a high mean D-IRAP (estimate) score ($M = .22$); the positive framing produced the highest mean D-IRAP (estimate) score ($M = .45$), and negative framing had the lowest ($M = -.29$). LSD pairwise comparisons showed that all the framing condition D-IRAP (estimate) scores differed significantly from each other (all $t_s > 2.70$, $p_s < .01$). The main effect of IRAP category was not significant, $F(1,57) = .10$, $p > .05$, $\eta^2 = .00$, and nor were any of the interactions (all $F_s < 2.03$, $p_s > .14$).

Follow-up t-tests tested whether each absolute D-IRAP (estimate) score differed from zero. A positive value indicates a positive attitude to the stimulus category and a negative value indicates a negative attitude. For the standard framing condition, scores were greater than zero when associating the Non-word stimuli with positive words (all $t_s > 5.19$, $p_s < .001$). Scores

did not differ from zero for associating Non-word stimuli with negative words (all $t_s < .69$, $p_s > .50$). For the positive framing condition, all D-IRAP (estimate) scores were significantly greater than zero (all $t_s > 3.12$, $p_s < .01$). For the negative framing condition scores were significantly below zero for associating Non-words with negative words (all $t_s > 3.44$, $p_s < .01$) but did not differ for associating Non-words with positive words (all $t_s < 1.58$, $p_s > .13$)³³.

Social System IRAP

The absolute D-IRAP (estimate) Social System results showed a similar pattern to those of the Non-word IRAP and are shown in Figure S3.1 (Top Right). A 2 (stimulus category: Capitalism, Socialism) \times 2 (word valence: positive, negative) \times 3 (framing condition: standard,

³³ Two pilot studies were conducted that used various response options to ensure that the word frequency or length of these response options was not causing the PFB. Pilot Study 1 tested 20 participants and was similar in structure to the standard condition. Each participant completed three IRAPs which incorporated various combinations of Non-words and the positive and negative target words were taken from Barnes-Holmes, Barnes-Holmes, Power, Hayden, Milne, and Stewart (2006). The response options used in each IRAP were: ‘Similar’ and ‘Opposite’; ‘Similar’ and ‘Different’; ‘True’ and ‘False’. Pilot study 2 had two conditions, standard and negative framing, similarly to the current experiment, with fifteen different participants in each condition. Various Non-words were used and the same positive and negative words used in the Non-word IRAP above were presented. Each participant completed 5 IRAPs that had different response options (i.e. ‘Confirmation’ and ‘No’; Symbol of a Thumb Up and a Thumb Down; Picture of a Happy Face and a Sad Face; ‘Similar’ and ‘Opposite’; ‘Not Different’ and ‘Different’). These pilot studies produced comparable results to the Non-word IRAP.

positive, negative) mixed ANOVA revealed a significant main effect of IRAP word valence, $F(1, 57) = 27.51, p < .001, \eta^2 = .33$; positive IRAP words had higher scores than did negative IRAP words. There was also a significant main effect of framing condition, $F(2, 57) = 17.46, p < .001, \eta^2 = .38$. A high mean D-IRAP (estimate) score was found for standard framing ($M = .16$), with the positive framing producing the highest ($M = .34$), and negative framing producing the lowest ($M = -.14$) values. LSD pairwise comparisons showed that all the framing condition D-IRAP (estimate) scores differed significantly from each other (all $t_s > 2.24, p_s < .05$). The main effect of IRAP category was not significant, $F(1,57) = 1.76, p > .05, \eta^2 = .03$, and all the interactions were non-significant (all $F_s < 2.46, p_s > .09$).

Follow-up t-tests showed that in the standard condition scores were greater than zero for associating Capitalism and Socialism with positive words (all $t_s > 3.45, p_s < .01$), but scores did not differ from zero for associating Capitalism and Socialism with negative words (all $t_s < .44, p_s > .67$). In the positive framing condition, all but the Capitalism negative word association D-IRAP (estimate) scores ($t = 1.81, p = .09$) were significantly greater than zero (all $t_s > 3.91, p_s < .01$). In the negative framing condition scores were significantly below zero for associating Capitalism and Socialism with negative words (all $t_s > 2.64, p_s < .05$) but did not differ for associating Capitalism and Socialism with positive words (all $t_s < 1.61, p_s > .13$).

Nature IRAP

A 2 (stimulus category: Flower, Insect) \times 2 (word valence: positive, negative) \times 3 (framing condition: standard, positive, negative) mixed ANOVA showed a significant main effect of IRAP word valence, $F(1,57) = 49.98, p < .001, \eta^2 = .47$: positive IRAP words had higher scores than did negative IRAP words (see Figure S3.1; Bottom Left). This time a significant main effect was found for stimulus category, $F(1,57) = 76.67, p < .001, \eta^2 = .54$: more positive attitudes were found for flowers than for insects. A significant main effect of framing condition was also found, $F(2, 57) = 9.01, p < .001, \eta^2 = .24$. A mean D-IRAP

(estimate) score of .12 was found in the standard framing conditions. This score was elevated in the positive framing condition ($M = .22$), and reduced in the negative framing condition ($M = -.01$). LSD pairwise comparisons showed that the standard framing D-IRAP (estimate) score was significantly different from the negative framing score ($t = 2.30, p < .05$), and marginally different from the positive framing score ($t = 1.94, p < .06$). The positive and negative framing D-IRAP (estimate) scores also differed from each other ($t = 4.25, p < .001$). None of the interactions were found to be significant (all $F_s < 1.40, p_s > .24$).

Follow-up t-tests showed that in the standard condition scores were significantly greater than zero for associating flowers with positive words ($t = 7.69, p < .001$), and below zero for associating insects with negative words ($t = 2.28, p < .05$), but scores did not differ from zero when associating flowers with negative words or insects with positive words (all $t_s < 1.54, p_s > .14$). In the positive framing condition, scores were greater than zero for the flower and insect categories presented with positive words as well as the flower category presented with negative words (all $t_s > 2.75, p_s < .05$). The score for the insect category presented with negative words was significantly below zero ($t = 2.36, p < .05$). For the negative framing condition, scores were significantly above zero for the flower category presented with positive words ($t = 5.87, p < .001$) and below zero for the insect category presented with negative words ($t = 5.02, p < .001$). The other two D-IRAP (estimate) scores were not significantly different from zero (all $t_s < 1.10, p_s > .29$).

Weight IRAP

A 2 (stimulus category: Thin Person, Fat Person) \times 2 (word valence: positive, negative) \times 3 (framing condition: standard, positive, negative) mixed ANOVA showed a significant main effect of IRAP word valence $F(1, 57) = 26.86, p < .001, \eta^2 = .32$: positive IRAP words produced higher scores than negative IRAP words (see Figure S3.1; Bottom Right). A significant main effect was found for stimulus category, $F(1, 57) = 37.79, p < .001, \eta^2 = .39$:

more positive attitudes were found for the Thin Person category than for the Fat Person category, and a significant main effect was also found for framing condition, $F(1, 57) = 3.28$, $p < .05$, $\eta^2 = .10$. A positive mean D-IRAP (estimate) score was found in the standard framing condition ($M = .16$); this score was higher in the positive framing condition ($M = .22$) and lower in the negative framing condition ($M = .07$). LSD pairwise comparisons revealed that only the positive and the negative framing condition D-IRAP (estimate) scores differed from each other ($t = 2.56$, $p < .05$). The remaining interactions were non-significant (all F s $< .83$, p s $> .44$).

Follow-up t-tests showed that in the standard framing condition the absolute D-IRAP (estimate) scores for the Thin Person category presented with positive words and Thin Person category presented with negative words were significantly greater than zero (all t s > 2.68 , p s $< .05$). The scores for the Fat Person category presented with negative words was significantly below zero ($t = 2.17$, $p < .05$) and the scores for the Fat Person category presented with positive words did not differ from zero ($t = .371$, $p > .05$). In the positive framing condition the scores for Thin Person positive words, Thin Person negative words and Fat person positive words were all significantly greater than zero (all t s > 2.31 , p s $< .05$). Score for Fat Person negative words did not differ from zero ($t = 1.24$, $p > .05$). In the negative framing condition the scores for the Thin Person category presented with positive word was greater than zero ($t = 3.99$, $p < .01$) and the scores for the Fat Person category presented with negative words was below zero ($t = 3.08$, $p < .01$). Both Thin Person negative word and Fat Person positive word scores did not differ from zero (all t s < 2.05 , p s $> .05$).

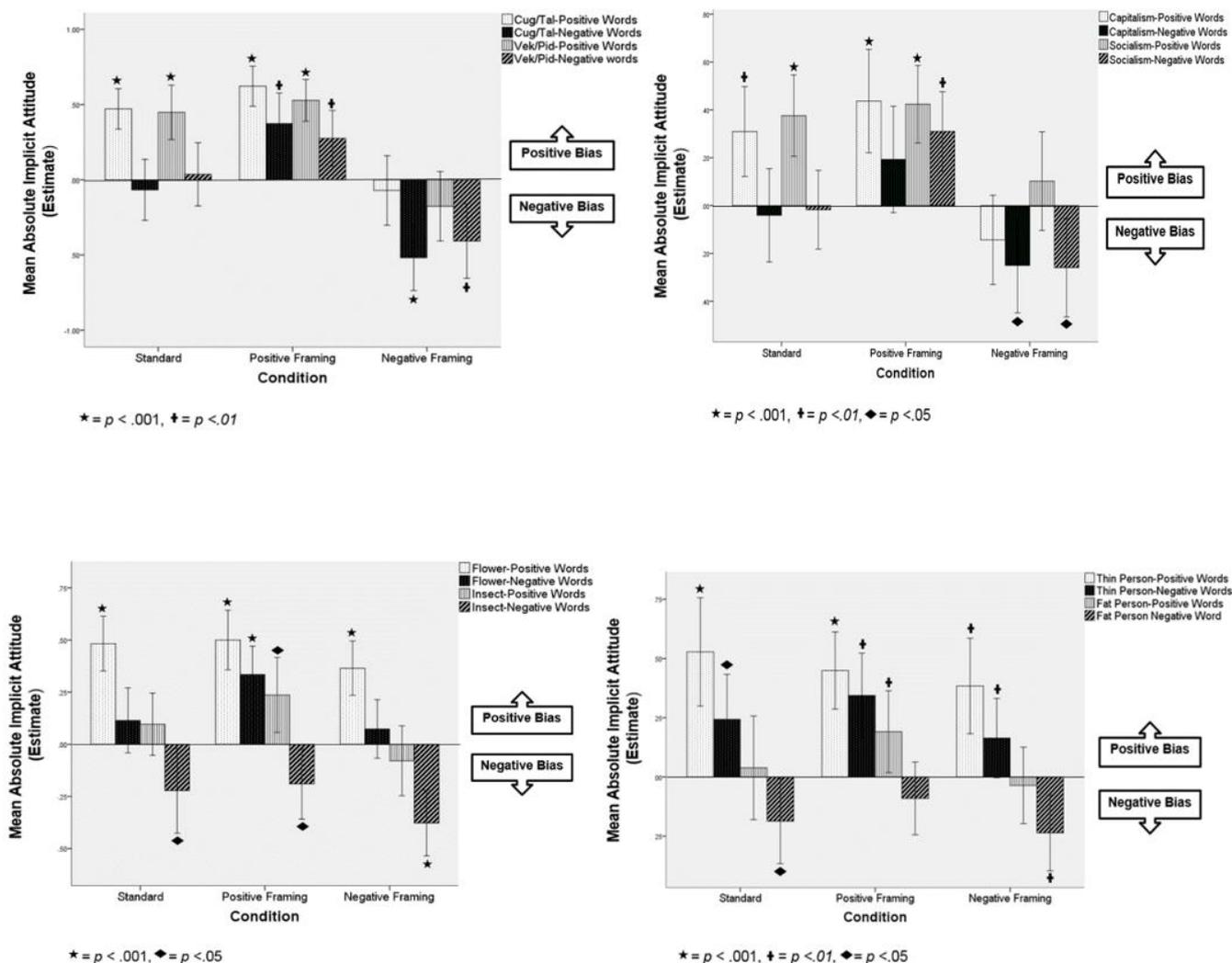


Figure S3.1: (Top Left): The four mean absolute Non-word D-IRAP (estimate) scores for each of the three framing conditions. Bars above the zero line indicate a positive attitude towards the category under investigation. This occurs when *True* responses are faster than *False* responses for categories presented with positive words and also when *False* response are faster than *True* response for categories presented with negative words. Points below the zero line indicate a negative attitude. This occurs when *False* responses are faster than *True* response for positive word associations and also when *True* response are faster than *False* response for negative word associations. Error bars with 95% confidence intervals have been included. Error bars that cross the zero mark indicate a statistically neutral attitude; those that do not cross zero indicate a significant positive or negative attitude (see stars).

(Top Right): The four mean absolute Social System D-IRAP (estimate) scores for each of the framing conditions.

(Bottom Left): The four mean absolute Nature D-IRAP (estimate) scores for each of the framing conditions.

(Lower Right): The four mean absolute Weight D-IRAP estimate scores for each of the framing conditions.

Appendix 3: Chapter 4

- Study 1: Qualtrics demographic and explicit questionnaire link.

https://warwickpsych.qualtrics.com/jfe/form/SV_agEITuIdboAbPeJ

- Study 2: Informed consent, demographic information and questionnaire (see below).

Informed Consent

Study Title: Reaction Time Test

Experimenters: Brian O'Shea, PhD Candidate Email: b.oshea@warwick.ac.uk

Supervisor: Dr. Gordon Brown Lecturer Email: G.D.A.Brown@warwick.ac.u

Description of Experiment: This experiment is a computer based task which requires participants to quickly associate positive and negative words towards different categories. The words Murderer and Rapist will be used in this study. A short questionnaire will be completed.

In order to participate in this research study, it is necessary that you give your informed consent. By signing this you are indicating that you understand the nature of the research study, your role and that you agree to participate in the research. Please consider the following points before signing:

- I understand that I am participating in psychological research;
- I understand that my identity will not be linked with my data, and that all information I provide will remain confidential;
- I understand the research team will use anonymised data and quotes in presentations and publications;
- I understand that I will be provided with an explanation of the research in which I participated;
- I understand that participation in research is not required, is voluntary, and when the study has begun, I may refuse to participate further without penalty or prejudice.
- I understand the anonymised data will be archived, to enable follow-up research, and for training future researchers.

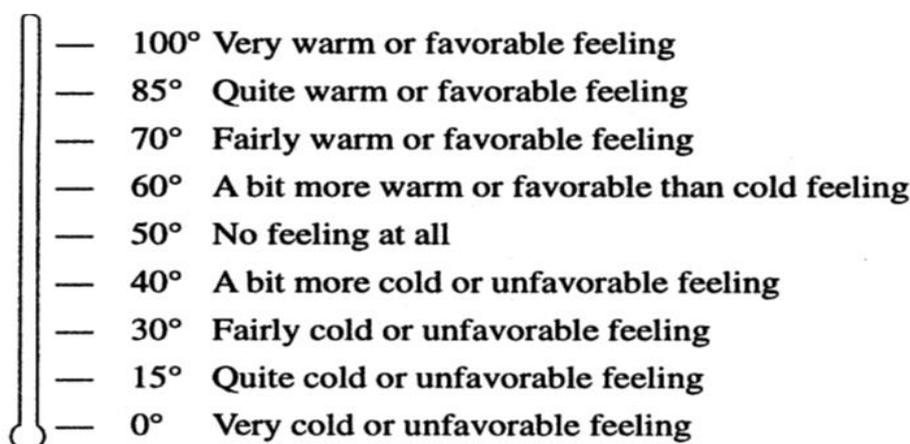
By signing this form I am stating that I am over 18 years of age, and that I understand the above information and consent to participate in this study being conducted.

Signature: _____
(of participant)

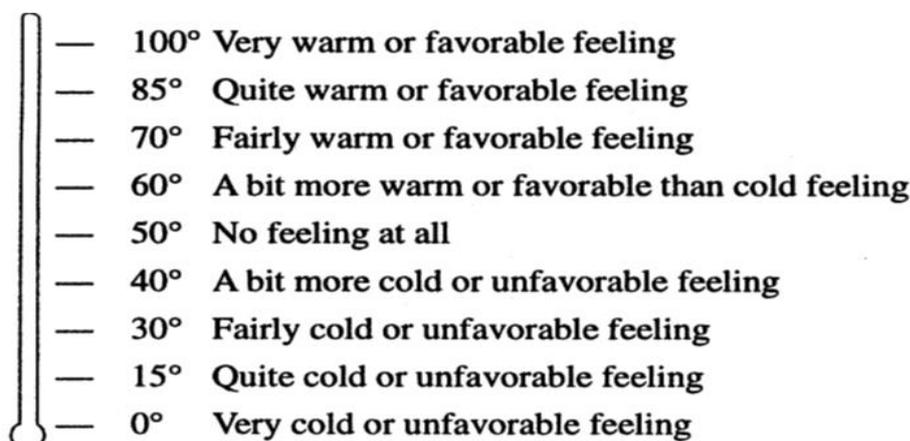
Today's Date: _____

Age: _____ Nationality: _____ (EXP.4)	Sex: M / F (circle) Ethnicity (e.g. White British, Black British, Mixed, Asian Chinese): Write Ethnicity here: _____
--	--

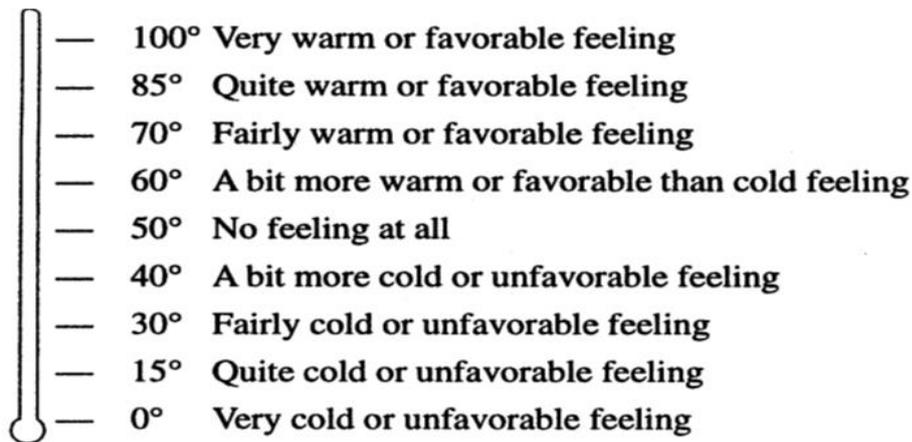
Please rate how positive or negative you feel towards a **Criminal?** (Mark somewhere on the thermometer).



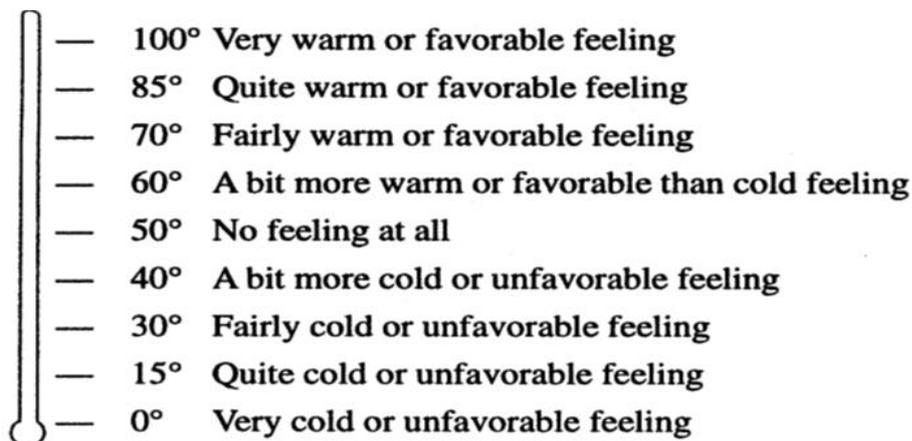
Please rate how positive or negative you feel towards a **Carer?** (Mark somewhere on the thermometer).



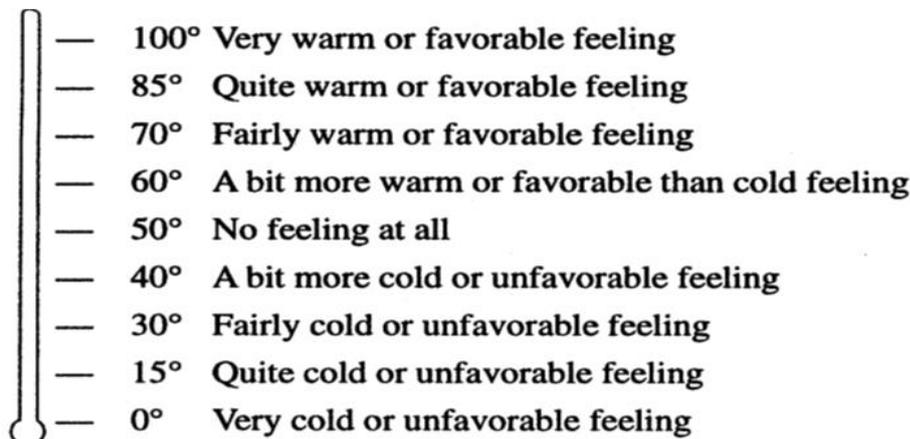
Please rate how positive or negative you feel towards **words with ‘O/o’?** (Mark somewhere on the thermometer).



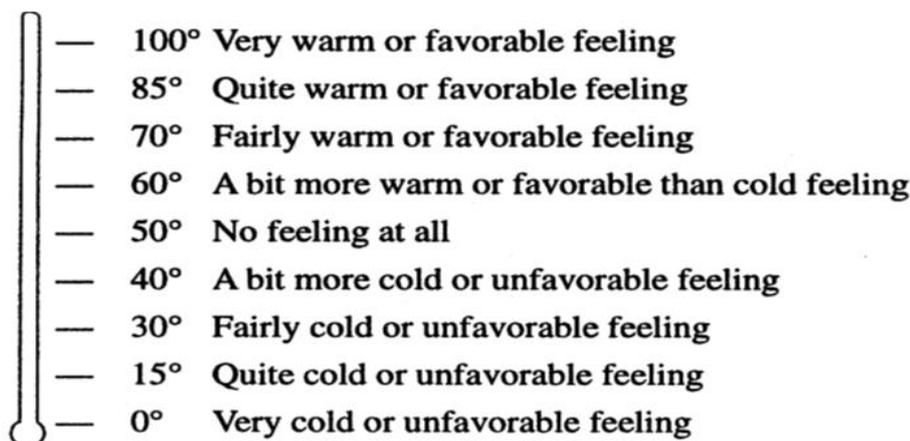
Please rate how positive or negative you feel towards words with **‘E/e’?** (Mark somewhere on the thermometer).



Please rate how positive or negative you feel towards **Insects?** (Mark somewhere on the thermometer).



Please rate how positive or negative you feel towards the word **Flowers?** (Mark somewhere on the thermometer).



Appendix 4: Chapter 5

Study 2 Replication (Amazon Turk Sample)

This study replicated Study 2 but recruited participants from a different source in which all participants receive a monetary reward for taking part. This study also used a fully randomised design.

Method

Participants: 222 participants (125 males, 95 females, 1 other and 1 declined to disclose gender) were recruited through Amazon Mechanical Turk and received \$1 for their participation. The average age of the participants was 35.9 years ($SD = 10.59$). Only participants from the US were selected to take part, 79% of the sample were white and the majority had an educational degree above high school level (>85%).

Materials: The materials were the same as those in Study 2 with an additional question in the demographic information section to obtain the participants' Amazon worker ID.

Procedure: The procedure was essentially the same as that of Study 2 apart from the recruitment method, provision of monetary payment for participation and the use of a fully randomised design with respect to the word type factor.

Results

H1 nouns vs. verbs vs. adjectives. A mixed 3×2 ANOVA with word type as the between-subject factor and valence as the within-subject factor revealed the presence of a significant valence \times word type interaction, $F(2, 209) = 3.72, p = .026, \eta^2 = .03$. Positive nouns were responded to more quickly than negative nouns but the reverse held for adjectives (replicating the results of Study 1 and 2), and verbs acted more like adjectives than nouns (replicating the results of Study 2). The main effect of valence approached significance, $F(1, 209) = 3.67, p = .057, \eta^2 = .02$, and the main effect of word type was non-significant, $F(2, 209) = .834, p > .250, \eta^2 = .00$ (see Figure S5.1).

H2 affirming words vs. negating words. A 3×2 mixed ANOVA with word type as the between subject factor and response option (support/affirm, oppose/negate) as the within subject factor, did not find the expected main effect of response option, $F(1, 211) = 2.03, p = .156, \eta^2 = .01$, or a main effect of word type, $F(1, 211) = .56, p > .250, \eta^2 = .00$. However, there was a significant word type \times response option interaction, $F(2, 211) = 6.06, p = .003, \eta^2 = .05$. As shown in Figure S5.1, participants were faster to respond to affirming words than to negating words when performing the task with nouns, $t(73) = 2.77, p = .007$, as well as during the task with adjectives, $t(70) = 2.17, p = .033$ but not during the verb task, $t(68) = 1.72, p = .09$. Of note, if stricter data trimming procedures are used that cut participants who have mean RTs that are 3 SD above and below the global mean, then the predicted main effect of response option is found, $F(1, 202) = 7.95, p = .005, \eta^2 = .04$. No significant main effect of word type or interaction were shown, $F_s > 2.36, p_s > .097$.

H3 affirming-positive, negating-negative association biases. Applying the D-IAT algorithm, 22 participants were removed for not performing to the required criteria. Analysing the remaining D-IAT scores, we again found a significant difference between sorting *affirming and positive words* together and *negating and negative words* together rather than the reverse sorting task, $t(201) = 40.41, p < .001, d = 2.84$. Again, participants were faster to sort affirming and positive words ($M = 1242.08, SD = 168.88$). then to sort negating and negative words ($M = 1253.792, SD = 171.92; t(211) = 2.43, p < .05, d = 0.17$).

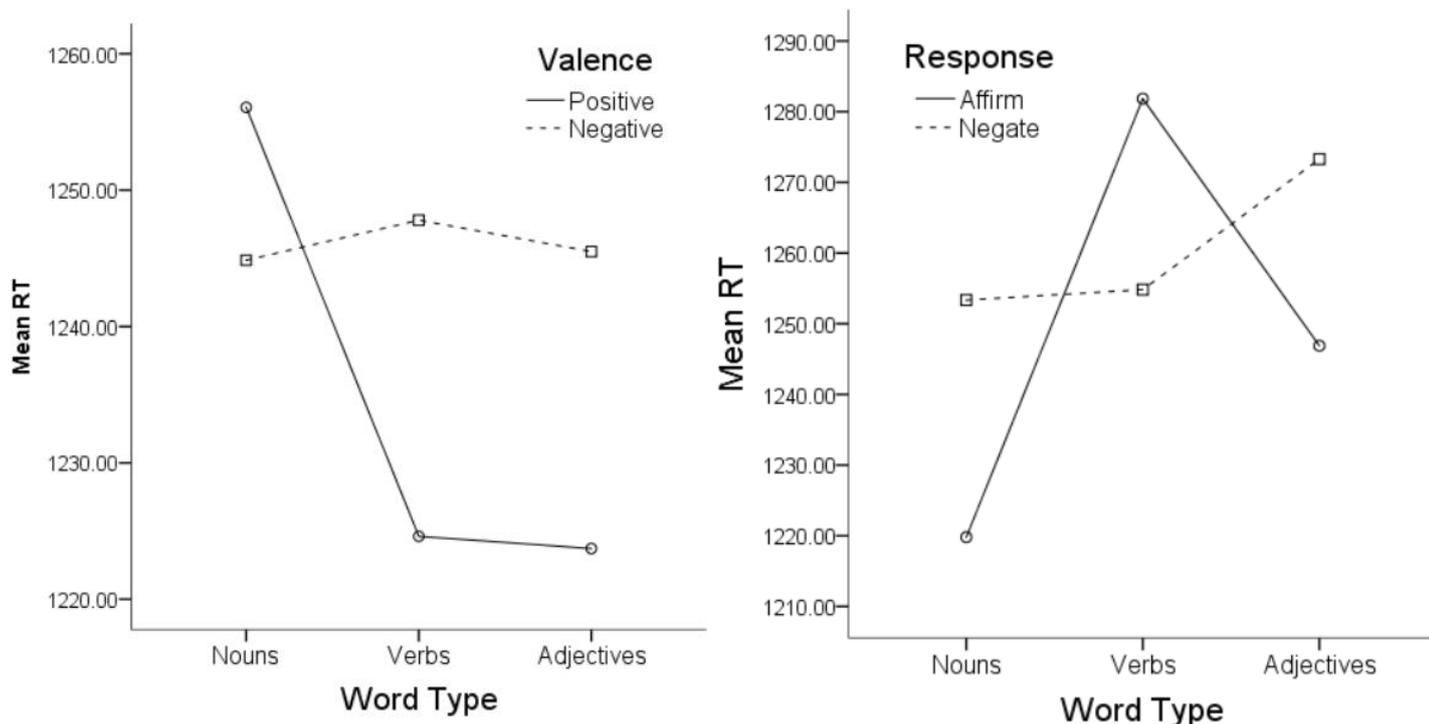


Figure S5.1: (left) Interaction between word type and valence. (right) Interaction between response type and word type.

Comparing the demographics and participants' performance between Study 2 (volunteered Reddit sample) and this replication (paid Amazon Turk sample) might indicate why the current study did not exactly replicate Study 2. We used a 2×2 chi-square test of independence to determine if the number of participants removed from the sample based on the D-IAT algorithm was equally distributed between Study 2 and this replication. The test revealed that participants were significantly more likely to be removed from the analysis in the Amazon Turk Sample than in the Reddit sample, $X^2(1, N = 390) = 8.71, p = .003$. When analysing the variability in RTs (3 standard deviations above minus 3 standard deviation below the mean) to each of the word types, we found that the Amazon Turk sample had more variability in RTs when sorting the positive and negative words, $t(10) = 11.20, p < .001, d = 6.47$, and the affirming and negating words, $t(5.80) = 13.94, p < .001, d = 8.05$. Of note, the

mean age of the Amazon Turk participants was higher than the Reddit participants ($M = 35.9$, $SD = 10.59$ vs $M = 24.94$, $SD = 6.86$), $t(378.66) = 12.29$, $p < .001$, $d = 1.22$).

Discussion

This study replicated the main findings from Study 1 and Study 2 (H1 and H3), except that participants did not respond faster to affirming words than to negating words in the verb condition (H2). This result might reflect differences in age or motivation between the self-interested volunteer (Reddit) and paid (Amazon Turk) participants, with the latter producing more variable and less accurate responses. Importantly, faster responses towards affirming words than to negative words were shown in the noun and adjective condition.

Table S5.1: Stimuli used in the Race IAT

Category Labels	Stimuli
African American	6 greyscale images of black individuals (3 males, 3 females)
European American	6 greyscale images of white individuals (3 males, 3 females)
Good	Glorious, Wonderful, Joy, Love, Peace, Pleasure, Laughter, Happy
Bad	Terrible, Evil, Horrible, Agony, Nasty, Awful, Failure, Hurt

Table S5.2: Stimuli used in the Nature IAT

Category Labels A	Stimuli A
Flower	carnation, clover, daffodil, dandelion, gladiola, lily, pansy, sunflower
Insect	tarantula, bee, beetle, fly, grasshopper, hornet, mosquito, wasps
Pleasant	amazing, brilliant, fantastic, wonderful, pleasure, happy, joy, superb
Unpleasant	agony, terrible, brutal, disgusting, destroy, tragic, awful, hurt
Category Labels B	Stimuli B
Flower	aster, bluebell, iris, magnolia, orchid, petunia, poppy, violet
Insect	caterpillar, cricket, dragonfly, earwig, lice, maggot, bug, termite
Pleasant	splendid, great, friend, love, glorious, laughter, excellent, peace
Unpleasant	nasty, failure, evil, horrible, disaster, painful, sad, barbaric
Category Labels C	Stimuli C
Flower	buttercup, jasmine, marigold, peony, rose, shamrock, tulip
Insect	bedbug, moth, cockroach, flea, horsefly, locust, ant, tick
Pleasant	adore, delicious, enjoy, positive, fabulous, healthy, joyful, kind,
Unpleasant	abandoned, bleak, crushed, rotten, injure, hostile, suffer,

Appendix 5: Chapter 6

Qualtrics demographic and explicit questionnaire links

QA Terror: https://warwickpsych.qualtrics.com/jfe/form/SV_bxBbVnsL7auEC9f

QA Disease: https://warwickpsych.qualtrics.com/jfe/form/SV_eRFX468jlI2B1ch

QB Terror: https://warwickpsych.qualtrics.com/jfe/form/SV_9ABnTc9sjACppqD

QB Disease: https://warwickpsych.qualtrics.com/jfe/form/SV_7aO2Qg9GdvrQ9oh