AN INVESTIGATION INTO THE INFLUENCE OF TARGET CATEGORY MANIPULATION ON THE RESULTS OBTAINED IN THE IMPLICIT ASSOCIATION TEST (IAT) IN RACE AND GENDER DOMAINS

by

LARRY FRANK TOOKE

Submitted in partial fulfilment of the requirements for a Masters of Social Science Degree in the School of Psychology

University of Kwa-Zulu Natal, Pietermaritzburg, RSA

June 2008

Abstract

The Implicit Association Test (IAT) is a computer-based psychological test that measures implicit attitudes, stereotypes and beliefs. In an effort to better understand the applicability and limitations of the IAT researchers have investigated the effects of manipulating a variety of procedural variables that comprise the IAT, not least the IAT categories and the exemplars that are instances of those categories. This study investigated the effects of manipulating the IAT's target categories that define the attitudinal domain that the IAT measures. Experiments were devised to determine the IAT's sensitivity to minor and major semantic manipulations to its target categories while keeping exemplars and attribute categories constant. It was found that the IAT was sensitive to major semantic differences in its target categories, but was apparently insensitive to minor semantic category differences, implying that it is unable to discriminate between subtle distinctions in attitude. It was hypothesised that this latter finding could have been partly due to a temporary cognitive re-definition of the categories in accordance with the salient characteristics of the exemplars.

Acknowledgements

With thanks to Doug Mansfield who first introduced me to the IAT and to Kevin Durrheim who encouraged me along the way.

Table of Contents

Abstract ii

Acknowledgements iii

Table of Contents iv

List of Tables vii

List of Figures viii

Chapter 1: Introduction 1

Chapter 2: Introduction to the Implicit Association Test 4

IAT Terminology 4

IAT Procedure 5

IAT Scoring 8

Interpretation of the IAT Effect 9

Chapter 3: Literature Review 11

Measuring Cognition and Mental States 12

The IAT as a Measurement Instrument of Implicit Cognition 13

IAT Research 14

Applied Research Domains 14

Meta-Research 15

The Influence of Exemplar Manipulation on the IAT effect 16

The Influence of Category Manipulation on the IAT effect 22

Positioning this Research 27

Chapter 4: Aim and Rationale 30

Chapter 5: Methodology 32

Experiments in Target Category Manipulation 32

Minor Semantic Changes in the Target Dimension 32

Experiment 1 – Minor Semantic Changes in a Race/Culture Domain 33

Experiment 2 – Minor Semantic Changes in a Gender Domain 35

Major Semantic Changes in the Target Dimension 36

Experiment 3 – Major Semantic Changes across Domains 37

Exemplar Selection 38

Experimental Method and Procedure........39
Sample........39
Sample Details for Experiment 1........39
Sample Details for Experiment 2........39
Sample Details for Experiment 3........40
Measurement Instrument........40
Apparatus........42
Research Design........42
Procedure........46
Data Preparation........48
Data Analysis........49
Chapter 6: Results........50
Applied vs. Meta-research........50
Design Integrity........50
Statistical Integrity........52
Overview of Results Presentation........53
Results: Experiment 1........54
Hypothesis Testing: ANOVA Results........54
Tabular Summary........55
Graphical Representation........56
Additional Observations........58
Trend Comparison........58
Supplementary Results........59
Results: Experiment 2........60
Hypothesis Testing: ANOVA Results........60
Tabular Summary........61
Graphical Representation........62
Additional Observations........63
Trend Comparison........64
Supplementary Results........64
Results: Experiment 3........64
Hypothesis Testing: ANOVA Results........65
Tabular Summary........66

Graphical Representation ... 67
Additional Observations ... 69
Trend Comparison ... 70
Supplementary Results ... 70
Results Summary ... 71
Chapter 7: Discussion ... 73
Study Limitations ... 73
Interrogating the Results ... 74
Major Semantic Variations in the Target Dimension ... 74
Minor Semantic Variations in the Target Dimension ... 76
IAT Limitations and Capabilities ... 82
Extending this Study ... 83
Conclusion ... 84
References ... 85
Appendix A: Exemplars ... 89
Appendix B: Battery/Test Allocation Distributions ... 90
Appendix C: Between-Test Results ... 94
Appendix D: IAT Results for Experiment 3 IATs on converting the Female/Male IAT results to Male/Female IAT results through a reversal of sign. ... 97
Appendix E: Supplementary Between-Group Results ... 99

List of Tables

Table 6-1 Summary of Race/Culture IAT results: IAT means, grand mean and ANOVA statistics 56

Table 6-2 Race/Culture IATs: Supplementary between-group ANOVA results 60

Table 6-3 Summary of Gender IAT results: IAT means, grand mean and ANOVA statistics 61

Table 6-4 Gender IATs: Supplementary between-group ANOVA results 64

Table 6-5 Summary of Race/Gender IAT results: IAT means, grand mean and ANOVA statistics 67

Table 6-6 Race/Gender IATs: Supplementary between-group ANOVA results 71

List of Figures

Figure 2-1 Schematic of IAT user interface..6

Figure 2-2 Schematic overview of IAT block presentation..8

Figure 6-1 Race/Culture IATs: Results by participant race..57

Figure 6-2 Race/Culture IATs: Results by participant race and gender 57

Figure 6-3 Gender IATs: Results by participant gender..62

Figure 6-4 Gender IATs: Results by participant race and gender.................................63

Figure 6-5 Race/Gender IATs: Results by participant race 68

Figure 6-6 Race/Gender IATs: Results by participant gender......................................68

Figure 6-7 Race/Gender IATs: Results by participant race and gender 69

Chapter 1: Introduction

The Implicit Association Test (IAT) is a computer-based psychological test pioneered by Greenwald and his colleagues (Greenwald, McGhee, & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2007) for the measurement of implicit attitudes, stereotypes and beliefs. Technically, the IAT is said to measure the relative strength of the automatic associations between pairs of concepts.

IAT scores are based on response latencies in a discriminative task in which exemplar stimulus words or images must be sorted to one of a pair of conceptual categories. The core IAT assumption is that rapid responses to the assignment task indicate a stronger association with a category, whereas slower responses indicate a weaker association(Greenwald et al., 1998; Nosek et al., 2007). By a procedural manipulation of the category-pairs in IAT tasks, differences in scores across tasks may be obtained. By applying a standard algorithm (Greenwald, Nosek, & Banaji, 2003) this difference may be converted into a score termed the *IAT effect* that gives an indication of the relative strength of association between the concepts. In practice the IAT effect is usually interpreted as a measure of attitude or preference.

The IAT is a relative newcomer to the world of psychological testing. It is now approaching 10 years since it was originally formally introduced to the psychological community (Greenwald et al., 1998). In the decade since its introduction, the IAT has generated a great deal of interest and debate. During this time IAT research has, broadly speaking, been divided into two main arenas: *applied research* which has investigated implicit attitudes in a wide variety of attitudinal domains, and *meta-research* [1]which has focussed on investigating the capabilities of the IAT itself. This latter research has been concerned with interrogating the IAT's validity and measuring the effects of manipulating a variety of experimental conditions or variables, not least the conceptual categories and exemplars which delineate the attitudinal domain that the IAT is being used to evaluate. Meta-research has typically been conducted either in an effort to discredit the IAT or in an attempt to better understand its limitations, its potential scope and the areas of its applicability.

This study is located in the area of meta-research and is conducted in a pioneering spirit - in an attempt to further the research community's understanding of the IAT's applicability and limitations. This research concentrates on determining the sensitivity of the IAT to semantic changes in the IAT's target categories while attribute categories and all exemplars (instances of the categories) are kept constant.

What this means in practice is difficult to appreciate in the abstract and is best understood by means of an example. (The reader who is unfamiliar with the IAT is encouraged to try the test online at https://implicit.harvard.edu/implicit/demo/ before proceeding further). Consider an IAT that might be devised in order to investigate attitudes towards race. Such an IAT might have a racial target dimension with category labels *Black* and *White* and an attribute dimension with category labels *Good* and *Bad*. The exemplars for the IAT's target categories might be typical names of *Blacks* and *Whites*. Exemplars for the attribute categories would

[1] The term *meta-research* in this thesis should not be confused with *meta-analysis*, the statistical technique that combines the results of several studies that address similar hypotheses. Rather the term *meta-research* is used here to distinguish between research that focuses on investigating the validity, limitations and capabilities of the IAT itself as opposed to *applied research* that uses the IAT to research implicit attitudes towards particular social objects.

be words that are associated with *Good* (such as *happy*, *pleasure*) and *Bad* (such as *hurt*, *pain*). Now consider a second IAT, identical to the first IAT in all respects, except that instead of a racial target dimension of *Black* and *White*, it has a more culturally oriented emphasis with target categories *African* and *Western.* Both IATs share the same attribute dimension and the same set of exemplars (typical names of *Blacks* and *Whites* for the target dimension and words associated with *Good* and *Bad* for the attribute dimension). The question of interest is whether or not these two tests will produce similar or divergent results to one another. That is, to what degree is the IAT sensitive to differences in the target categorisation frame? Naturally, changes to the target dimension may be fairly subtle, such as a change from a racial to a cultural dimension, or more obvious such as a change from a racial to a gender dimension (if exemplars are suitably chosen to permit this). It may be that the IAT is capable of detecting differences in the latter case, but not the former, or in other words, the IAT may have a limited sensitivity to subtle or minor semantic categorisation shifts, while detecting more dramatic or major semantic changes to its target categories.

This study seeks to investigate the limits of the IATs sensitivity to changes in the target categorisation frame in an effort to extend the research community's understanding of the test's capabilities and limitations. The attitudinal domains used in this investigation are those of race and gender, two domains frequently evaluated in IAT research, with a wealth of applied findings. To those new to the IAT, much of this introduction is likely to have been elusive. The test is best understood by walking through an example to explain its mechanics, assumptions and the basic rationale behind the claim that it is capable of measuring implicit attitudes. In the interests of lucidity, an introduction to the IAT, its terminology and its method of scoring and interpretation will be presented in the next chapter, prior to a more extensive literature review.

Chapter 2: Introduction to the Implicit Association Test

The Implicit Association Test is a psychological test that was designed to indirectly measure the relative strength of association between pairs of concepts. This relative score, known as the IAT effect may be seen as a measure of an individual's implicit attitudes, beliefs or preferences with reference to the attitudinal domain under investigation (Greenwald et al., 1998). This chapter seeks to clarify the preceding statements. It introduces IAT terminology, describes the IAT procedure and gives a brief explanation of how the IAT effect is calculated.

IAT Terminology

The IAT literature is not always consistent in its use of terminology, although there seems to be a generally accepted IAT parlance. This section covers the IAT terminology that is used in this study.

In the literature, mention is sometimes made of two IAT dimensions: the *target* and *attribute* dimensions. A *dimension* is defined by a pair of categories that are commonly the polar opposites of one another. The *target* dimension refers to the pair of categories that are targeted for comparison with one another and delineates the domain of interest. For example, *Black* and *White* would be the categories for the target dimension where race was the domain of interest. The *attribute* dimension refers to the pair of categories that define in what respect the target dimension will be compared or contrasted. For example, *Good* and *Bad* would be categories for the attribute dimension where the contrasting valence (or relative preference) of the target dimension categories was of interest. A short-hand for referring to how an IAT is defined is to list the categories of the target dimension followed by the categories of the attribute dimension thus: *Black/White*, *Good/Bad.*

Each IAT category has a set of *stimulus items* or *exemplars* which are words or images that represent or are instances of a category. In the race IAT, typical names or images of *Blacks* and *Whites* would be exemplars of the target categories. Words or images associated with *Good* (such as *happy, pleasure*) and *Bad* (such as *hurt, pain*) might be exemplars of the attribute categories. Each exemplar should have membership in only one of the four categories.

During the administration of the IAT, exemplars are presented on the computer screen one at a time in a series of tasks usually referred to as *blocks*. For each block, participants are required to sort exemplars to their parent categories. Some of the blocks present a simple task, in which exemplars must be sorted to one of two parent categories, either the target categories or the attribute categories. Two of the IAT blocks present a more complex task, in which exemplars must be sorted to a target-attribute category pair sometimes referred to as an *associative pair*. An associative pair is represented with a plus symbol as a conjunction between categories thus: *target category + attribute category*. In the race IAT, *Black+Bad* and *White+Good* are examples of associative pairs. In a scored block, a participant would be required to sort exemplars belonging to either the *Black* category or the *Bad* category to the *Black+Bad* associative pair and exemplars belonging to either the *White* category or the *Good* category to the *White+Good* associative pair.

IAT Procedure

The IAT is not a test that measures a specific construct, but rather represents a procedural format for measuring implicit cognition (Lane, Banaji, Nosek, & Greenwald, 2007). The construct to be measured is determined by the researcher and then represented in the IAT through the judicious selection of the categories that describe the target and attribute dimensions and of the exemplars that are instances of those categories.

During execution of the IAT, subjects are presented with a series of tasks or *blocks* in which they must classify exemplars into the categories to which they belong as rapidly as possible. The user interface is presented on the test taker's computer screen as in the figure below:

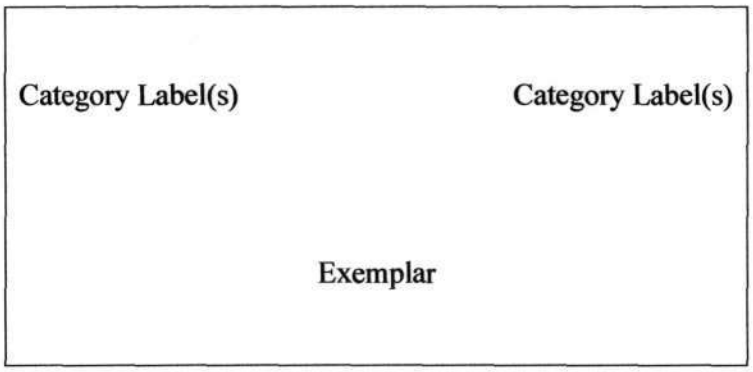

Figure 2-1 Schematic of IAT user interface

Category labels are displayed on the computer screen in the positions indicated for categories in the figure above. These may be single categories (e.g. *Black* on the left-hand side and *White* on the right-hand side) or an associative target-attribute category pair (e.g. *Black+Bad* on the left-hand side and *White+Good* on the right-hand side) depending on the IAT block being presented. Exemplars are displayed one at a time in a random sequence of *trials* in the bottom centre of the screen. When they appear, the subject must press the relevant key as quickly as possible to sort the exemplar either to the left or to the right category (or associative category-pair). When the subject makes a sorting error, an 'X' is temporarily displayed above the exemplar until the subject successfully completes the trial. The response time and whether or not a sorting error occurred are recorded for each trial.

An IAT participant is presented with five distinct IAT blocks for completion. Three of these are practise blocks to allow the participant to gain familiarity with the test categories and their associated exemplars. In these practise blocks the categories are singular. The two remaining blocks involve sorting exemplars to associative category pairs. Only the response times obtained in these blocks are used in the calculation of the IAT effect. Each scored block defines a set of two associative pairs, with the second block counterbalancing the first in the attribute category. (For example, in the race IAT if the first scored block had associative pairs *Black+Bad*, *White+Good*, the second scored block would have associative pairs *Black+Good*, *White+Bad*.)

The IAT procedure is best illustrated by means of an example. The schematic below indicates the sequence of blocks that must be completed during the execution of the race IAT. Block 1 is a practice task that requires participants to classify names into the singular target categories *Black* (left key press) or *White* (right key press). Block 2 is also a practice task but for the attribute categories *Bad* and *Good.* In block 3 the two prior tasks are combined into a set of associative category pairs. Now subjects must press the left key for exemplars belonging to the associative category-pair *Black+Bad* (that is, exemplars that belong to the *Black* category or to the *Bad* category are sorted to the left) and the right key for exemplars belonging to the associative category-pair *White+Good* (that is, exemplars that belong to the *White* category or to the *Good* category are sorted to the right). This is the first *scored block.* Block 4 is the reversal of block 2, with *Good* words now requiring the left key response and *Bad* words the right key response. This is a practice block. Block 5 is the combination of blocks 2 and 4 into a second set of associative category pairs. This is the same as block 3 except for the counter-balancing of the attribute categories. *Black+Good* now share the left key response with *White+Bad* sharing the right key response. This is the second scored block.

Block	Left key assignment	Right Key Assignment	Type
1	*Black* names	*White* names	Practice
2	*Bad* words	*Good* words	Practice
3	*Black* names + *Bad* words	*White* names + *Good* words	Scored
4	*Good* words	*Bad* words	Practice
5	*Black* names + *Good* words	*White* names + *Bad* words	Scored

Figure 2-2 Schematic overview of IAT block presentation

The basic assumption at the heart of the IAT is that "when two concepts that share a response are strongly associated, the sorting task is considerably easier than when two response-sharing concepts are either weakly associated or bipolar-opposed" (Greenwald et al., 2002, p. 8). In other words, when a participant finds that the associative paired categories are more compatible (or congruent) a more rapid sorting of the exemplars is facilitated, whereas incompatible (or incongruent) category associations will disrupt the sorting response, yielding slower response times.

What this means in the case of the race IAT is that the difference in the latency to respond to the first associative pairing (*Black+Bad* and *White+Good*) as compared to the second associative pairing (*Black+Good* and *White+Bad*) provides a measure of the relative strength of association between these two sets of pairings. If the first set (block 3) produces a *faster* response than the second (block 5) this indicates that the strength of association of (*Black+Bad* and *White+Good*) is *stronger* (or more congruent to the participant) than that of (*Black+Good* and *White+Bad*) implying an implicit preference for *White* over *Black*, that is an attitude that is pro-white.

IAT Scoring

For the IAT to measure the relative strength of association between concepts, it provides a procedural mechanism for evaluating a participant's associations of the target dimension with respect to the attribute dimension. Having explained the basic IAT procedure, it is a simple enough matter to explain how the IAT effect is calculated. Leaving aside the matters of how error responses are handled and how to deal with latencies that fall outside of a reasonably expected range, the IAT effect is computed simply enough by taking the difference in the mean response times between the two scored blocks and dividing by the standard deviation of the response times for both blocks combined (Greenwald et al., 2003). This effectively condenses all response data to a single statistic, analogous to Cohen's d statistic, the IAT effect (sometimes represented as D) which is a measure of the relative strength of association between the pairs of categories.

Interpretation of the IAT Effect

In practice the IAT effect is usually interpreted as an implicit measure of attitude with reference to the target domain under investigation. The creators of the IAT use a basic rule of thumb that is analogous to Cohen's definition of small, medium and large effect sizes. An IAT effect below 0.20 is interpreted to indicate the absence of a preference for either of the social objects of the target dimension, a score between 0.20 and 0.50 indicates a slight preference, scores between 0.50 and 0.80, reveal a moderate preference and scores in excess of 0.80, a strong preference (Nosek, personal communication, August 2002 in Blanton & Jaccard, 2006).

Usually, researchers who are using the IAT to investigate attitudes in particular social domains, are interested in determining whether mean IAT scores within a group differ significantly from the no preference zero base-line. Other applied researchers compare attitudes between groups having particular demographic characteristics by using statistical tests to compare group IAT means. The main thrust of this study is not primarily concerned with such comparisons. Rather, it focuses on variations between tests administered in a battery in order to ascertain the sensitivity of the IAT to semantic variations in its target dimension.

Having laid the foundation for a basic understanding of IAT terminology, procedure, scoring and the interpretation of the IAT effect, a closer inspection of the literature regarding implicit measures in general and the IAT in particular now follows.

Chapter 3: Literature Review

Since it began as a human science, psychology has been interested in measuring human behaviour and its determinants. Measurements taken from a sample population are frequently used to make inferences about the larger population from which the sample was drawn. In quantitative studies this usually involves the rigorous testing of hypotheses and the use of inferential statistics. The reasonableness and accuracy of these inferences depend upon the appropriateness of the measurement instrument that is used to obtain the data, or in psychometric terms, its reliability and validity. This study is concerned with a particular measurement instrument, the Implicit Association Test, which in the space of a single decade, has become the source of much interest, debate and publication in the psychological research community. The IAT has even spilled over into the public arena ("General Information," 2007) in newspapers such as the New York Times and the Washington Post and on television on The Discovery Channel, CNN and the popular talk show Oprah.

While there is a considerable body of research around the IAT, many questions remain unanswered regarding exactly what the IAT measures and through what cognitive processes it does so (De Houwer, 2006, in press; Fazio & Olson, 2003). Although the IAT has been widely researched, it applicability and limitations have not yet been fully investigated. It is incumbent on the research community to continue to investigate these questions to ensure that the IAT as a psychological instrument for the measurement of implicit cognitions is used in an accurate and responsible manner.

What follows is a description of the historical context out of which the IAT arose, a brief mention of the applied and meta-research conducted on the IAT by the research community and a more detailed review of one particular branch of IAT meta-research that investigates the influence of exemplar and category selection on IAT results. It is in this last area of meta-research, particularly with regards to the effect of semantic changes in IAT categorisation, that this study is located. Other studies that have investigated similar questions will be presented and this study's relationship to them explained. Finally, the manner in which this research might contribute to the determination of the IAT's applicability and limitations will be explained.

Measuring Cognition and Mental States

Cognitive and social psychologists have long been interested in the role of human mental states and their relation to behaviour, but have been faced with the difficulty of how best to obtain accurate measurements of such states. Initial attempts at measuring constructs such as attitudes, beliefs, stereotypes, values and motives relied largely on self-report questionnaires and other similar survey instruments (De Houwer, 2006; Kihlstrom, 2004). However, such psychometric instruments are limited by their reliance on the willingness and ability of respondents to disclose their personal views and feelings (Greenwald et al., 2002). As such, they frequently lack validity, especially when socially sensitive constructs are under investigation (Kihlstrom, 2004).

Deficiencies in self-report measurements led psychologists to search for alternative measurement strategies that did not require introspection. Advances in research into types and function of memory revealed the existence of an implicit or unconscious memory that was largely dissociated from explicit or conscious memory (Schacter, 1987). This implicit-explicit distinction was found to extend into a variety of cognitive domains including: perception, thinking, problem solving, learning and motivation and pointed to the possibility of circumventing the measurement difficulties inherent in self-report by directly accessing implicit cognitions (Kihlstrom, 2004). The recognition of this distinction led psychologists to create a variety of psychological tests that were aimed at bypassing explicit cognitive processes to access implicit cognitions.

The IAT as a Measurement Instrument of Implicit Cognition

A popular focus of research into implicit cognition has centred on the measurement of implicit attitudes, with particular attention devoted to investigating implicit stereotyping, prejudice and bias. Psychologists have devised a variety of measurement techniques in an effort to measure implicit attitudes. Of these the most widely used and researched is the Implicit Association Test (Fazio & Olson, 2003), a test which has provided considerable impetus to implicit research since its introduction a decade ago (Greenwald et al., 1998).

Defining implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favourable or unfavourable feeling, thought, or action toward social objects." (1995, p. 8), Greenwald and his colleagues designed the IAT claiming that it was capable of measuring "implicit attitudes by measuring their underlying automatic evaluation" (1998, p. 1464). This claim has led to a great deal of applied research into implicit social attitudes using the IAT, but also to a burgeoning body of literature that has attempted to critique the IAT and to determine what it in fact measures. While the early literature refers to the IAT as measuring implicit attitudes, later literature more precisely describes the IAT as measuring relative strengths of association (Greenwald et al., 2002). Despite this technical precision, in practice IAT results are usually interpreted as indicative of attitudes or preferences towards particular social objects. The assumptions underlying the IAT, its procedural features and the basic formula for the calculation of the IAT effect were described in Chapter 2 above.

IAT Research

Ever since Greenwald et al. (1998) published their first paper on measuring individual differences in implicit cognition, which formally introduced the IAT to the research community, there has been a steadily growing interest in the test as a means of measuring implicit attitudes. Over the last decade more than 200 academic papers and hundreds of conference papers have reported making use of the IAT and almost 5 million individual tests have been completed on the project implicit website (Lane et al., 2007).

The IAT has been used primarily within the disciplines of social and cognitive psychology (Fazio & Olson, 2003; Greenwald & Nosek, 2001; Kihlstrom, 2004) but has spread into other areas such as developmental psychology (Baron & Banaji, 2006), clinical psychology (Teachman, Gregg, & Woody, 2001), neuroscience (Cunningham et al., 2004; Phelps et al., 2000) and market research (Maison, Gregg, & Bruin, 2002).

Applied Research Domains

A wide variety of attitudinal domains have been investigated using the IAT including attitudes towards age (Jelenec & Steffens, 2002), weight (Rudman, Feinberg, & Fairchild, 2002), nationality (Greenwald et al., 1998), religion (Rudman et al., 2002), sexual orientation (Banse, Seise, & Zerbes, 2001) and many others. Applied research into social attitudes using the IAT typically reveals in-group over out-group preferences (Nosek et al., 2007).

A survey of the literature reveals that the domain most frequently investigated using the IAT is that of racial attitudes (Baron & Banaji, 2006; Dasgupta, McGhee, Greenwald, & Banaji, 2000; Greenwald et al., 1998; Mitchell, Nosek, & Banaji, 2003; Ottaway, Hayden, & Oakes, 2001; Smith-Mclallen, Johnson, Dovidio, & Pearson, 2006). Racial attitude research has repeatedly revealed a strong pro-white preference amongst white participants with a tendency to neutrality among black participants (Banaji, Nosek, & Greenwald, 2004; Dasgupta & Greenwald, 2001; Greenwald et al., 1998; Ottaway et al., 2001). Gender attitude research also features prominently in the literature (Aidman & Carroll, 2003; Greenwald & Nosek, 2001; Mast, 2004; Rudman & Goodwin, 2004) with women typically preferring females over males and men evidencing a more neutral attitude (Rudman & Goodwin, 2004).

This study is one of meta-research into the IAT. It is *not* primarily interested in examining the attitudes of a particular group towards social objects as is the norm in applied research. Rather it aims to examine the effects on the IAT of manipulating a single part of the IAT definition, the target dimension, in experiments having a within-subjects design. Race and gender were selected as the domains for this meta-research investigation in part because existing publications in these domains allow for an evaluation of the reasonableness of the IAT results obtained.

Meta-Research

In parallel with the applied research that investigates attitudes towards social objects, the research community has rigorously turned the microscope on the IAT itself (Lane et al., 2007; Nosek et al., 2007). Inter alia, researchers have investigated: the IAT's relationship to other implicit measures (Fazio & Olson, 2003; Greenwald & Nosek, 2001), convergence and divergence with explicit measures (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005; Nosek, 2005; Nosek, Banaji, & Greenwald, 2002; Olson & Fazio, 2003), the effects of changing various procedural elements of the IAT (Greenwald et al., 1998; Nosek, Greenwald, & Banaji, 2005), the analysis of and improvements to the IAT algorithm used in calculating the IAT effect (Greenwald et al., 2003), the malleability of IAT results through manipulation of the experimental context (Dasgupta & Greenwald, 2001; Lowery, Hardin, & Sinclair, 2001), and the effects of varying the categories and exemplars that comprise the IAT's content (De Houwer, 2001; Fazio & Olson, 2003; Govan & Williams, 2004; Mitchell, Nosek, & Banaji, 2003). With regards to this latter area of investigation, it has been of a matter of considerable debate amongst researchers as to whether it is the nature of the IAT categories or the features of the IAT exemplars that are responsible for influencing the IAT effect (Bluemke & Friese, 2006; Nosek et al., 2007). It is within this particular niche of investigation that this study is located. To gain a proper appreciation for how this research is positioned it is important to first review in some detail the research to date into the influence of category and exemplar manipulation on IAT results.

The Influence of Exemplar Manipulation on the IAT effect

A variety of experiments have investigated the effect of the manipulation of exemplars on IAT results. De Houwer (2001) noted that exemplars used in the IAT have relevant features that associate them with the category of which they are members, but also irrelevant features that might attract them to a category to which they do not belong. For example, in a *Flower/Insect* IAT with attribute dimension *Good/Bad*, the exemplar "cockroach" is relevantly a member of the *Insect* target category, but inasmuch as cockroaches are perceived negatively, it may be irrelevantly considered a member of the *Bad* attribute category. De Houwer pointed out that the IAT typically confounds relevant and irrelevant features: *Flowers* are generally evaluated as irrelevantly *Good* and many *Insects* are usually considered irrelevantly *Bad*. To test whether IAT performance is a function of relevant or irrelevant exemplar features, De Houwer devised an IAT with target dimension *British/Foreign* and an attribute dimension *Positive/Negative* for administration to a sample of British participants.

Exemplars for the *British* and *Foreign* target categories were comprised of a set of names, half having *Positive* valence (e.g. British positive = Princess Diana, Foreign positive = Einstein) and half having *Negative* valence (e.g. British negative = Rosemary West - a mass murderer, Foreign negative = Hitler). In examining the response data generated by the IAT, it was found that whether the individual *British* or *Foreign* exemplars had positive or negative valence had no appreciable effect on sorting times of these exemplars in the IAT scored blocks. This null effect of the irrelevant features of the exemplars led De Houwer to conclude that it is only the IAT categories and not the exemplars that influence the IAT effect.

Challenging De Houwer's conclusions, Mitchell et al. (2003, Experiment 2) and Govan and Williams (2004, Experiment 1) found that manipulating the valence of exemplars that belonged to the target categories *can* affect IAT results. These researchers conducted IAT experiments in the domain of racial attitudes, an area of investigation that has historically revealed a strong in-group preference amongst white participants (Banaji, Nosek, & Greenwald, 2004; Dasgupta & Greenwald, 2001; Greenwald et al., 1998; Ottaway et al., 2001). In a departure from De Houwer's method, these studies investigated the effect on IAT results on a white sample when *all* exemplars for the *Black* and *White* target categories were selected to affectively favour *Blacks* over *Whites*. In a within-subjects design, Mitchell et al. (2003) found a notably weaker pro-white IAT effect emerged for the IAT having admired *Black* and disliked *White* exemplars compared to the IAT with disliked *Black* and admired *White* exemplars. In a similar experiment, Govan and Williams (2004) had even more dramatic results, eliminating (but not reversing) the usual pro-white bias in an IAT having admired *Black* and disliked *White* exemplars. Significantly, these experiments differed from De Houwer's (2001) in that *all* exemplars in the *Black* target category were positive and all exemplars in the *White* target category were negative whereas in De Houwer's experiment half of the exemplars in each target category were positive and half were negative.

To explain their findings Mitchell et al. (2003) suggested that the experimental context influences exemplar evaluation. Accordingly, a preponderance of positive *Black* stimuli would activate positive mental associations towards blacks, leading to a facilitation of sorting the admired *Black* exemplars (as compared to neutral *Black* exemplars) to the *Black+Good* category-pair. Govan and Williams (2004) proposed a different hypothesis, that the atypical *Black* and *White* exemplars in their study had caused a temporary mental re-definition of the target categories, so that the *Black* exemplars that were presented were evaluated in terms of a sub-type of the *Black* category (in this case, '*nice Blacks*') with a resultant decrease in the time required to sort the exemplars to the *Black+Good* category-pair. Both groups of researchers noted that only a reduction or elimination and not a reversal of the typical IAT trend of a preference for whites over blacks was obtained. This was attributed to an enduring strong association with *White+Good* among participants, despite the negativity of the *White* exemplars.

Govan and Williams (2004, Experiment 2) conducted a further experiment to support their conclusion that exemplar valence had affected IAT scores through a temporary cognitive re-definition of the IAT categories. In this experiment, two separate samples were presented with an *Animal/Plant*, *Pleasant/ Unpleasant* IAT. The first group was presented with positive *Animal* and negative *Plant* target exemplars (e.g. puppy, poison ivy) and the second with negative *Animal* and positive *Plant* target exemplars (e.g. crocodile, daisy). Having completed the IAT and a brief filler task, both groups then repeated the test, but this time the target exemplars previously used for the *Animal* and *Plant* categories were replaced with the neutral words 'plant' and 'animal'. Results showed that the sample that had received the pro-Animal exemplars in their first IAT showed a preference for *Animal* over *Plant* in *both* IATs whereas the sample with the pro-Plant exemplars in their first IAT showed a preference for *Plant* over *Animal* in *both* IATs. Govan and Williams reasoned that in accordance with the valence of the target exemplars, the first sample had mentally re-defined the IAT target categories as *Nice Animals/Nasty Plants* and the second sample had re-defined the categories as *Nasty Animals/Nice Plants*. Further they suggested that this category re-definition had persisted when participants completed the second IAT despite the neutrality of its target exemplars. Taken together, Govan and Williams' (2004) experiments appear to give credence to their category re-definition hypothesis and as will be argued later has some relevance to the interpretation of results obtained in this study.

The experiments discussed above have all investigated the effects of manipulating exemplars related to the IAT's target dimension. Steffens and Plewe (2001), adopted a different approach in which they studied the effect of manipulating exemplars belonging to the IAT's attribute dimension. Using a counterbalanced *Masculine/Feminine, Pleasant/Unpleasant*, they conducted an experiment in the domain of gender attitudes in which they investigated the effect of exemplars that exhibit what they termed 'cross-category associations', that is, stereotypic associations with a target category[2]. For the first IAT, they selected pleasant attribute exemplars associated with the feminine stereotype (e.g. *beautiful*, *empathic)* and unpleasant attribute exemplars associated with the masculine stereotype (e.g. *violent, brutal*). For the second IAT they chose pleasant attribute exemplars stereotypical of males (e.g. *logical, independent)* and unpleasant attribute exemplars stereotypical of females (e.g. bitchy,

[2] In essence, cross-category associations are to attribute exemplars what De Houwer's (2001) irrelevant exemplar features are to target exemplars.

hysterical). Consistent with findings on gender based IAT's (Rudman & Goodwin, 2004), the results for both IAT's revealed a gender preference for women over men in a female sample, but there was a substantially weaker pro-Feminine IAT effect in the IAT having pleasant attribute exemplars that were positively associated with the *Masculine* target category and unpleasant attribute exemplars that were negatively associated with the *Feminine* target category. Steffens and Plewe (2001), concluded that manipulating exemplars in the attribute dimension to introduce cross category associations (stereotypic associations with a target category), can influence IAT results. As a consequence of their findings they cautioned that attribute exemplars should ideally be normed to ensure their relative neutrality in relation to the target dimension under investigation.

Finally, (Bluemke & Friese, 2006), conducted an experiment in which they manipulated target exemplars in a manner similar to Govan and Williams (2004) and attribute exemplars in a manner similar to Steffens and Plewe (2001). In their experiment they used *West/East* (Germany) as the IAT target dimension and *Positive/Negative* as the IAT attribute dimension. The sample consisted of West German participants. A number of different IATs were defined:

- A control IAT in which all exemplars of both target and attribute dimensions were neutral.
- An IAT having attribute exemplars with positive West Germany cross-category associations and negative East Germany cross-category associations while keeping target exemplars neutral.
- An IAT having positive-West Germany and negative-East Germany target exemplars while keeping attribute exemplars neutral.

- An IAT with target and attribute exemplars with pro-West Germany and anti-East Germany associations.
- Three IATs defined similarly to the three preceding IATs above, but favouring East Germany and derogating West Germany.

Bluemke and Friese (2006) found that they obtained the lowest IAT effect in the all neutral test, a larger IAT effect where the attribute exemplars were pro-West and anti-East, a still larger IAT effect where the target exemplars were pro-West and anti-East and the highest IAT effect where both target and attribute exemplars favoured the West and derogated the East. All of these IATs revealed a significant preference for West over East by the West German sample. In the three IATs that favoured the East over the West, a reduction in the IAT effect (as compared to the neutral test condition) occurred when attribute exemplars were manipulated in favour of East Germany and to the detriment of West Germany, a reversal of sign of the IAT effect resulted when target exemplars were pro-East and anti-West (although the magnitude of the IAT effect was small) and the greatest change in the IAT effect was in evidence when both target and attribute exemplars where simultaneously manipulated in favour of the East. In this last case a strong pro-East result was obtained in the West German sample. Results thus ranged from a strong in-group to a strong out-group preference even though IAT categories remained constant. Bluemke and Friese (2006) concluded that both attribute and target exemplars are capable of influencing the IAT effect.

The foregoing has served to summarise the research into the effects of manipulating either target or attribute exemplars or both on the IAT effect. Certainly the deliberate manipulation of either target or attribute exemplars has repeatedly been demonstrated to influence IAT results. However the degree of influence seems to be constrained by the nature of the domain under investigation. This is apparent in the experiments reviewed above. In the race and gender IATs, a manipulation of exemplars led to IAT effects being *reduced* but not *reversed* (Govan & Williams, 2004; Mitchell et al., 2003; Rudman & Goodwin, 2004). In the West/East German IAT (Bluemke & Friese, 2006) a negligible IAT reversal was achieved (except in the unusual circumstance where both target and attribute exemplars were manipulated simultaneously). Finally in the case of the Animal/Plant IAT, the IAT effect was easily manipulated either in favour of animals or of plants (Govan & Williams, 2004). It does appear that certain attitudinal domains are less susceptible to being influenced by exemplar manipulation than others. In this regard, De Houwer, who effectively began the investigation into the effects of exemplar manipulation with the suggestion that the valence of exemplars has little effect on IAT results (De Houwer, 2001) has since qualified his conclusions. He states: "When the categories are clearly positive or negative, category valence might be much more salient than the valence of the exemplars. Hence, exemplar valence might not have much effect on performance. But when the categories are fairly neutral, exemplar valence might be salient and have an effect on performance" (De Houwer, in press). This observation seems to best sum up the findings on the effects of exemplars on IAT results. Perhaps an afterword based on the work of Govan and Williams (2004) is worth adding, that the mechanism by which exemplar valence affects performance may be through a cognitive re-definition of the category to more accurately fit the exemplars.

The Influence of Category Manipulation on the IAT effect

Far less attention has been focused on the influence of category selection on IAT results than has been devoted to the effects of exemplar selection despite the acknowledgement that the category labels are more salient than the exemplars under usual test conditions (De Houwer, in press; Olson & Fazio, 2004) and critical in their role of constraining exemplar interpretation (Nosek et al., 2007). Indeed, only two published studies have investigated the effects of category manipulation on IAT results in original ways (Mitchell et al., 2003; Olson & Fazio, 2004). This is curious, because as Olson and Fazio (2004, p. 654) observe: "The IAT's operation at the level of the category label, instead of the individual exemplar, suggests that researchers should consider the category labels and not the individual exemplars to be the objects most directly relevant to the IAT". As it happens the two investigations into this question approach it from different angles. Olson and Fazio (2004) investigated the effects of modifying *attribute* category labels, while Mitchell et al. (2003) were interested in the effects of modifying the *target* category labels.

Olson and Fazio (2004) took as their point of departure a study conducted by Karpinski and Hilton (2001). These researchers demonstrated a divergence between an *Apple/Candy Bar, Pleasant/Unpleasant* IAT and an explicit measure of preference for apples versus candy bars. Moreover, unlike the IAT measure, the explicit measure predicted participant behaviour, in that, when offered a choice between an apple and a candy bar at the close of the experiment, participants tended to choose according to their explicit preference. Noting that society tends to view apples in a positive light whereas candy bars are portrayed with some ambivalence, and recognising that their experimental IAT results seemed in line with such a societal view, Karpinkski and Hilton (2001) proposed that it is perhaps "the associations a person has been exposed to in his or her environment, not that individual's level of endorsement regarding the attitude object" (2001, p. 786) that are measured by the IAT.

Drawing on Karpinkski and Hilton's (2001) work, and noting that the IAT appears to report higher rates of participant prejudice than do other implicit measures, Olson and Fazio (2004) proposed a related idea. They suggested that the IAT as a measure of personal attitude may be contaminated by what they refer to as extrapersonal associations, that is, associations that are available in memory as general knowledge but to which the individual has not personally acquiesced. This view supposes that "personal attitudes may stand in contrast to the valence implied by other information that individuals possess, such as cultural knowledge" (Olson & Fazio, 2004, p. 664). From this standpoint, Olson and Fazio argued that "given Blacks' negative portrayal by much of the media, even people for whom positivity is automatically activated in response to Blacks ought to have readily available in memory a host of negative associations with Blacks. This would inflate estimates of prejudice on the IAT if the IAT is contaminated by this general knowledge" (2004, p. 655). To 'decontaminate' the IAT in order to allow for a 'purer' measure of personal attitude, they suggested a simple modification – the replacement of the familiar *Pleasant/Unpleasant, Good/Bad,* or *Positive/Negative* categories of the IAT attribute dimension with the personalised attribute categories *I Like/I Don't Like*. Conducting a counterbalanced experiment in which they compared results obtained from the new personalised IAT with the traditional IAT in a racial attitude domain they found that the personalised IAT revealed less racial bias than did the traditional IAT (Olson & Fazio, 2004, Experiment 1). Also, in a repetition of Karpinkski and Hilton's (2001) apples versus candy bar experiment (Olson & Fazio, 2004, Experiment 3) the personalised IAT revealed an equal liking for apples and candy bars and correlated highly with explicit measures of liking and behaviour whereas the traditional IAT indicated a clear preference for apples over candy bars and failed to correlate with explicit measures. In their experiments Olson and Fazio demonstrated how re-defining the attribute categories may markedly affect IAT scores and concluded that the personalised IAT may be a more accurate measure of individual attitudes than is the traditional IAT.

The relevance of Olson and Fazio's personal-extrapersonal dichotomy and the conclusions drawn from their research have been challenged in a recent publication (Nosek & Hansen, 2005) but further evidence has also been offered in support of their extrapersonal hypothesis (Han, Olson, & Fazio, 2006). Whatever the outcome of this debate, it is clear that Olson and Fazio's (2004) manipulation of the attribute categories from their traditional labelling to a more personalised labelling was sufficient to significantly influence results obtained on the IAT in the domains under investigation. In the language of this study, a semantic variation in the attribute dimension of the IAT resulted in a significantly different IAT effect, even when exemplars and the target dimension were kept constant.

Of particular interest from the point of view of this study is a paper by Mitchell et al. (2003, Experiment 1) which describes their investigation into the effects of target category manipulation on IAT results. Referring to Dasgupta and Greenwald's (2001) research that demonstrated that pre-test exposure to positive African American exemplars resulted in white participants evaluating blacks less negatively in a race based IAT than did their control condition counterparts, Mitchell et al. argued that attitudes are contextually variable rather than inherently stable. With this as their point of departure, they designed an experiment to test whether the same exemplars would be evaluated differently if their group membership (or social context) was experimentally manipulated. In this experiment, disliked white politicians (e.g. Richard Nixon) and liked black athletes (e.g. Michael Jordan) were used as category exemplars in a within-subject counterbalanced IAT experiment. Both IATs had a constant attribute dimension (*Good/Bad*) and constant target and attribute exemplars. The IATs differed only in their target dimension. The first IAT was defined with an occupational target dimension (*Athletes/Politicians*) and the second with a racial target dimension (*Black/White*). The IAT that was framed in occupational terms showed a preference for

Athletes (who were black) over *Politicians* (who were white) whereas the IAT framed in terms of race showed a preference for *Whites* (who were politicians) over *Blacks* (who were athletes) even though the exemplars were identical for both tests. Statistically significant different between-test IAT effects were obtained, indicating a marked difference in exemplar evaluation between the two IATs. Mitchell et al. (2003) described their results as follows:

> "Experiment 1 demonstrated that social objects evoked different automatic attitudes as a function of the context in which they were encountered. When highly regarded Black athletes such as Michael Jordan were categorized by occupation, positive automatic attitudes were elicited, in line with consciously reported attitudes of liking. However, when the exemplars were categorized by race, the elicited attitude was qualitatively different from the one observed under occupation categorization."

In essence this experiment demonstrated that a semantic modification of the target categories (in this case from an occupational to a racial domain), can lead to statistically different IAT results even when all other aspects of the IAT are kept constant.

In summary, the research into the effects of the influence of category manipulation on IAT results, although sparse, reveals that IAT scores can be affected by the manipulation of either the target or the attribute dimensions even when all other features of the IAT are kept constant.

Positioning this Research

The research reviewed demonstrates that the manipulation of the exemplars or categories of either the target or the attribute dimension can be sufficient to significantly affect IAT scores. In this regard, Nosek et al. state:

> "In sum, IAT design requires careful attention to the selection of both category label and stimulus items. Category labels are clearly of great importance for the IAT, but the stimulus exemplars can nevertheless influence the construal of those categories." (2007, p. 282)

At time of writing, there is general agreement that the IAT is primarily influenced by its category definitions (De Houwer, in press; Lane et al., 2007; Olson & Fazio, 2004). By contrast the IAT appears to be fairly robust when it comes to the manipulation of exemplars. De Houwer (2001) showed that when half of the target exemplars in a category are given a positive valence and half a negative valence, there was no appreciable difference in the sorting time of the exemplars. It appears that it is only when a sufficient proportion of exemplars are biased in a particular direction that a cognitive re-definition of a category may occur (Govan & Williams, 2004) with a resultant effect on IAT scores.

While there is evidence that the IAT is relatively robust with regards to the selection of its exemplars, less is know about the sensitivity of the IAT regarding the selection of its categories. As mentioned above, Olson and Fazio (2004), showed that the IAT was able to detect differences between what they termed the traditional and the personalised IAT, in which the attribute dimension was modified from *Pleasant/Unpleasant* to *I Like/I Don't Like*. This may be considered a moderate semantic variation in attribute category definition because, as Olson and Fazio (2004) argue, while the meanings of the categories are similar, *Pleasant* and *Unpleasant* carry a specifically normative implication whereas *I Like* and *I Don't Like* are explicitly personal. In a similar vein, Mitchell et al. (2003) showed that the IAT was capable of detecting differences between major semantic modifications in target dimensions from an occupational (*Athlete/Politician*) to a racial (*Black/White*) target categorisation.

Taking cognisance of the primacy of category influence on IAT scores, and the scarcity of the research into the influence of category selection, the study at hand extends Mitchell et al.'s (2003) work, concentrating on further exploring variations in the IAT *target dimension* while keeping the attribute dimension and all exemplars constant. This study takes two main tacks. The first is to endeavour to determine whether minor changes in categories of the target dimension might yield significantly different IAT results across tests. To this end two experiments will be conducted, one in a racial/cultural domain with target dimensions *Black/White*, *African/Western* and *Previously Disadvantaged/Previously Advantaged* and the other in a gender domain with target dimensions *Female/Male*, *Girl/Boy* and *She/He*. This is a new avenue of investigation in that it is *minor* semantic changes in the target dimension that are under investigation. The second is to attempt to replicate the findings of Mitchell et al.'s (2003) first experiment to validate their hypothesis, but using race and gender as target dimensions rather than race and occupation. To this end, this study will conduct a third experiment with categories *Black/White* as the race target dimension for the first IAT and *Female/Male* as the gender target dimension for the second IAT. In addition, this study introduces the notion of a composite target-dimension in which category labels define more than one feature. It investigates how an IAT with a composite target dimension (e.g. *Black*

Male/White Female) might compare to IATs with the usual singular target dimensions (*Female/Male* or *Black/White*).

Is the IAT capable of detecting subtle semantic changes to its target categories, or is it only able to detect more major semantic changes? This is the main question addressed by this study.

Chapter 4: Aim and Rationale

Typically, before a psychometric test is accepted by the psychological community it must be rigorously evaluated to determine whether or not it can be considered to be a good psychological measure of the construct under investigation. The sample population for which it is relevant is stipulated and the test's applicability and limitations are outlined. There is a clear methodology laid out for selecting test items and for determining the reliability and validity of a new test.

Not all psychological tests fit neatly into a prescribed methodology for determining whether or not, or under what circumstances they are good psychological measures of particular constructs. For such tests psychologists are required to be creative in exploring the test's boundaries through the testing of hypotheses that help to highlight the applicability and limitations of the test for the constructs under consideration. The Implicit Association Test (IAT) is just such a test. It cannot be neatly pigeon-holed or evaluated as is the case with many other psychological tests. There is still much debate surrounding how the IAT works in terms of the cognitive processing that it activates and what it is that the test actually measures, whether social perceptions or individual mental states or some mixture of the two (Arkes & Tetlock, 2004; Blanton & Jaccard, 2006; Karpinski & Hilton, 2001; Olson & Fazio, 2004; Tetlock & Arkes, 2004) . In 2004, an informal document entitled the "Revised Top Ten List of Things Wrong with the IAT" was drafted (Greenwald, 2004). This document focused on areas of theoretical concern and issues raised regarding the IAT and its capability of measuring implicit attitudes. It also pointed to areas of progress in IAT research and theory, but candidly recognised that much still needs to be investigated in addressing outstanding questions.

What is beyond doubt is that the IAT has captured the attention of the psychological research community who vie to either build up or tear down the credibility of the claim that the IAT is capable of measuring implicit attitudes or preferences. This sometimes acrimonious difference of opinion within the research community has led to the publication of a considerable body of research that has served to accelerate the evaluation of the IAT as an instrument for measuring implicit cognition. The result is that the IAT's applicability and limitations are gradually clarifying. A growing body of empirical findings is allowing researchers to propose theoretical models for how the IAT interacts with implicit cognitions (De Houwer, 2001, 2006, in press; Fazio & Olson, 2003; Greenwald et al., 2002; Greenwald, Nosek, Banaji, & Klauer, 2005; Olson & Fazio, 2003; Rothermund & Wentura, 2004; Rothermund, Wentura, & De Houwer, 2005).

One of the areas of considerable investigation into the IAT has been into the effect of the manipulation of IAT categories and exemplars on IAT results. The aim of this study is to make a small contribution to this particular niche of IAT research, in particular, to investigate the effects of manipulating the categories of the IAT target dimensions on IAT scores with a view to determining the sensitivity of the IAT to semantic variations of the target categorisation frame. It is hoped that these findings might assist in delimiting the applicability and limitations of the IAT as a psychological instrument for investigating implicit cognitions.

Chapter 5: Methodology

This chapter presents the three IAT experiments conducted for this study and the methods and procedures used in obtaining and analysing the experimental IAT data. The IAT experiments shared a common methodology, differing primarily in the specifics of the labels used to describe the target IAT categories, the number of participants in each experiment and the particulars of the sample composition (although all participants were drawn from the same predominately South African student population). The nature of the three experiments will first be described, then, in the interests of brevity, the features common to the method of all three experiments will be explained. Finally the unique particulars of each experiment's sample will be presented.

Experiments in Target Category Manipulation

Three within-subjects experiments were designed to test the effects of the manipulating the categories of the IAT *target* dimensions while keeping exemplars and the attribute dimension constant. Each experiment consisted of a battery of three counterbalanced IATs with each IAT differing from the others only in the labels of its target categories.

Minor Semantic Changes in the Target Dimension

The first two experiments were concerned with the same goal, to determine whether or not *minor* semantic modifications in the IAT's target dimension would be discernable by the IAT. The primary difference between the two experiments was in their attitudinal domains, the first being a racial/cultural domain and the second a gender domain. The decision to conduct two experiments was prompted by the fact that this is a new area of IAT investigation. Concordance in experimental results from two different attitudinal domains would carry a greater weight than if only a single experiment were conducted. The details of experiments 1 and 2 are given below:

Experiment 1 – Minor Semantic Changes in a Race/Culture Domain

This experiment involved a counterbalanced within-subject design with each participant in the sample being presented with a battery of three IATs having a target dimension in the domain of race/culture. For all three IATs, the attribute categories (*Good/Bad*) and all exemplars (for both the target and the attribute dimensions) were kept constant. Only the target categories were changed between tests. Target categories were as follows:

- IAT 1: *Black* and *White.*
- IAT 2: *African* and *Western.*
- IAT 3: *Previously Disadvantaged* and *Previously Advantaged.*

Semantically, the first of these is a racial target dimension, the second carries connotations of culture and race and the third, in South African society, suggests a racial domain similar to that represented by the categories *Black* and *White* except that this categorisation is more ambiguous in that it encompasses groups such as Indians and Coloureds who, in apartheid South Africa, were historically advantaged compared to blacks and historically disadvantaged compared to whites.

These target dimensions were selected to have minor semantic differences from one another for a predominately South African sample. Since this is new ground as far as the IAT is concerned it was difficult to hypothesise as to whether or not the IAT would discriminate between the three tests with a degree of statistical significance. The literature points to the fact that under most circumstances the IAT categories are primarily responsible for influencing IAT scores (De Houwer, 2001, in press). This implies that for this experiment a variation in the IAT scores should be expected, although these may not reach statistical significance because of the fact that the target categories are semantically similar to one another. On the other hand, the nature of the exemplars can sometimes result in a re-definition of the categories (Govan & Williams, 2004). If this effect were to predominate, it is possible that the target categories could be cognitively redefined to a common categorisation that might be applied over all three tests. For example, all three IATs may be cognitively redefined as referring to race. If this were the case it is likely that results of all three IATs in the battery would converge as they would in effect be measuring the same construct. Given the above, it was hypothesised that the likelihood was that the IAT would be unable to discriminate between the three tests, at least not with statistical significance, although a reasonable mount of variation in IAT means was anticipated.

In addition a secondary hypothesis that within the white sub-sample the *White/Black* IAT would result in a preference for *White* over *Black* that exceeded the preference of *Western* over *African* or of *Previously Advantaged* over *Previously Disadvantaged*, although a statistically significant difference was not anticipated. This prediction was based on the assumption that the category *Black* could be conflated with negative connotations that are sometimes attached to the word 'Black' in terms like black (evil) magic, black market, black plague and blackmail and in associations between 'Black' and darkness or the night and the fears that such phenomena invoke (Smith-Mclallen, Johnson, Dovidio, & Pearson, 2006). The word 'White' by way of contrast is often associated with terminology such as white (good) magic and positive phenomena like purity and light. It was hypothesised that these alternative conceptualisations of *White* and *Black* could facilitate the sorting of *Good* exemplars to the *White+Good* associative pair and *Bad* exemplars to the *Black+Bad* associative pair while disrupting the sorting of *Bad* exemplars to the *White+Bad* associative pair and *Good* exemplars to the *Black+Good* associative pair. This facilitation in the first instance and disruption in the second would increase the magnitude of the IAT effect for the *Black/White* IAT relative to the other two IATs in the experiment.

A confirmation of the primary hypothesis would suggest that subtle semantic differences in the target dimension either have little effect on IAT results or are difficult to detect in a within-subject design, possibly because of a cognitive confound in which a temporary redefinition of categories influences exemplar evaluation.

The confirmation of the secondary hypothesis that a greater magnitude in the IAT effect was expected for the *White/Black* IAT as compared to the two other IAT's, while not conclusive, would support recent findings that evaluative associations with colours 'White' and 'Black' are related to evaluative racial associations (Smith-Mclallen et al., 2006).[3]

Experiment 2 – Minor Semantic Changes in a Gender Domain

This experiment was similar to experiment 1, except that it was less susceptible to ambiguity and was located in the gender domain. It involved a counter-balanced within-subject design with each participant in the sample being presented with three IATs having a target dimension in the gender domain. For all three tests, the attribute dimension and all exemplars were kept constant. Only the target categories were changed between tests. Target categories were as follows:

[3] Smith-Mclallen et al. found that controlling for the influence of implicit colour preferences nevertheless did not alter implicit racial preferences. As they noted: "although Whites' implicit preferences for the colour white over the colour black were consistently correlated with their racial preferences, implicit racial preferences remained significant beyond any effect of colour preferences." (Smith-Mclallen et al., 2006, pp. 65-66)

- IAT 1: *Female* and *Male*
- IAT 2: *Girl* and *Boy*
- IAT 3: *She* and *He*

These target dimensions were selected to have minor semantic differences from one another. Here, the *Girl/Boy* categorisation might be considered the most semantically different within the battery as is specifically refers to children whereas the other IAT target dimensions are neutral with respect to age.

For similar reasons to those identified for experiment 1, it was hypothesised that the IAT would be unable to discriminate between the three tests. A secondary hypothesis was that within the female sub-sample the commonly encountered pro-female findings amongst female participants (Rudman & Goodwin, 2004) would be diminished in the *Girl/Boy* IAT as compared to the *Female/Male* and the *She/He* IATs, although a statistically significant difference was not anticipated. This hypothesis was based on an expectation that gender evaluations in the case of the *Girl/Boy* IAT might be moderated by the explicit association of the target dimension to children. In accordance with Rudman and Goodwin's (2004) findings that women who perceived men as intimidating or threatening tended to have greater implicit pro-female preferences than their non-threatened counterparts, it was supposed that the term *Boy* as compared to the word *Male* or *He* might provoke this sense of threat to a lesser degree, potentially resulting in a lesser pro-female result in the Girl/Boy IAT in the female sub-sample.

Major Semantic Changes in the Target Dimension

The third experiment was similar to that of Mitchell et al's (2003, Experiment 1) in that major semantic modifications were made to the target dimension of the tests in the IAT battery. The purpose of this experiment was to replicate Mitchell et al's findings but in an IAT battery having race and gender domains. This would have the dual purpose of confirming Mitchell et al's findings (2003, Experiment 1) and providing a contrast to the aforementioned experiments that investigated minor semantic modifications to the target dimension.

Experiment 3 – Major Semantic Changes across Domains

This experiment entailed a counterbalanced within-subject design with each participant in the sample being presented with three IATs having target dimensions that were substantially semantically different for each test. For all three tests, attribute categories (*Good/Bad*) and all exemplars (for both the target and the attribute dimensions) were kept constant. Only the target categories were changed between tests. Target categories were as follows:

- IAT 1: *Black* and *White*
- IAT 2: *Female* and *Male*
- IAT 3: *Black/Male* and *White/Female*

Semantically, the first of these is in the domain of race, the second in the gender domain and the third in a composite race-gender domain. These target dimensions were selected to have substantial semantic differences from one another. It was hypothesised that, as found by Mitchell et al. (2003), the IAT would discriminate between these three tests. A confirmation of this hypothesis would corroborate prior findings that major semantic differences in the target dimension alone are sufficient to significantly affect IAT scores. A composite target domain was included to allow a peripheral investigation into how a composite domain might affect IAT results relative to the single domain IATs from which it was derived. The hope was that a pattern between the single and composite IATs might be discernable (although naturally not conclusive).

Exemplar Selection

A word is required on exemplar selection for the three experiments. With regards to the *Good/Bad* attribute dimension all three experiments used the same exemplars, drawn from those normed by Greenwald et al (1998), with ten exemplars selected for each category. For the target dimension, the exemplars used in all three experiments were typical South African names. In the first two experiments an equal number of names were chosen from each of the four groups: black males, black females, white males, and white females. In fact, the same exemplars were used for both experiments, but exemplar category membership was determined by race in experiment 1 and by gender in experiment 2. In the third experiment exemplars were constrained by the requirements of the composite categories of the third IAT in the battery. That is, only exemplars belonging to the groups black females and white males were permissible. In experiment 3, the exemplars for the target categories were repeated twice meaning that there were only five unique exemplars per target category for this experiment. According to Nosek et al. (2005) limiting the number of unique exemplars per category in this way would be expected to have a negligible effect on IAT results. In fact, the study upon which experiment 3 was based (Mitchell et al., 2003) only used three unique exemplars per target category repeating them a number of times within the requisite IAT blocks. The exemplars for each of the IAT experiments are listed in Appendix A.

Experimental Method and Procedure

This study's three experiments and their related hypotheses have been described in some detail above. A description of the methodological features of these experiments and the assignment algorithm that dynamically allocated each participant to an experiment now follows:

Sample

All participants in the study were registered psychology students at South African university at the second year level or higher. Convenience sampling was used to recruit subjects, with students being invited to voluntarily participate in the research as an optional practical component in their psychology course. Participants did not receive any remuneration. All participants were fluent in English, but not all were first language English speakers. All participants had not previously encountered the IAT.

Sample Details for Experiment 1

Sixty-three subjects participated in the first experiment. The sample comprised of 13 black females, 9 black males, 12 Indian females, 5 Indian male, 15 white females and 9 white males. The average age of the participants was 21. Fifty-nine of the sixty-three participants (94%) were South Africans. All but one was from the African continent. A few participants who did not identify themselves as black, Indian or white participated in the study but their data was excluded from the analysis because of the scarcity of their numbers.

Sample Details for Experiment 2

Fifty-nine subjects participated in the second experiment. The sample comprised of 14 black females, 5 black males, 9 Indian females, 6 Indian males, 17 white females and 8 white males. The average age of the participants was 20. Fifty-three of the fifty-nine participants (89%) were South Africans. All but one was from the African continent. A few participants who did not identify themselves as black, Indian or white participated in the study but their data was excluded from the analysis because of the scarcity of their numbers.

Sample Details for Experiment 3

Forty-seven subjects participated in the third experiment. The sample comprised 13 black females, 5 black males, 10 Indian females, no Indian males, 15 white females and 4 white males. The average age of the participants was 21. Forty-one of the forty-seven participants (87%) were South African. All but one was from the African continent. No participants from other ethnic groups participated in the study. A disappointing characteristic of this sample was the relative scarcity of males. Only 19% of the sample was male.

Measurement Instrument

The measurement instrument used in the three experiments was the Implicit Association Test. Participants accessed the computer based tests over the internet at the website http://www.webiat.com (no longer operational). The website and software for delivering the IAT over the internet were developed by the author of this research for the purpose of conducting these experiments. The web-based IAT was developed using the Microsoft programming technologies, VB.NET and ASP.NET with results being recorded to a Microsoft Access database on the web-server. Unlike the Project Implicit website, which has been used for conducting internet based IAT experiments by the research community (Nosek, Banaji, & Greenwald, 2002), this website delivered IATs to participants using Javascript web browser technology without the need for the installation of web browser plug-ins.

The web-server was programmed to take participants through a process of informed consent, gather their user demographics, administer a web-based questionnaire on explicit racial and gender attitudes, assign subjects to one of the three experiments and administer an appropriately counterbalanced battery of three IATs per experiment.

For each experiment, the IATs were administered in accordance within the guidelines of standard IAT conventions. Category labels, in concordance with left and right assignment keys, were displayed in the mid-left and mid-right of the screen respectively. Exemplar stimulus words were displayed towards the bottom of the screen in the centre. The computer recorded response times from the initial appearance of the exemplar stimulus to the point at which a response key was pressed. An inter-trial delay of 100 ms was used between user response and exemplar display. If an incorrect response key was pressed, a red 'X' appeared above the exemplar to indicate an error and the user was still required to press the correct key to complete the trial. Error responses were flagged and recorded to file.

It should be noted that timing precision for measuring response latencies is dependent upon a number of factors, in particular the user's PC hardware, operating system and choice of web browser. A consistent timing precision could therefore not be guaranteed across users. However, as Greenwald et al (2003) point out, this is not a serious drawback because of the non-systemic nature of the resulting noise and the averaging of response latencies over multiple trials. For this study, it is likely that timer precision varied from approximately 10 ms to 20 ms, which would have a negligible effect on results obtained.

To test the measurement instrument, an informal pilot study of the web-based IAT was conducted prior to enlisting research subjects. During this process a number of minor problems related to internet access were detected and addressed. No apparent problems with the instrument were uncovered and predictable IAT effects were obtained in this pilot phase.

Apparatus

Participants made use of PCs with internet access to complete the IATs. Most subjects made use of computers located in the local university campus laboratories. These computers were relatively up to date using the Windows XP operating system. A few participants made use of their personal home computers. Statistics gleaned from the web-site reveal that a large majority of subjects used the Internet Explorer 6.0 browser to connect to the IAT website with only 3 percent using some other browser. All browser versions of Internet Explorer were version 5.0 or above. It seems reasonable to assume that the hardware/software configurations used by participants could be considered equivalent for the purposes of this research.

Research Design

The research involved three experiments having a within-subject design, each requiring the administration of a battery of three counterbalanced IATs to a different sample. Since the attitudinal domains of interest in these experiments were those of race/culture and gender it was desirable to examine not only the results obtained over the entire experimental sample, but also within the gender and race groups to which the participants belonged. The race groups analysed were limited to black, Indian and white. Participants of other races were excluded from the analysis owing to the scarcity of their numbers.

For each of the experiments, the order in which the three tests in the battery were presented to participants was counterbalanced to factor out effects of test familiarity, fatigue and most importantly the documented phenomenon that IAT effect magnitudes tend to decline with repeated administration both in longitudinal studies and in multiple testing sessions such as those used in this study (Greenwald & Nosek, 2001; Greenwald et al., 2003). There were thus six different combinations in which the three tests could be presented. With regards to counterbalancing considerations, there was also a procedural effect to consider. Greenwald et al. (1998) showed that the procedural order of administration of the scored IAT blocks *within each test* has a small impact on IAT scores. The performance of the first scored block appears to interfere somewhat with the performance of the second (Nosek et al., 2007). Thus, for example, the presentation of the scored block *White+Good* and *Black+Bad* before the scored block *White+Bad* and *Black+Good* usually yields an IAT effect that is more pro-white (or less pro-black) than when the presentation order of these blocks is reversed. As a result, researchers that are interested in measuring the strength of group attitudes typically counterbalance the within-test order of the scored blocks to counteract this procedural effect. In a between-test study such as this one, it is important that the IATs in a battery all have the *same* within-test ordering to eliminate variations in score that are a consequence of this procedural effect. That is, for a given participant the order of the scored blocks must be the same for each IAT in the battery. This requirement appears to suggest that counterbalancing of the scored blocks should be omitted from this study. However, a supplementary interest of this research was to examine the strength of group attitudes validating them against trends in the literature, a goal that made the implementation of scored block counterbalancing desirable. Happily both requirements could be accommodated by carrying out the counterbalancing of the scored blocks across participants. Thus a participant had *all three* IATs in their test battery assigned to either the one or the other scored block condition. This arrangement ensured the integrity of the data for between-test comparisons while satisfying the procedural requirements for the calculation of group attitudes. What this meant for the experimental design in practice was that a secondary 2-way counterbalancing was introduced for each battery. (It should be noted that this addition was not necessary for the primary purpose of this study and was therefore subordinate to the counterbalancing of test order within the IAT battery). Together with the six combinations of test order, this additional counterbalancing resulted in 12 different combinations in which the three tests could be presented in the IAT battery.

A further complication arose from the fact that it was of interest to examine the experimental results by participant race and gender. It was therefore desirable to apply this 12-way counterbalancing within each of the combinations of race and gender of the sample. Since there were two gender groups (male and female) and three race groups (black, white and Indian) to be analysed, this represented six distinct groups of interest. In all, to counterbalance the twelve different combinations of test presentation over each of these six groups would yield the requirement that 72 appropriately sequenced batteries be administered to obtain a perfectly balanced design.

To summarise:

- A battery of 3 tests can be counterbalanced in 6 combinations.
- Each battery can further be counterbalanced in terms of the order of presentation of the 2 scored IAT tasks (combined blocks) within the test.
- There are thus 12 sequences in which the tests can be presented to achieve a balanced design.
- There are 2 genders and 3 race groups of interest within the sample, yielding 6 distinct groups of interest.

- A perfectly balanced experiment (at the level of the group) would require each of the 6 distinct groups to have the 12 unique sequences presented to them, i.e. the administration of a total of 72 appropriately sequenced batteries.
- The above pertains to each experiment. This research encompasses three separate experiments with participants gleaned from the same pooled sample.

Balancing the data as described above, although desirable is constrained by a number of factors. Principal among these are

- The ability to implement an effective algorithm to allocate participants to the appropriate experiment and battery.
- Access to a sample that is sufficiently large to ensure that each distinct group defined by race and gender is adequately represented in the experimental data.

A basic description of the algorithm that was devised to implement the research design by allocating participants to an experiment and battery can be summarised as below:

- Allocations to an experiment and battery were made in the order that participants took the test on the website.
- Allocations were made on the basis of participant race and gender, first to experiment 1, then to experiment 2 and then to experiment 3. For each experiment, a complement of six batteries counterbalanced by test order (and having the same secondary counterbalancing condition) was allocated per participant race and gender. Thereafter subsequent participants were allocated to the next experiment.

- Once allocations to experiment 3 were complete, the algorithm was reset to allocate to experiment 1 (with a switch made to the secondary counterbalancing condition).

Post hoc it became evident that this algorithm and its implementation suffered from a number of deficiencies. In particular a finer grained allocation scheme would have produced a more optimal distribution of allocations. Further details will be given in a later chapter when the research results are presented. It will be shown that despite certain shortfalls, the algorithm performed sufficiently well to assure design integrity.

Procedure

Prior to participating in this research, the target student population in a social psychology course were given a formal lecture in which they were educated on the differences between explicit and implicit measures of cognition in general and introduced to the Implicit Association Test in particular. It was not expected that this introduction would influence IAT scores as the IAT has been shown to be resistant to attempts at faking amongst first time test-takers in the absence of explicit instructions as to how IAT scores can be controlled (Nosek et al., 2007)[4]. These students were then invited to take part in a research study in which they would complete a battery of three IATs online and comment on their experience and opinion of their test results in order to obtain credit for their social psychology course. (If they preferred they could complete an alternative exercise which excluded the online use of the IAT.) Confidentiality was guaranteed and participants were informed that while certain

[4] It has been shown, for example that many white IAT participants show an implicit preference for *White* over *Black* despite an explicit desire not to do so and many black participants show an implicit preference for *White* over *Black* despite explicitly desiring not to do so (Nosek et al., 2002).

demographic information such as race and gender would be requested from them when they participated in the research online, no names or other such identifying data would be required of them. Participants were asked to indicate their informed consent on-line before proceeding with the test and were advised both verbally and online that they could opt out at any time during the test, an action that would result in the deletion of all of their associated test data. Finally, subjects were informed of the possibility that they might find their IAT results uncomfortable or disturbing as they could potentially suggest that their attitudes towards race or gender might differ from their idealised view of these attitudes. Helpful literature in this regard was made available to all participants and they were offered the opportunity to discuss their test results should they desire to do so.

Participants who elected to take part in the research were given the address of the research website and could visit the website at any time within a two week period to participate in the research. Following an online process of informed consent, participants provided demographic information and were dynamically allocated to one of the three experiments. This allocation was in accordance with a programmed algorithm on the web-server that attempted to distribute participants equitably (.i.e. in a balanced manner) to an experiment and to a suitably counterbalanced test battery depending on the participant's race and gender.

The allocated battery of three IATs was administered with a participant-controlled pause between each test and each test block. The tests followed the traditional IAT sequence of practise and scored blocks with practise blocks requiring the sorting of exemplars to single categories and scored blocks requiring the sorting of exemplars to associative category-pairs. A left response was effected by pressing the 'Q' key and a right response by pressing the 'P' key.

Data Preparation

For all of the experiments, the IAT effect was calculated for all three tests in the IAT battery in accordance with the specifications of the algorithm laid out by Greenwald et al. (2003). To briefly summarise, this involved:

- The elimination of trial response latencies greater than 10 000 ms or less than 400 ms.
- The exclusion of any participant whose response times were less than 300 ms for more than 10% of trials.
- The inclusion of trials that were initially error responses, with the latencies for such trials adjusted in accordance with Greenwald et al.'s (2003) recommendations to replace error latencies with the mean latency of the block.
- The calculation of the IAT effect as the difference between the latencies of the scored blocks divided by the standard deviation of the latencies of both scored tasks.

Participants who wished to do so could repeat the experiment. In such cases the data were excluded from the analysis. These cases were detectable if the participant repeated the experiment during the same session on the website (that is this could be determined programmatically) or if they indicated on-line that they were repeating the experiment.

IAT results were calculated on the web-server once a participant concluded the experiment and were saved to a database on the web-server. The programmatic calculation of IAT scores was verified by a comparison with results obtained using the SPSS statistical package executing the algorithm available on Greenwald's website (Greenwald, 2005).

Data Analysis

The data from each of the experiments were analysed using a one-way ANOVA. This was done in order to determine whether or not there were statistically significant differences between the IAT effects obtained across the three-IAT battery. A repeated measures ANOVA was not used because the three tests in the IAT battery were not equivalent, differing as they were in their target dimension. For each experiment, the ANOVA was conducted over the entire sample and within sub-groups of the sample that were derived from the race and/or gender demographics of the participants. This sub-group analysis was conducted in order to investigate whether the various experimental hypotheses were consistent in more homogenous sub-samples.

Chapter 6: Results

This chapter presents the experimental results. Before doing so, a brief recap on how this study differs from more traditional IAT research is given in order to re-iterate the primary focus of this research. Thereafter, issues of research design integrity arising out of the sampling procedure are examined, followed by a comment on the integrity of the statistical testing used in this research. Next, an overview of the format that is used to present the results is described. Finally the results are presented.

Applied vs. Meta-research

The IAT is most commonly used in applied research to investigate group attitudes towards social objects by using statistical tests to compare IAT group means to a no preference zero base-line or to compare group means using statistical tests such as t-tests and ANOVA. This study, by contrast, is characterised as meta-research in that its main concern is *between-test* comparisons of IATs administered in an IAT battery in order to gain insight into the capabilities and limitations of the IAT itself.

Design Integrity

In the previous chapter the rationale for the research design was presented, the counterbalancing requirements to realise the design were discussed and the allocation algorithm to implement it was described. Since it is known that IAT effects are influenced by the order in which multiple tests are presented in a battery study (Greenwald & Nosek, 2001; Greenwald et al., 2003), it is important that the allocation algorithm that executed on the web-server resulted in an equitable distribution of the order in which the tests were presented in the experiment, without which the integrity of the experimental data would be questionable.

An analysis of the allocation results revealed that the algorithm and its implementation suffered from a number of deficiencies. In particular, four deficiencies were apparent. These were:

- The requirement that a full complement of six batteries be allocated per participant race and gender before making allocations to the next experiment resulted in an uneven distribution of completed batteries over the three experiments (n=63, 59 and 47 respectively).
- The algorithm sought to balance the order of test presentation at the level of participant race and gender under the assumption that this would result in a reasonable balancing over the entire sample. (As it happens this was the case, but a finer grained allocation scheme would have produced more optimal results).
- When a participant opted to abort a test battery the algorithm did not recognise the need to re-introduce the aborted battery for allocation. (An infrequent condition).
- When a participant opted to repeat a test battery (which was then omitted from the analysis) the algorithm allocated the participant to a battery as if he/she were a new participant and did not recognise the need to retain that battery for allocation.

Despite these flaws and the resulting imperfections in the distribution of participants to experiments and batteries, the algorithm performed reasonably effectively, especially with regards to the sub-samples that were defined by participant race and gender. Appendix B gives a breakdown of the allocation results for each experiment over the entire sample and the various sub-samples of interest. Two levels of allocation detail are provided in separate tables. Respectively these itemise the number of times each *battery sequence* was allocated and the resultant number of times each *test* was presented first, second and third. Importantly the latter breakdown reveals a relatively even distribution of test presentation order, meaning that in general the IATs were equally influenced by the ordering effects that are known to affect IAT scores in multi-test experiments (Greenwald & Nosek, 2001; Greenwald et al., 2003). The reader is referred to the introduction in Appendix B for a more comprehensive explanation of the allocation results.

The counterbalancing algorithm also attempted to carry out a secondary within-test counterbalancing of IAT block presentation order. This was important only for the supplementary investigation of this study that aimed to compare IAT results obtained by groups defined by participant race or gender to those published in the literature for such groups. On average the algorithm achieved a 3:2 distribution ratio for block presentation order which would have resulted in a slight (but most likely negligible) procedural bias towards a pro-white result in the race/culture IATs and a slight procedural bias towards a pro-male result in the gender IATs.

Statistical Integrity

The statistical test used to analyse the data for the various samples and sub-samples of this study was the one-way ANOVA. Since ANOVA testing assumes that sample data is normally distributed and that the variance in the data of the samples being compared is homogenous, these assumptions were tested to ensure the statistical integrity of the data analysis. An inspection of boxplots for all samples and sub-samples revealed that the sample data approximated a normal distribution in all cases, although for the smaller sub-samples ($n = 5$) this was a rather coarse approximation. Levene's test verified that the assumption of homogeneity of variance was satisfied for all samples and sub-samples. Thus, the basic ANOVA assumptions were met and the analysis could proceed.

Overview of Results Presentation

For each of the three experiments the following will be presented:

- The ANOVA results for the between-test comparisons of the IAT means over:
    - the entire sample
    - the homogenous sub-samples defined by participant race *or* gender
    - the homogenous sub-samples defined by participant race *and* gender

  All significance tests were conducted with $\alpha < 0.05$. Sample sizes refer to the number of participants to complete a battery of three tests. That is, the number of IATs completed was 3 times the sample size.
- A table, summarising the results of the statistical tests. This in effect presents the ANOVA results in a tabular format, but with the addition of means for each IAT in the battery for all samples and relevant sub-samples.
- Line-graphs plotting the means for each test in the battery over the entire sample and for the relevant sub-samples. While p-scores are the arbiter of whether or not there is a statistically significant difference between tests, the magnitude of the range between the IAT means, when represented graphically, gives a more tangible sense of the similarity or difference in test results. For the reader to better appreciate the line graphs, the rule of thumb used by IAT researchers to interpret the magnitude of the IAT effect can be used as a guide. Based on the rule of thumb, a qualitative difference in attitude is discernable where the difference in magnitude of the IAT effect exceeds 0.2 units. A difference in magnitude in IAT

scores of 0.1 units (the size of the y axis scale in all of the graphs) should be considered negligible.[5]

- Additional observations regarding the experimental results.
- A comparison of the IAT results for the race and gender sub-samples with trends published in the literature.
- Supplementary results that present *between-group* comparisons that may be of interest from an applied research perspective. These results demonstrate the ability of the IAT to discriminate between groups having relatively small sample sizes. As these results are of peripheral interest only they are presented in brief.

Results: Experiment 1

Experiment 1 targeted a race/culture domain to investigate whether the IAT would discriminate between tests having target dimensions defined by the categories *White/Black*, *Western/African* and *Previously Advantaged, Previously Disadvantaged* when the attribute dimension and all exemplars were kept constant.

Hypothesis Testing: ANOVA Results

An ANOVA showed that the between-test IAT results over the entire sample were not significantly different from one another (n=63, Mean=0.322, SD=0.373, F(2,186)=0.362, p=0.697).

Similar non-significant results were obtained for the sub-samples defined by participant race:

[5] See the paragraph entitled *Interpretation of the IAT Effect* in Chapter 2 for information on the rule of thumb used for interpreting the magnitude of the IAT effect. The author has taken some liberty in extending this rule to interpret *differences* in IAT effects. This is simply an aid for the reader to better appreciate the graphical representation of the experimental results.

- Black sub-sample (n=22, Mean=0.091, SD=0.378, F(2,63)=0.747, p=0.478)
- Indian sub-sample (n=17, Mean=0.389, SD=0.284, F(2,48)=0.221, p=0.802)
- White sub-sample (n= 24, Mean=0.486, SD=0.373, F(2,69)=0.080, p=0.923)

The pattern of non-significance continued for all sub-samples that were defined by a combination of race and gender:

- Black female sub-sample (n=13, Mean=0.071, SD=0.341, F(2,36)=0.259, p=0.773)
- Black male sub-sample (n=9, Mean=0.122, SD=0.430, F(2,24)=0.786, p=0.467)
- Indian female sub-sample (n=12, Mean=0.363, SD=0.235, F(2,33)=0.343, p=0.712)
- Indian male sub-sample (n=5, Mean=0.451, SD=0.381, F(2,12)=0.074, p=0.929)
- White female sub-sample (n=15, Mean=0.452, SD=0.349, F(2,42)=0.110, p=0.896)
- White male sub-sample (n=9, Mean=0.542, SD=0.255, F(2,24)=0.303, p=0.741)

These results confirmed the null hypothesis that minor semantic changes in the target dimension of the IATs would not result in statistically significant between-test differences.

Tabular Summary

The table below summarises the ANOVA results and gives the mean IAT scores for each test in the battery, together with the range of these means. (A more complete table that also gives standard deviations and confidence intervals for each IAT can be found in Appendix C). IATs 1, 2 and 3 in the table correspond to the *White/Black*, *Western/African* and the *Previously Advantaged, Previously Disadvantaged* IATs respectively.

Sample Race	Sample Gender	n	IAT 1 Mean	IAT 2 Mean	IAT 3 Mean	Mean Range	Battery Mean	F	p
All	Both	63	0.292	0.349	0.325	0.057	0.322	0.362	0.697
Black	Both	22	0.011	0.128	0.136	0.125	0.091	0.747	0.478
Indian	Both	17	0.404	0.412	0.351	0.061	0.389	0.221	0.802
White	Both	24	0.471	0.507	0.480	0.036	0.486	0.080	0.923
Black	Female	13	0.031	0.055	0.126	0.095	0.071	0.259	0.773
Black	Male	9	-0.019	0.233	0.150	0.252	0.122	0.786	0.467
Indian	Female	12	0.398	0.372	0.319	0.079	0.363	0.343	0.712
Indian	Male	5	0.417	0.508	0.429	0.091	0.451	0.074	0.929
White	Female	15	0.417	0.464	0.475	0.058	0.452	0.110	0.896
White	Male	9	0.560	0.579	0.488	0.091	0.542	0.303	0.741

Table 6-1 Summary of Race/Culture IAT results: IAT means, grand mean and ANOVA statistics

From the table it is evident that there is little variation in the IAT means with a range of less than 2.0 units for all sub-samples except for the black male sub-sample (range=0.252). According to the IAT rule of thumb, this suggests a negligible attitudinal difference between the tests in the battery. The one exception is the black male sub-sample for which the rule implies a small attitudinal difference between the *Black/White* (-0.019) and *Western/African* (0.233) IATs. However, a small sample size (n = 9) and a lack of between-test statistical significance for this sub-sample (p=0.467) raise questions about the accuracy of such an interpretation.

Graphical Representation

The IAT means for each of the tests in the experimental battery are plotted in the figures below. In figure 6-1, line-graphs are plotted for the entire sample (All) and for each sub-sample defined by the participant's race.

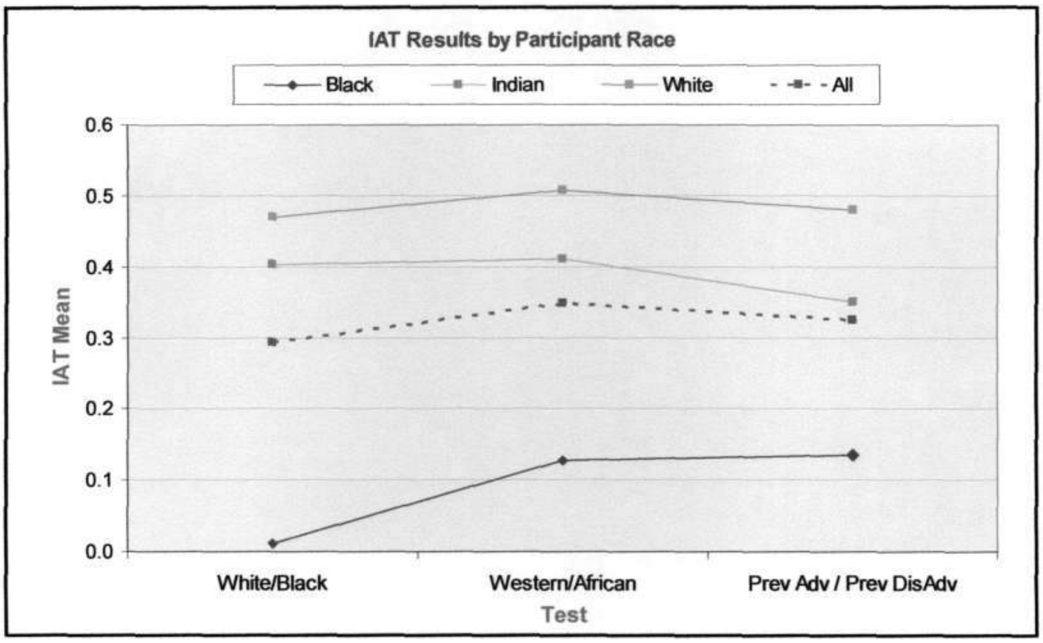

Figure 6-1 Race/Culture IATs: Results by participant race

In figure 6-2, line-graphs are plotted for each sub-sample defined by the participant's race and gender. The abbreviations BF, BM, IF, IM, WF and WM in the legend stand for black female, black male, Indian female, Indian male, white female and white male respectively.

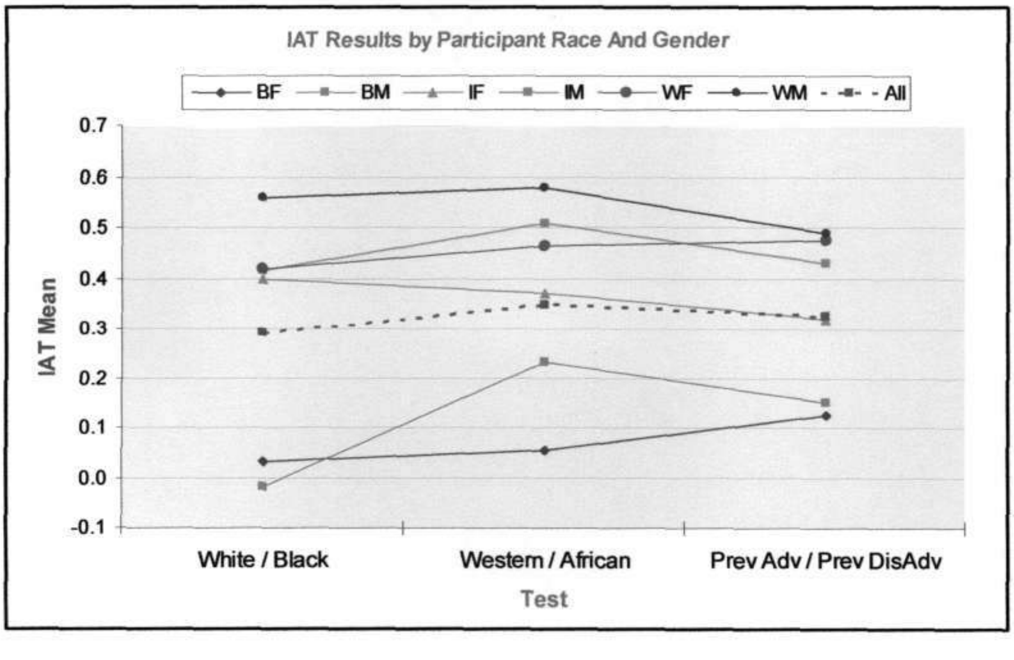

Figure 6-2 Race/Culture IATs: Results by participant race and gender

Remembering that a difference in IAT effect magnitude of 0.1 units is considered negligible, these graphical representations visually underscore that there is little variation in the mean IAT scores between tests of the IAT battery for all samples with the exception of the black male sub-sample. By contrast, it is evident from an inspection of the graphs that there is a considerably larger *between-group* variation in the IAT means. This difference is explored further in the section on supplementary results.

Additional Observations

Contrary to expectations, when ranking mean IAT scores, the *White/Black* IAT scored the lowest of the three tests for the white sub-sample (*White/Black* Mean=0.471, *Western/African* Mean= 0.480, *Previously Advantaged/Previously Disadvantaged* Mean=0.507). It was thought that in line with the research conducted by Smith-Mclallen et al. (2006) positive associations with the colour white and negative associations with the colour black amongst white participants would result in higher mean IAT scores being recorded for the *White/Black* IAT as compared to the IATs where colour associations were not present.

Trend Comparison

The mean IAT effects for the *White/Black* IAT for sub-samples defined by participant race were as follows: black (Mean=0.011), Indian (Mean=0.404), white (Mean=0.471). According to the IAT rule of thumb these scores show no racial preference amongst blacks and small to moderate pro-white preferences amongst whites and Indians. These results follow similar trends in the literature for white and black samples although some studies have reported higher pro-white attitudes amongst whites and slight pro-white preferences amongst black samples (Govan & Williams, 2004; Greenwald et al., 1998; Mitchell et al., 2003; Nosek et al., 2002). No study of Indian attitudes was available for comparison. It is possible that IAT results were of a lesser magnitude than would have resulted if participants had completed just a single IAT (Greenwald & Nosek, 2001; Greenwald et al., 2003).

Supplementary Results

Supplementary to the between-test results that have been presented above, a between-group ANOVA was carried out for each IAT with groups defined by participant race and by participant race and gender. Results are summarised in the table below. All tests showed statistically significant between-group differences except for the *Previously Advantaged/ Previously Disadvantaged* IAT when groups were defined by race and gender. These results are not integral to this study but may be of interest to applied researchers. They show the IAT's ability to discriminate between demographic groups for the sample sizes obtained in this study.

Test	Grouped By	Mean	SD	F	p
IAT 1: White/Black	Race	0.292	0.380	13.097	< 0.001
	Race and Gender	0.292	0.380	5.324	< 0.001
IAT 2: Western/African	Race	0.349	0.371	7.711	0.001
	Race and Gender	0.349	0.371	3.614	0.007
IAT 3: Prev Adv/Prev Disadv	Race	0.325	0.373	5.690	0.005
	Race and Gender	0.325	0.373	2.250	0.062

Table 6-2 Race/Culture IATs: Supplementary between-group ANOVA results

Results: Experiment 2

The second experiment targeted a gender domain to investigate whether the IAT would discriminate between tests having target dimensions defined by the categories *Female/Male*, *Girl/Boy* and *She/He* when the attribute dimension and all exemplars were kept constant.

Hypothesis Testing: ANOVA Results

An ANOVA showed that the between-test IAT results over the entire sample were not significantly different from one another (n=59, Mean= 0.264, SD=0.382, F(2,174)=0.279, p=0.757).

Similar non-significant results were obtained for the homogenous sub-samples:

- Female sub-sample (n=40, Mean=0.422, SD=0.310, F(2,117)=0.374, p=0.689)
- Male sub-sample (n=19, Mean=-0.068, SD=0.298, F(2,54)=0.340, p=0.713)

The pattern of non-significance continued for all sub-samples where a combination of participant race and gender defined the sub-samples:

- Black female sub-sample (n=14, Mean=0.450, SD=0.330, F(2,39)=0.433, p=0.651)
- Black male sub-sample (n=5, Mean=-0.143, SD=0.331, F(2,12)=1.471, p=0.268)
- Indian female sub-sample (n=9, Mean=0.412, SD=0.221, F(2,24)=1.178, p=0.325)

- Indian male sub-sample (n=6, Mean=-0.123, SD=0.296, F(2,15)=0.166, p=0.849)
- White female sub-sample (n=17, Mean=0.405, SD=0.337, F(2,48)=0.466, p=0.630)
- White male sub-sample (n=8, Mean=0.019, SD=0.265, F(2,21)=0.943, p=0.405)

These results confirmed the null hypothesis that minor semantic changes in the target dimension of IAT tests would not result in statistically different results across the tests.

Tabular Summary

Table 6-3 below summarises the ANOVA results over the entire sample and each of the sub-samples. In addition it gives the mean IAT scores for each of the IATs in the battery and the range of these means. (A more complete table with standard deviations and confidence intervals for each IAT can be found in Appendix C). IATs 1, 2 and 3 in the table correspond to the *Female/Male*, *Girl/Boy* and *She/He* IATs respectively.

Sample Race	Sample Gender	n	IAT 1 Mean	IAT 2 Mean	IAT 3 Mean	Mean Range	Battery Mean	F	p
All	Both	59	0.266	0.290	0.237	0.053	0.264	0.279	0.757
All	Female	40	0.440	0.439	0.387	0.053	0.422	0.374	0.689
All	Male	19	-0.102	-0.024	-0.079	0.078	-0.068	0.340	0.713
Black	Female	14	0.480	0.488	0.382	0.106	0.450	0.433	0.651
Black	Male	5	-0.340	-0.011	-0.077	0.329	-0.143	1.471	0.268
Indian	Female	9	0.324	0.478	0.434	0.154	0.412	1.178	0.325
Indian	Male	6	-0.182	-0.100	-0.086	0.096	-0.123	0.166	0.849
White	Female	17	0.469	0.377	0.367	0.102	0.405	0.466	0.630
White	Male	8	0.106	0.026	-0.076	0.182	0.019	0.943	0.405

Table 6-3 Summary of Gender IAT results: IAT means, grand mean and ANOVA statistics

The table shows that for most of the samples, the variations in IAT means (Mean Range) were small in magnitude, less than 2.0, showing a negligible attitudinal difference between the tests according to the IAT rule of thumb. This is with the exception of the black male sub-sample (range = 0.329) for which the rule implies a small attitudinal difference between the *Female/Male* (D= -0.340) and *Girl/Boy* (D= -0.011) IATs. However the small sample size (n = 5) and a lack of statistical significance for this sub-sample (p=0.268) raise questions concerning the accuracy of such an interpretation.

Graphical Representation

The IAT means for each of the tests in the experimental battery are plotted in the graphs below. In figure 6-3, line-graphs are plotted for the entire sample (Both) and for each sub-sample defined by the participant's gender.

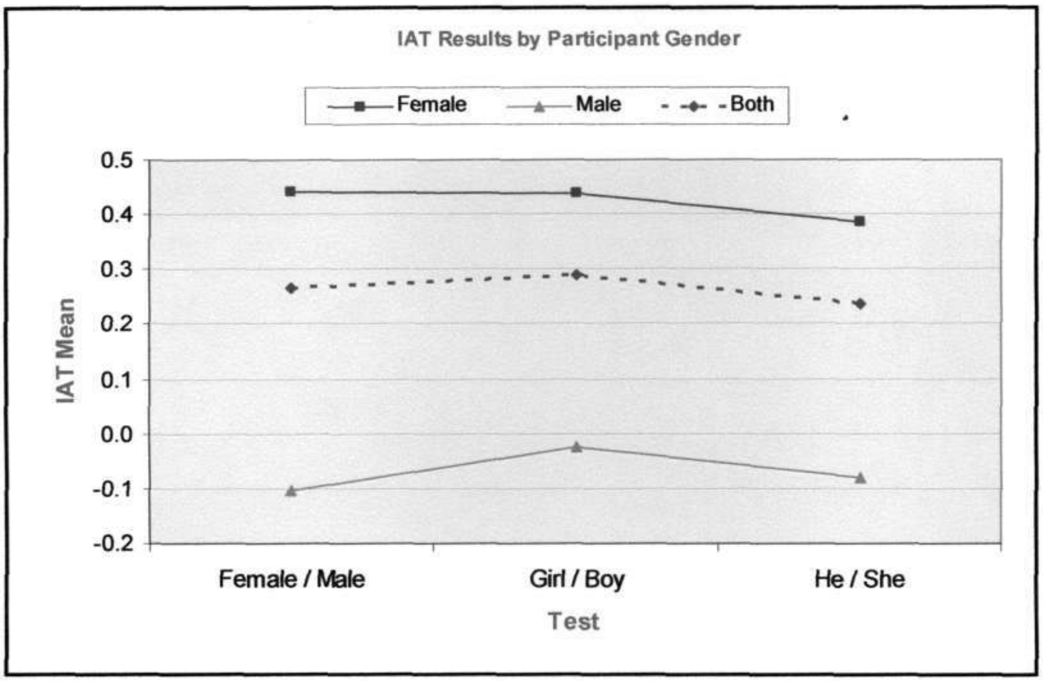

Figure 6-3 Gender IATs: Results by participant gender

In the figure 6-4, line-graphs are plotted for each sub-sample defined by participant race and gender.

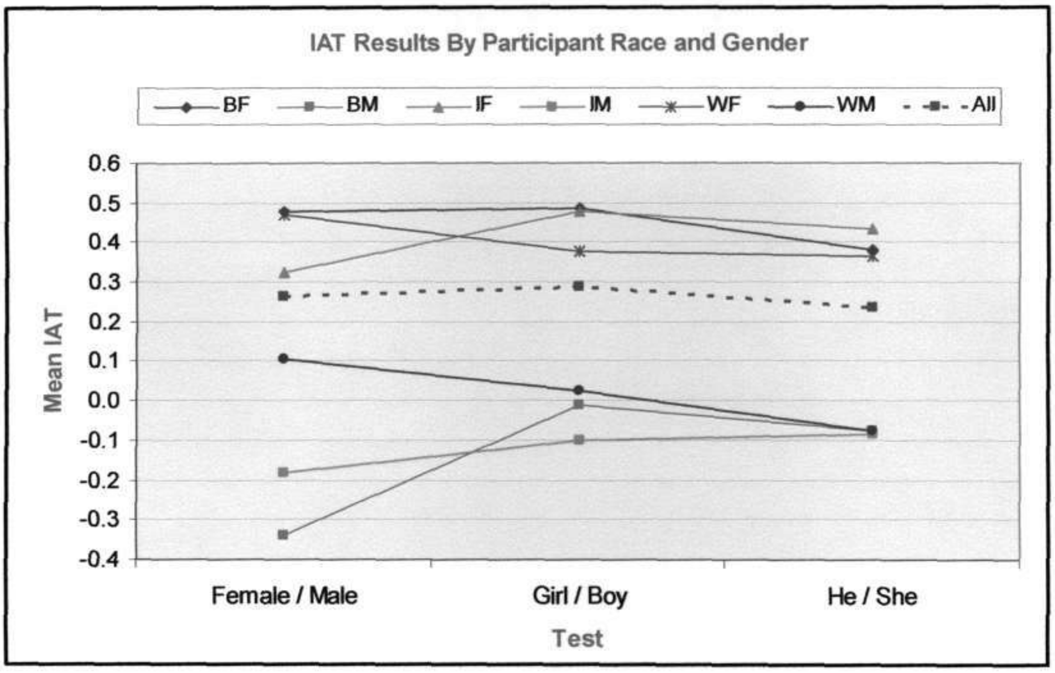

Figure 6-4 Gender IATs: Results by participant race and gender

As for experiment 1, these graphical representations visually confirm that there was little variation in the mean IAT scores between the tests of the IAT battery (except for the black male sub-sample) whereas there was considerable between *between-group* variation in the IAT means.

Additional Observations

Contrary to expectations, when ranking mean IAT scores for female participants, the *Girl/Boy* IAT did not show the smallest pro-female result (*She/He* Mean= 0.387, *Girl/Boy* Mean = 0.439, *Female/Male* Mean= 0.440). This expectation was based on the premise that the term *Boy* as compared to the word *Male* or *He* might provoke less of a sense of threat amongst women with negative experiences of men resulting in a reduced pro-female result (Rudman & Goodwin, 2004).

Trend Comparison

The mean IAT effects for the *Female/Male* IAT for sub-samples defined by participant race were as follows: male (Mean=-0.102), female (Mean=0.440). According to the IAT rule of thumb these scores show no gender preference amongst males and a moderate pro-female attitude amongst females. These results follow trends in the literature for male and female samples (Rudman & Goodwin, 2004).

Supplementary Results

For each IAT in the experiment, t-tests were carried out to compare IAT results between groups defined by participant gender. Similarly ANOVAs were conducted to compare IAT results grouped by participant race and gender. Results are summarised in the table below. All tests showed statistically significant between-group differences. These results are not integral to this study but may be of interest to applied researchers. They show the IAT's ability to discriminate between groups for the sample sizes obtained in this study.

Test	Grouped By	Mean	SD	t	F	p
IAT 1: Female/Male	Gender	0.266	0.418	5.827	33.953	< 0.001
	Race and Gender	0.266	0.418		8.841	< 0.001
IAT 2: Girl/Boy	Gender	0.290	0.376	5.367	28.802	< 0.001
	Race and Gender	0.290	0.376		5.861	< 0.001
IAT 3: She/He	Gender	0.237	0.354	5.982	35.782	< 0.001
	Race and Gender	0.237	0.354		6.761	< 0.001

Table 6-4 Gender IATs: Supplementary between-group ANOVA results

Results: Experiment 3

Experiment 3 investigated whether the IAT would discriminate between tests having target dimensions with major semantic differences defined by the categories *White/Black*, *Female/Male* and *White Male, Black Female* when the attribute dimension and all exemplars were kept constant. An unfortunate aspect of this experiment was that males comprised only 19% of the sample.

Hypothesis Testing: ANOVA Results

An ANOVA showed that the between-test IAT results over the entire sample were significantly different from one another (n=47, Mean=0.264, SD=0.417, F(2,138)=9.235, p<0.001).

Amongst Indian and white participants there were statistically significant between-test differences, but this was not true for the black sub-sample. Results were as follows:

- Black sub-sample (n=18, Mean=0.124, SD=0.405, F(2,51)=1.158, p=0.322)
- Indian sub-sample (n=10, Mean=0.327, SD=0.341, F(2,27)=3.780, p=0.036)
- White sub-sample (n=19, Mean=0.364, SD=0.433, F(2,54)=11.443, p<0.001)

A statistically significant between-test difference was found amongst female participants but not amongst their male counterparts.

- Female sub-sample (n=38, Mean=0.289, SD=0.396, F(2,111)=15.634, p<0.001)
- Male sub-sample (n=9, Mean=0.162, SD=0.490, F(2,24)=2.152, p=0.138)

Finally, with regards to sub-groups defined by participant race and gender, Indian and white females showed significant between-test differences whereas black females, black males and white males did not.

- Black female sub-sample (n=13, Mean=0.149, SD=0.408, F(2,36)=0.931, p=0.403)
- Black male sub-sample (n=5, Mean=0.060, SD=0.403, F(2,12)=2.04, p=0.173)
- Indian female sub-sample (n=10, Mean=0.327, SD=0.341, F(2,27)=3.780, p=0.036)

- White female sub-sample (n=15, Mean=0.384, SD=0.392, F(2,42)=31.389, p<0.001)
- White male sub-sample (n=4, Mean=0.289, SD=0.574, F(2,9)= 0.530, p=0.606)

These results confirmed the hypothesis that major differences in the IAT's target dimension could result in statistically different between-test results. This was the case over the entire sample and a majority of the larger sub-samples. The black male (p=0.173), white male (p=0.606), male (p=0.138), black female (p=0.403) and black (p=0.322) sub-samples obtained results that were contrary to the hypothesis. Of these, all of the male sub-sample had small sample sizes (n <10).

Tabular Summary

Table 6-5 below summarises the ANOVA results over the entire sample and each of the sub-samples. It also gives the mean IAT scores for each of the IATs in the battery and the range of those means. (A more complete table that includes addition statistics can be found in Appendix C). Respectively IAT means 1, 2 and 3 in the table are for the *White/Black*, the *Female/Male* and the *White Male/Black Female* IATs respectively. The table shows that for most of the samples, the variations in IAT means are relatively large in magnitude, greater than 2.0 units, showing a qualitative difference in the attitudinal strength of the tests according to the IAT rule of thumb. The exception to this is the black sub-sample (range = 0.188). Also, although the black male and white male sub samples have sizeable ranges in their test means (range= 0.466 and 0.436 respectively) the small size of these sub-samples and a lack of between-test statistical significance make it difficult to assert with any authority that there is a qualitative difference in the attitudinal strengths recorded for these sub-samples.

Sample Race	Sample Gender	n	IAT 1 Mean	IAT 2 Mean	IAT 3 Mean	Mean Range	Battery Mean	F	p
All	Both	47	.385	.344	.064	.321	.264	9.235	.000
Black	Both	18	.242	.077	.054	.188	.124	1.158	.322
Indian	Both	10	.321	.523	.138	.385	.327	3.78	.036
White	Both	19	.553	.504	.034	.519	.364	11.443	.000
All	Female	38	.390	.447	.030	.417	.289	15.634	.000
All	Male	9	.364	-.088	.210	.452	.162	2.152	.138
Black	Female	13	.235	.186	.026	.209	.149	.931	.403
Black	Male	5	.259	-.207	.128	.466	.060	2.04	.173
Indian	Female	10	.321	.523	.138	.385	.327	3.78	.036
Indian	Male	0	-	-	-	-	-	-	-
White	Female	15	.569	.622	-.040	.662	.384	31.389	.000
White	Male	4	.496	.060	.312	.436	.289	.530	.606

Table 6-5 Summary of Race/Gender IAT results: IAT means, grand mean and ANOVA statistics

Graphical Representation

The three figures below plot the IAT means for each of the tests in the experimental battery. Line-graphs are plotted for each sub-sample defined by participant race, gender and combination of race and gender respectively. In contrast to the prior two experiments, all three figures show considerable variation in the mean IAT scores between the tests of the IAT battery. The different shapes of the graphs for the various sub-samples are an indication of how the domains were evaluated differently by each sub-sample.

In figure 6-5 below, line-graphs are plotted for the entire sample (All) and for each sub-sample defined by participant race.

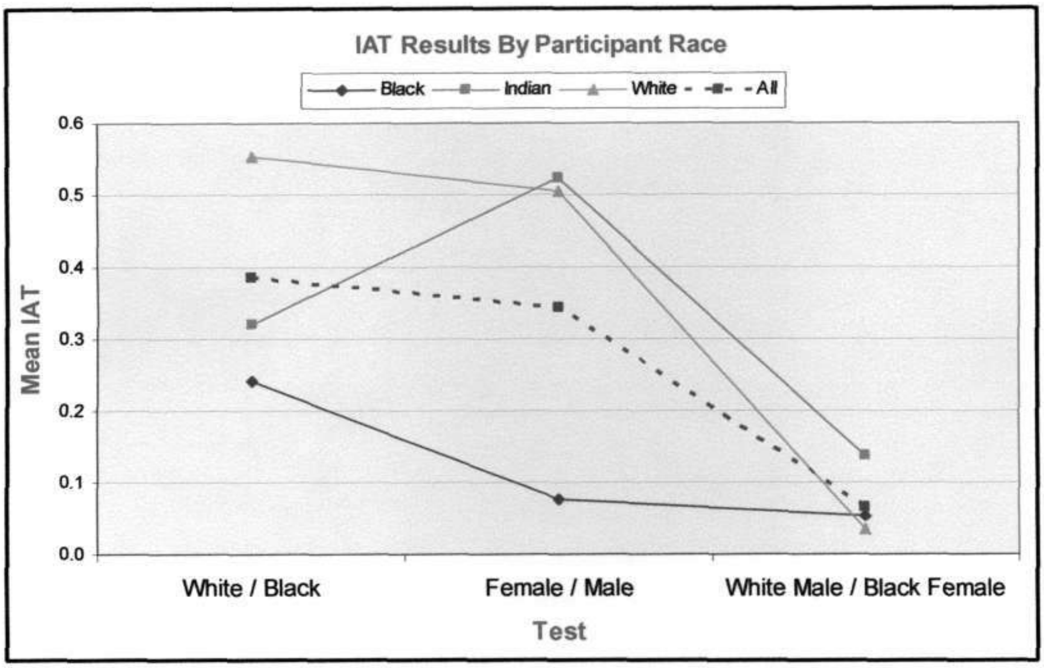

Figure 6-5 Race/Gender IATs: Results by participant race

Figure 6-6 plots the mean IAT results for each sub-sample defined by participant gender.

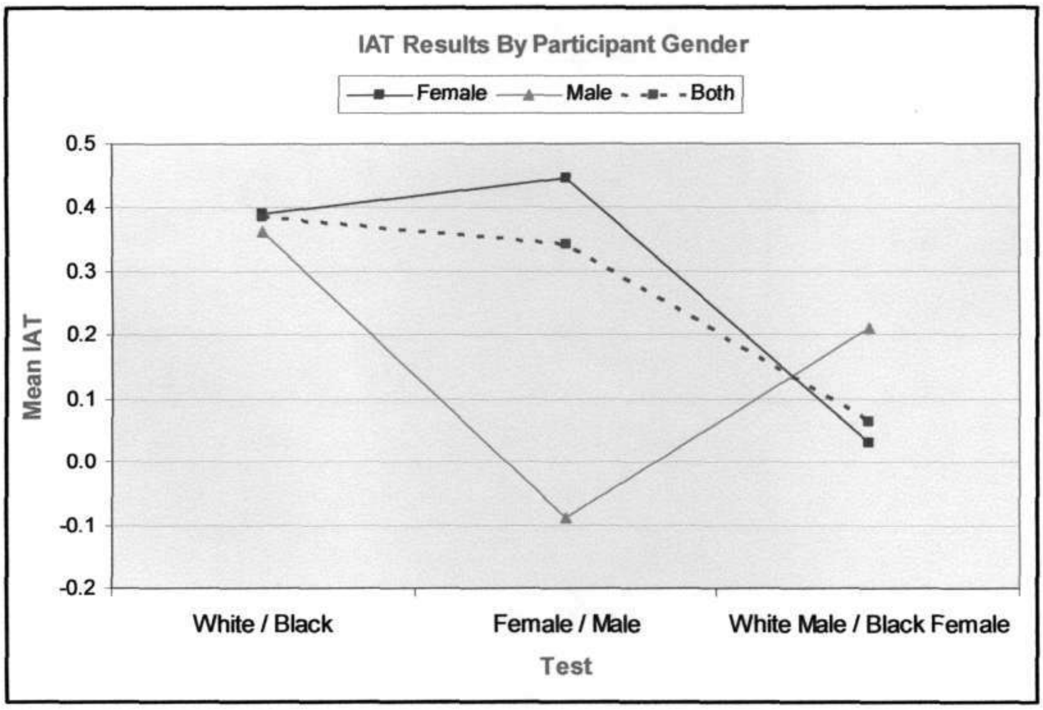

Figure 6-6 Race/Gender IATs: Results by participant gender

Finally, figure 6-7 plots the results obtained for each sub-sample defined by participant race and gender.

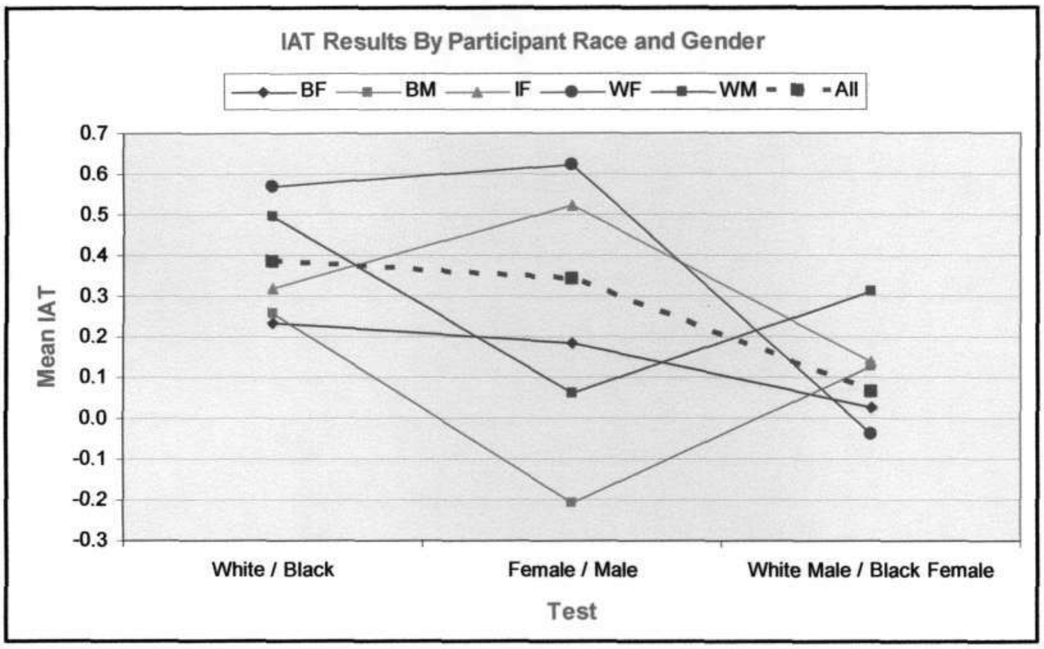

Figure 6-7 Race/Gender IATs: Results by participant race and gender

Additional Observations

One of the secondary investigations of this study was to consider the effects of a composite target dimension on IAT results. With a little manipulation, a pattern becomes apparent. When the direction of the sign of the *Female/Male* IAT results is reversed to effectively obtain *Male/Female* IAT results it becomes evident that the *White Male / Black Female* test results consistently fall between the *White/Black* and *Male/Female* IAT results for all samples and sub-samples with the exception of the black male sample. (Appendix D illustrates this finding graphically in three figures.) This suggests that both race and gender features of exemplars were evaluated by participants completing the composite IAT.

Trend Comparison

The mean IAT effects for the *White/Black* IAT for sub-samples defined by participant race were as follows: black (Mean=0.242), Indian (Mean=0.321), white (Mean=0.553). According to the IAT rule of thumb these scores show slight pro-white preferences amongst blacks and Indians and moderate pro-whites attitudes amongst whites. These results are similar to trends in the literature for white and black samples (Greenwald et al., 1998; Nosek et al., 2002).

The mean IAT effects for the *Female/Male* IAT for sub-samples defined by participant race were as follows: male (Mean=-0.088), female (Mean=0.447). These scores show no gender preference amongst males and slight to moderate pro-female attitudes amongst females. These results are similar to trends in the literature for male and female samples (Rudman & Goodwin, 2004).

Supplementary Results

A between-group ANOVA was carried out for each IAT with groups defined by participant race, by participant gender and by participant race and gender. The ANOVA results are summarised in the table below. These results are not integral to this study and will not be discussed in detail. A brief point of interest is that while the *White/Black* IAT showed significant differences between groups defined by race (p=0.024) and the *Female/Male* IAT showed significant differences between groups defined by gender (p<0.001) the composite *White Male/Black* Female IAT showed no significant differences between any of the groups under investigation. This suggests further evidence that neither race nor gender were exclusively salient in evaluations in the composite IAT otherwise results would have tended to be similar to those obtained for either the race or the gender IATs.

Test	Grouped By	Mean	SD	F	p
IAT 1: White/Black	Race	0.385	0.363	4.073	0.024
	Gender	0.385	0.363	0.034	0.854
	Race and Gender	0.385	0.363	1.989	0.114
IAT 2: Female/Male	Race	0.344	0.421	7.608	0.001
	Gender	0.344	0.421	15.469	< 0.001
	Race and Gender	0.344	0.421	8.520	< 0.001
IAT 3: Black Male/White Female	Race	0.064	0.396	0.227	0.797
	Gender	0.064	0.396	1.518	0.224
	Race and Gender	0.064	0.396	0.783	0.542

Table 6-6 Race/Gender IATs: Supplementary between-group ANOVA results

Results Summary

Three experiments were conducted primarily to investigate the effects of making minor and major semantic changes to the target dimension of the Implicit Association Test. The results that are pertinent to the main aims of this study are summarised below.

The first experiment targeted a race/culture domain and compared results obtained for IAT with target categories *White/Black*, *Western/African* and *Previously Advantaged, Previously Disadvantaged.* A between-test ANOVA was carried out over the entire sample and homogenous sub-samples defined by participant race and by participant race and gender. In all cases the hypothesis that the IAT would be unable to discriminate between the tests at a level of statistical significance was confirmed.

The second experiment targeted a gender domain and compared results obtained for IATs with target categories *Female/Male*, *Girl/Boy* and *He, She*. A between-test ANOVA was carried out over the entire sample and homogenous sub-samples defined by participant gender and by participant race and gender. In all cases the hypothesis that the IAT would be unable to discriminate between the tests at a level of statistical significance was confirmed.

The third experiment investigated whether the IAT would discriminate between target dimensions with major semantic differences, defined by the categories *White/Black*, *Female/Male* and *White Male, Black Female*. An ANOVA was carried out over the entire sample and homogenous sub-samples defined by participant race, gender and participant race and gender. The hypothesis that the IAT would be able to discriminate between the tests at a level of statistical significance was confirmed over the entire sample and most sub-samples that had sufficiently large sample sizes.

These experimental results are discussed in more detail in the following chapter.

Chapter 7: Discussion

This chapter discusses the experimental findings of this study and makes some suggestions as to what these tell us about the IAT's limitations and capabilities. While at first glance the findings appear conclusive, there are various considerations that should be taken into account before inferences are made.

Study Limitations

Before discussing the experimental results, a number of points should be made about the limitations of this study.

Probably the most pertinent criticism of this study is its lack of a formal approach for selecting semantically major and minor differences in categories. This study, investigated IATs having 'minor' and 'major' semantic differences in their target dimension, but selection of IAT categories was based on the author's judgement. This study could have been enhanced by providing a more compelling rationale for asserting that target dimensions have minor or major semantic differences. For example, an experiment could have been devised whereby participants ranked words and phrases (amongst them those used as the IAT target categories), according to their semantic closeness to a base category word. This criticism notwithstanding, the distinctions between the 'minor' IAT categories are inarguably more subtle than the distinctions between the 'major' categories.

With regards to the sampling, three limitations were identifiable. Firstly, the sample was drawn from a student population and may not be representative of a more general South African population. This is by no means a problem unique to this study. Much of the IAT literature is based on experiments with student samples. Secondly, a larger sample size would have been preferable for the experiments, to provide sufficient statistical power for sub-sample hypothesis testing especially for those sub-samples that were defined by participant race *and* gender. In general sample sizes were sufficiently large, to provide the power required for testing over the entire experimental sample and the sub-samples defined by race *or* gender[6]. Experiment 3, especially suffered from a scarcity of males in the sample (n=9). Thirdly, the algorithm that allocated participants to experiments and batteries, while performing reasonably well, could have been more optimal.

Interrogating the Results

At face value, the experimental results confirm the hypothesis that the IAT does not discriminate between tests having minor semantic differences in their target dimension (at least for race and gender domains) but is able to do so for tests having major semantic differences in their target dimensions alone. However, a close inspection of the experimental results raises a number of points that qualify these findings. Breaking the usual pattern, discussion of the experiment that investigated major semantic variation in the target dimension will precede discussion of the experiments that investigated minor semantic variation in the target dimension.

Major Semantic Variations in the Target Dimension

Experiment 3 emulated Mitchell et al.'s (2003) experiment that compared results for IATs having major semantic differences from one another in their target dimension alone. However, whereas their data was analysed over their entire sample only, this study analysed the data over the entire sample and for the various sub-samples defined by race, gender and a

[6] More detail regarding statistical power is provided in later discussion.

combination of race and gender. For this study, the alternate hypothesis that the IAT would discriminate between the tests was confirmed over the entire sample and for a majority of the larger sub-samples. However, there were a number of sub-samples for which the alternate hypothesis was not confirmed, in particular: the black male (p=0.173), white male (p=0.606), male (p=0.138), black female (p=0.403) and black (p=0.322) sub-samples. That is, all male and black sub-samples yielded results contrary to expectations.

With regards to the male sub-samples it is likely that the divergence from expectations was due to the small sizes of the sub-samples: black males (n=5), white males (n=4), and no Indian males, giving a total of only 9 males. Since small sample sizes are deficient in statistical power, this would increase the probability of an erroneous rejection of the alternate hypothesis. Given that the range between the IAT battery means (a coarse measure of between-test differences) was relatively large for the black male (range= 0.466), white male (range= 0.436) and male (range=0.422) sub-samples, larger sample sizes could well have led to statistically significant findings.

The black female sub-sample (p=0.403) was also small in size (n=13) making it possible that a null result was influenced by a lack of statistical power. However, this result was in contrast to other experimental sub-samples of comparable size for which the alternate hypothesis was confirmed. A close inspection of the black female IAT results reveals that this sample obtained a slight pro-white result (Mean=0.235) and a slight pro-female result (Mean=0.186) with a resultant neutral result for the composite IAT (Mean=0.026). This is in keeping with findings in the literature which show that black samples sometimes evidence slight implicit preferences for whites over blacks (Nosek et al., 2002) and that female samples tend to show a preference for female over male (Rudman & Goodwin, 2004). Since the literature is clear that exemplars in the IAT are not evaluated independently of their parent categories[7], the similarity in IAT magnitudes for the race and gender IATs should be considered to be coincidental, analogous to a group of students happening to achieve similar average scores for tests in different subject domains such as Psychology and English. The null result of the black sub-sample as a whole may be attributed to the influence of the high proportion of females in that sub-sample (13 of 18).

In sum, these findings indicate that the IAT is capable of discriminating between IATs having major semantic variations in their target dimensions alone, although it is incorrect to suppose that a difference in IAT scores will always be in evidence. A coincidental similarity in IAT effect magnitudes may occur across disparate domains. On the whole, over the entire sample and for sub-samples that were sufficiently large the alternate hypothesis was confirmed for experiment 3. This study thus corroborates Mitchell et al.'s (2003) findings and supports the position in the literature that the IAT categories strongly influence IAT results (De Houwer, in press; Lane et al., 2007; Nosek et al., 2007).

Minor Semantic Variations in the Target Dimension

Experiments 1 and 2 compared results for IATs having minor semantic differences from one another in their target dimensions. There were no precedents for these experiments in the literature. Findings of non-significant between-test differences were found over the entire sample and for all relevant sub-samples in both experiments suggesting that the IAT is unable to discriminate between such IATs. There are however a number of questions to consider.

[7] Exemplars are always evaluated with reference to the categories to which they belong, but may, to a degree, influence the construal of those categories, although their ability to do so is constrained by the IAT category definitions (De Houwer, 2001, in press; Nosek et al., 2007).

Firstly, there is the question of statistical power. From the literature, the IAT is known to produce fairly *large* standardised effect sizes[8] (Greenwald et al., 1998) meaning that sample sizes can be relatively small and still have sufficient power from hypothesis testing using inferential tests. Certainly, fairly small sample sizes of 20 to 80 participants are common in the IAT literature. This suggests that the entire sample and the sub-samples defined by participant race *or* gender were large enough and had sufficient statistical power for hypothesis testing but not the sub-samples defined by a combination of race *and* gender. This notwithstanding, the post hoc examination of the results for the IATs of experiments 1 and 2 show such similar results to one another that it is reasonable to conclude that the acceptance of the null hypothesis for race and gender sub-samples was not attributable to a deficiency in statistical power.

The marked similarity of the between-test IAT results in experiments 1 and 2 bears further investigation as a greater variation between the tests was anticipated. Although the confirmation of the null hypothesis was predicted for experiments 1 and 2, it was anticipated that the IAT results for each experiment would show a reasonable amount of variation from one another. In the literature a moderate semantic change in the attribute dimension from *Pleasant/Unpleasant* to the more personalised *I Like/I Don't Like* resulted in statistically significant differences between IAT means (Olson & Fazio, 2004). With this precedent and the fact that the category definitions are believed to be the primary contributor to IAT scores (De Houwer, 2001, in press) it was expected that the IAT would show more variation than is evidenced in the results. This section advances the possibility that the surprisingly similar results were in part due to a cognitive confound that was not controlled for in the

[8] This statistic should not be confused with the IAT effect.

experimental design. That fact that test results show so little variation over the entire sample and almost all of the homogenous sub-samples in both experiments raises the question of whether the different IATs in the experimental batteries were in fact measuring essentially the same construct. It will be argued that this possibility arises from two sources:

- The unintended priming of participants as a consequence of the social psychology course in which they were enrolled at the university and the process of informed consent that they underwent.
- The cognitive re-definition of the IAT categories to represent the domain that was salient to the participant (Govan & Williams, 2004).

With regards to priming, students in a social psychology course at the university were introduced to the IAT in the context of a number of lectures that examined social attitudes, with a particular focus on sexism and racism. In addition, the process of informed consent made students aware of the possibility that they might find their IAT results uncomfortable or disturbing because their scores might suggest that their implicit racial or gender attitudes differed from their idealised view of these attitudes. It is likely that these factors primed students that the purpose of the experiment was to measure their attitudes towards race and gender in particular, making these constructs salient for participants at the time of completing the IAT. The effects of priming on IAT results have been demonstrated by Dasgupta et al. (2001) who showed that exposing a white participant sample to admired black exemplars, ostensibly as a test of general knowledge, resulted in a significant reduction of pro-white bias as compared to a control sample. Although not directly analogous to the circumstances of this study, this experiment shows that priming can have a considerable effect on a participant's response to the IAT.

Govan and Williams (2004) introduced the idea that under certain conditions IAT categories may be cognitively redefined by IAT participants to fit the exemplars that are instances of the categories. Moreover they showed that this category redefinition appears to persist in multi-test experiments even when the exemplars that effected the re-categorisation are modified to be neutral in subsequent tests of the experiment. Or, in other words, the nature of the exemplars can influence the construal of the category definitions, essentially redefining the construct that is being measured by the IAT. This re-categorisation may then remain fairly stable in subsequent IATs of a multi-test experiment.

A proper appreciation of the foregoing paragraph is best accomplished by revisiting Govan and Williams' (2004) experiment. In their experiment two sample groups completed an *Animal/Plant, Pleasant/Unpleasant* IAT with markedly different results through the differential manipulation of target exemplars. The first group was presented with an IAT having positive *Animal* exemplars and negative *Plant* exemplars, the second group with an IAT having negative *Animal* and positive *Plant* exemplars. Findings showed that the first group evidenced a pro-Animal attitude and the second a pro-Plant attitude that differed significantly from each another. Govan and Williams (2004) attributed this to a cognitive re-categorisation of the first group's IAT target dimension to *Nice Animal/Nasty Plant* and a cognitive re-categorisation of the second group's IAT target dimension to *Nasty Animal/Nice Plant* without which neutral results would have been expected in both groups. After completing the IATs as described above, both groups were then presented with a second IAT, having only a slight modification from the first - all of the prior *Animal* and *Plant* target exemplars were substituted with the neutral words 'animal' and 'plant', effectively eliminating the prior positive or negative connotations of the target exemplars. The experimental findings of the second IAT were identical to the first, with only a slight decline in t-scores. That is, the pro-Animal attitude persisted in the first group and the pro-Plant attitude persisted in the second, in spite of the fact that the target exemplars were no longer polarised. Govan and Williams (2004) saw this as evidence that the cognitive re-categorisation that had been initiated in the first experiment, continued to influence evaluations in the second experiment.

Govan and Williams' (2004) findings have an important bearing on this research. They present the possibility that in experiments 1 and 2 in this study, participants might have cognitively re-defined the IAT target categories in accordance with what was perceived as salient in the exemplars. Thus, for example, in the race/culture IAT, the salience of racial features of the exemplars might have resulted in all IATs in the experiment being evaluated according to the criterion of race, with any nuances in the target categories effectively factored out of the evaluation. The fact that there were only minor semantic differences in the IAT's target dimension, might have allowed this cognitive re-categorisation to remain stable for all three tests of the experimental battery, meaning in effect that for each participant, the same construct was measured by all IATs, with a consequent convergence of IAT results. The unintended pre-experimental priming that heightened the awareness of participants to race and gender domains might have been influential in activating or strengthening this re-categorisation effect.

This interpretation of the experimental results lends itself to explaining the failure of the secondary hypotheses. That is:

- For experiment 1, the expectation that the white sub-sample would evidence a greater pro-white result for the *White/Black* IAT than for the other IATs was not confirmed despite the fact that whites generally have more positive associations with the colour white and more negative associations with the colour black (Smith-Mclallen et al., 2006).

- For experiment 2, the expectation that the female sample would evidence a lesser pro-female result for the *Girl/Boy* IAT than for the other IATs was not confirmed despite the anticipation that the label *Boy* would be seen as less negative than the label *Male* for females who might perceive men as threatening or intimidating.

If, as has been argued, the target categories of the IATs in experiments 1 and 2 were cognitively redefined according to the salient characteristics of the exemplars and this re-categorisation remained stable owing to the relatively minor semantic variations in the IAT definitions, with the result that the same construct was in effect being measured for all tests in the experiment, no particular ranking of scores should be expected for the tests of the experimental batteries.

In summary, this section began with the recognition that experiments 1 and 2 showed no statistically significant between-test differences when comparing results obtained from a battery of three IATs. This implied that the IAT is unable to discriminate between tests having only minor semantic differences in their target dimension. It was established that this failure to discriminate between IATs was not due to a lack of statistical power, at least not for the larger sub-samples. The fact that the IAT results were so similar to one another that they might have been influenced by factors that had not been taken into account in the experimental design was considered. A technical explanation for the lack of variation in between-test scores was then advanced based on findings in the literature. It was hypothesised that priming effects and a cognitive redefinition of the IAT target dimension in accordance with the salient characteristics of the target exemplars could have resulted in all three IATs in the experimental batteries effectively measuring the same construct, with a resultant convergence of IAT scores. It was also shown how such an interpretation could explain the failures of the secondary hypotheses of these experiments. It should be noted that much of this reasoning is speculative and further experimentation would be required in order to test this hypothesis. Nonetheless, indications are that the IAT does not appear to be capable of discriminating between tests having minor semantic variations in their target dimension, at least not when a within-subject design is used.

IAT Limitations and Capabilities

The purpose of this study was to add to the body of knowledge that describes the limitations and capability of the IAT as a psychological instrument for investigating implicit cognitions. The particular focus of this research has been to gain a better understanding with regards to the sensitivity of the IAT to semantic variations in its target dimension.

In agreement with the findings of Mitchell et al. (2003), but investigating different social domains, this study found that the IAT is indeed sensitive to major semantic changes in the target dimension. It was noted, however, that major semantic differences in the target dimension do not necessarily guarantee differences in IAT results.

With regards to the IAT's sensitivity to minor semantic variations in its target dimension it would appear that the IAT is unable to discriminate between such variations. If this is the case, it suggests that the IAT is insensitive to subtle distinctions in attitude. However this finding is not conclusive. It is possible that uncontrolled effects might have influenced these results to a degree. In particular there might have been a temporary cognitive re-categorisation of the IAT categories in accordance with what was perceived as the salient characteristics of the exemplars, resulting in all tests in the battery being similarly evaluated. Further experimentation is required to test this hypothesis.

Extending this Study

The investigation into the IAT's sensitivity to minor semantic variations in its target dimension has raised a number of questions that might assist in furthering an understanding of the cognitive processes that underlie the IAT. The marked similarity in experimental results between the IATs on this study raises the possibility that they were in effect measuring the same construct. A hypothesis has been advanced that this might have been due to priming effects and the cognitive redefinition of the IAT categories to fit what was salient in the exemplars. This hints at the possibility that the IAT activates a basic automatic response that is resistant to change when only minor changes in IAT definition are introduced. Pursuing this hypothesis would help to clarify whether or not (and under what conditions) the IAT is capable of discriminating between minor semantic changes in it definition, but might, in addition, contribute to an understanding of the underlying cognitive mechanisms that the IAT activates.

A possible research avenue to follow would be to investigate the effects of *explicitly* defining the category labels before a participant completes the IAT, an approach used by Dewitte, De Houwer, & Buysse (in press) in a recent study. Deliberately emphasising the meanings of the category labels prior to testing could potentially inhibit the hypothesised category re-definition that might have affected results in this study. Experiments using this approach could compare the results obtained from an experimental group that received explicit definitions for category labels with those obtained from a control group that did not. Both within-subject and between-subject designs could yield results of interest.

Conclusion

The IAT is a psychological test that is widely used to measure implicit attitudes. While a great deal of research has investigated the IAT's capabilities and limitations, these are still not fully understood. This study has sought to contribute to a niche of IAT research that investigates the effects of exemplar and category manipulation on IAT results. Within this niche, the particular focus of this research has been on determining the IAT's sensitivity to major and minor semantic modifications to its target dimension. Findings revealed that the IAT is able to detect major semantic modifications in its target dimension, but appears unable to discriminate between tests having minor semantic modifications in this dimension with the implication that it is incapable of detecting subtle distinctions in attitude. However, this latter finding was inconclusive as the marked similarity in test results raised the possibility that uncontrolled factors that had not been taken into account in the experimental design might have impacted upon the test results. It was hypothesised that priming effects and a cognitive redefinition of the IAT target dimension in accordance with the salient characteristics of the target exemplars could have resulted in all three IATs in the experimental batteries effectively measuring the same construct, with a consequent similarity in IAT results. This hypothesis should be tested before the question of the IATs abilities with regards to discriminating between minor semantic differences in its target dimension is decided.

References

Aidman, E., & Carroll, S. (2003). Implicit individual differences: relationships between implicit self-esteem, gender identity, and gender attitudes. *European Journal of Personality, 17*(1), 19-37.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: reliability, validity, and controllability of the IAT. *Z Exp Psychol, 48*(2), 145-160.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science, 17*(1), 53–58.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*, 27–41.

Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology 42*(2), 163-176.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science, 15* (12), 806–813.

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. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37*(6), 443-451.

De Houwer, J. (2006). What are implicit measures and why are we using them? In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11-28): Thousand Oaks, CA: Sage Publishers.

De Houwer, J. (in press). Comparing measures of attitudes at the functional level and procedural level: Analysis and implications. In R. Petty, R. H. Fazio & P. Brinol (Eds.), *Attitudes:Insights from the new implicit measures.*: Erlbaum.

Dewitte, M., De Houwer, J., & Buysse, A. (in press). On the role of the implicit self-concept in adult attachment. *European Journal of Psychological Assessment.*

Fazio, R. H., & Olson, M. A. (2003). Implicit Measures in Social Cognition Research: Their Meaning and Use. *Annual Review of Psychology*, 297-328.

General Information. (2007). Retrieved 9 March, 2008, from http://www.projectimplicit.net/generalinfo.php

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology, 40*(3), 357–365.

Greenwald, A. G. (2004, January). *Revised top 10 list of things wrong with the IAT.* Paper presented at the Attitudes Preconference of the 5th annual meeting of the Society of Personality and Social Psychology, Austin, TX.

Greenwald, A. G. (2005). Generic iat zipfile download. Retrieved 15 June, 2006, from http://faculty.washington.edu/agg/iat_materials.htm

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., 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.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464-1480.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie, 48*(2), 85-93.

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.

Han, H. A., Olson, M. A., & Fazio, R. H. (2006). The influence of experimentally created extrapersonal associations on the Implicit Association Test. *Journal of Experimental Social Psychology, 42*(3), 259-272.

Jelenec, P., & Steffens, M. C. (2002). Implicit attitudes toward elderly women and men. *Current Research in Social Psychology, 7*(16), 275-293.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81*(5), 774-788.

Kihlstrom, J. F. (2004). Implicit methods in social psychology. In C. Sansone, C. C. Morf & A. T. Panter (Eds.), *The Sage Handbook of Methods in Social Psychology* (pp. 195–212): Thousand Oaks, CA: Sage.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: IV: Procedures and validity. In B. Wittenbrink & N. Schwarz (Eds.).

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.

Maison, D., Gregg, A. P., & Bruin, R. (2002). The Implicit Association Test as a measure of implicit consumer attitudes. *Polish Psychological Bulletin, 32*(1), 61–69.

Mast, M. S. (2004). Men Are Hierarchical, Women Are Egalitarian: An Implicit Gender Stereotype. *Swiss Journal of Psychology, 63*(2), 107-111.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General, 132*(3), 455-469.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics: Theory, Research, and Practice, 6*(1), 101-115.

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.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior.*: Psychology Press.

Nosek, B. A., & Hansen, J. J. (2005). The Associations in our Heads Belong to Us: Searching for Attitudes and Knowledge in Implicit Evaluation. University of Virginia.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extra-personal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86*(5), 653-667.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., et al. (2000). Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation. *Journal of Cognitive Neuroscience, 12*(5), 729-738.

Rudman, L. A., Feinberg, J., & Fairchild, K. (2002). Minority members' implicit attitudes: Automatic ingroup bias as a function of group status. *Social Cognition, 20*(4), 294-320.

Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men. *Journal of Personality and Social Psychology, 87*(4), 494-509.

Schacter, D. L. (1987). Implicit memory: history and current status (Vol. 13, pp. 501-518).

Smith-Mclallen, A., Johnson, B. T., Dovidio, J. F., & Pearson, A. R. (2006). Black and white: The role of color bias in implicit race bias. *Social Cognition, 24*(1), 46-73.

Steffens, M. C., & Plewe, I. (2001). Items'cross-category associations as a confounding factor in the Implicit Association Test. *Zeitschrift für Experimentelle Psychologie, 48*(2), 123-134.

Teachman, B. A., Gregg, A. P., & Woody, S. R. (2001). Implicit associations for fear-relevant stimuli among individuals with snake and spider fears. *Journal of Abnormal Psychology, 110*(2), 226–235.

Appendix A: Exemplars

Black African Prev. Disadv	White Western Prev. Adv	Good	Bad
bheki	andrew	freedom	awful
busi	jackie	gladness	evil
jabulani	johannes	happy	failure
lindiwe	maria	joy	horrible
nomvula	mark	laughter	hurt
sibusiso	michelle	love	nasty
sipho	patrick	peace	negative
thabo	sarie	pleasure	pain
thandi	stephanus	positive	sadness
zanele	susan	wonderful	terrible

Table A-1 Experiment 1 Exemplars

Male Boy He	Female Girl She	Good	Bad
andrew	busi	freedom	awful
bheki	jackie	gladness	evil
jabulani	lindiwe	happy	failure
johannes	maria	joy	horrible
mark	michelle	laughter	hurt
patrick	nomvula	love	nasty
sibusiso	sarie	peace	negative
sipho	susan	pleasure	pain
stephanus	thandi	positive	sadness
thabo	zanele	wonderful	terrible

Table A-2 Experiment 2 Exemplars

White Male White Male	Black Female Black Female	Good	Bad
andrew	busi	freedom	awful
andrew	busi	gladness	evil
johannes	lindiwe	happy	failure
johannes	lindiwe	joy	horrible
mark	nomvula	laughter	hurt
mark	nomvula	love	nasty
patrick	thandi	peace	negative
patrick	thandi	pleasure	pain
stephanus	zanele	positive	sadness
stephanus	zanele	wonderful	terrible

Table A-3 Experiment 3 Exemplars

Appendix B: Battery/Test Allocation Distributions

In a perfectly counterbalanced experiment every possible sequence of tests would be completed an equal number of times. More practically, because sample sizes are not always exactly divisible by the counterbalancing factor, an *optimal* allocation strategy ensures that each test sequence is completed at most once more than any other test sequence for a given sample.

This appendix gives the allocation distributions for experiments 1, 2 and 3, for each sample and the relevant sub-samples. The data are presented in tabular format at two levels, by *battery sequence* and by the number of times each *test* was presented first, second and third within a battery. The column labelled with a '?', indicates the number of re-allocations[9] that would be required to achieve optimal battery or test allocations for the sample in question. The greater this value, the less optimal the allocations.

In general, an inspection of the data revealed that at the *battery sequence* level the smaller sub-samples were optimally or near-optimally allocated, whereas the larger samples were less optimally (but still reasonably equitably) distributed. When inspecting the data at the level of *test presentation order*, almost all of the samples and sub-samples showed little deviation from an optimal allocation. This is particularly important because it is a known procedural effect of the IAT that test presentation order impacts upon IAT results in multi-test experiments (Greenwald & Nosek, 2001; Greenwald et al., 2003).

In summary, it appears that the computerised counterbalancing algorithm performed reasonably well and it is unlikely that the experimental results were unduly influenced by order effects.

The allocation distributions for each experiment are presented on the pages that follow.

[9] A re-allocation is defined as the removal of an allocation from a condition that is over-represented to a condition that is under-represented.

Experiment 1 – Battery/Test Allocation Distribution

In Table B-1 and Table B-2 below IATs 1, 2 and 3 refer respectively to the *Black/White*, *Western/African* and *Previously Advantaged/Previously Disadvantaged* IATs.

		Battery Test Sequence – Times Completed						
Sample	n	IATs 1,2,3	IATs 1,3,2	IATs 2,1,3	IATs 2,3,1	IATs 3,1,2	IATs 3,2,1	?
All	63	12	9	9	12	13	8	5
Black	22	5	3	3	4	4	3	1
Indian	17	4	2	3	2	3	3	1
White	24	3	4	3	6	6	2	4
Black Female	13	3	2	2	2	2	2	0
Black Male	9	2	1	1	2	2	1	0
Indian Female	12	3	1	2	2	2	2	1
Indian Male	5	1	1	1	0	1	1	0
White Female	15	2	2	2	4	4	1	2
White Male	9	1	2	1	2	2	1	0

Table B-1 Experiment 1: Battery sequence distribution

		Completed 1st				Completed 2nd				Completed 3rd			
Sample	n	IAT1	IAT2	IAT3	?	IAT1	IAT2	IAT3	?	IAT1	IAT2	IAT3	?
All	63	21	21	21	0	22	20	21	1	20	22	21	1
Black	22	8	7	7	0	7	8	7	0	7	7	8	0
Indian	17	6	5	6	0	6	7	4	1	5	5	7	1
White	24	7	9	8	1	9	5	10	3	8	10	6	2
Black Female	13	5	4	4	0	4	5	4	0	4	4	5	0
Black Male	9	3	3	3	0	3	3	3	0	3	3	3	0
Indian Female	12	4	4	4	0	4	5	3	1	4	3	5	1
Indian Male	5	2	1	2	0	2	2	1		1	2	2	0
White Female	15	4	6	5	1	6	3	6	2	5	6	4	1
White Male	9	3	3	3	0	3	2	4	1	3	4	2	1

Table B-2 Experiment 1: Test presentation order distribution

Experiment 2 – Battery/Test Allocation Distribution

In Table B-3 and Table B-4 below IATs 1, 2 and 3 refer respectively to the *Female/Male*, *Girl/Boy* and *She/He* IATs.

		Battery Test Sequence – Times Completed						
Sample	n	IATs 1,2,3	IATs 1,3,2	IATs 2,1,3	IATs 2,3,1	IATs 3,1,2	IATs 3,2,1	?
All	59	13	9	8	11	9	9	4
Female	40	10	5	6	8	5	6	4
Male	19	3	4	2	3	4	3	1
Black Female	14	3	2	2	3	2	2	0
Black Male	5	1	1	0	1	1	1	0
Indian Female	9	3	1	2	1	0	2	1
Indian Male	6	1	1	1	1	1	1	0
White Female	17	4	2	2	4	3	2	2
White Male	8	1	2	1	1	2	1	0

Table B-3 Experiment 2: Battery sequence distribution

		Completed 1st				Completed 2nd				Completed 3rd			
Sample	n	IAT1	IAT2	IAT3	?	IAT1	IAT2	IAT3	?	IAT1	IAT2	IAT3	?
All	59	22	19	18	2	17	22	20	2	20	18	21	1
Female	40	15	14	11	2	11	16	13	2	14	10	16	3
Male	19	7	5	7	1	6	6	7	0	6	8	5	1
Black Female	14	5	5	4	0	4	5	5	0	5	4	5	0
Black Male	5	2	1	2	0	1	2	2	0	2	2	1	0
Indian Female	9	4	3	2	1	2	5	2	2	3	1	5	2
Indian Male	6	2	2	2	0	2	2	2	0	2	2	2	0
White Female	17	6	6	5	0	5	6	6	0	6	5	6	0
White Male	8	3	2	3	0	3	2	3	0	2	4	2	1

Table B-4 Experiment 2: Test presentation order distribution

Experiment 3 – Battery/Test Allocation Distribution

In Table B-5 and Table B-6 below IATs 1, 2 and 3 refer respectively to the *Black/White*, *Female/Male* and *White Male/Black Female* IATs.

		Battery Test Sequence – Times Completed						
Sample	n	IATs 1,2,3	IATs 1,3,2	IATs 2,1,3	IATs 2,3,1	IATs 3,1,2	IATs 3,2,1	?
All	47	8	8	8	9	9	5	2
Black	18	3	3	3	4	4	1	2
Indian	10	2	2	2	1	1	2	0
White	19	3	3	3	4	4	2	1
Female	38	7	6	6	7	7	5	1
Male	9	1	2	2	2	2	0	1
Black Female	13	2	2	2	3	3	1	1
Black Male	5	1	1	1	1	1	0	0
Indian Female	10	2	2	2	1	1	2	0
Indian Male	0	-	-	-	-	-	-	0
White Female	15	3	2	2	3	3	2	0
White Male	4	0	1	1	1	1	0	0

Table B-5 Experiment 3: Battery sequence distribution

		Completed 1st				Completed 2nd				Completed 3rd			
Sample	n	IAT1	IAT2	IAT3	?	IAT 1	IAT2	IAT3	?	IAT1	IAT2	IAT3	?
All	47	16	17	14	1	17	13	17	2	14	17	16	1
Female	18	6	7	5	1	7	4	7	2	5	7	6	1
Male	10	4	3	3	0	3	4	3	0	3	3	4	0
	19	6	7	6	0	7	5	7	1	6	7	6	0
Black													
Indian	38	13	13	12	0	13	12	13	0	12	13	13	0
White	9	3	4	2	1	4	1	4	2	2	4	3	1
Black Female	13	4	5	4	0	5	3	5	1	4	5	4	0
Black Male	5	2	2	1	0	2	1	2	0	1	2	2	0
Indian Female	10	4	3	3	0	3	4	3	0	3	3	4	0
Indian Male	0	-	-	-		-	-	-		-	-	-	
White Female	15	5	5	5	0	5	5	5	0	5	5	5	0
White Male	4	1	2	1	0	2	0	2	1	1	2	1	0

Table B-6 Experiment 3: Test presentation order distribution

Appendix C: Between-Test Results

Sample	Sample		IAT 1: White / Black				IAT 2: Western / African				IAT 3: Prev Adv / Prev Disadv				Battery			ANOVA	
Race	Gender	n	Mean	SD	95% Conf Int		Mean	SD	95% Conf Int		Mean	SD	95% Conf Int		Mean	SD	Range	F	p
All	Both	63	.292	.380	.199	.385	.349	.371	.256	.442	.325	.373	.232	.418	.322	.373	.057	.362	.697
Black	Both	22	.011	.348	-.151	.173	.128	.368	-.034	.289	.136	.419	-.026	.297	.091	.378	.125	.747	.478
Indian	Both	17	.404	.228	.263	.545	.412	.316	.271	.553	.351	.314	.210	.492	.389	.284	.061	.221	.802
White	Both	24	.471	.353	.340	.602	.507	.319	.375	.638	.480	.293	.349	.611	.486	.373	.036	.080	.923
Black	Female	13	.031	.286	-.165	.227	.055	.352	-.141	.251	.126	.398	-.070	.322	.071	.341	.095	.259	.773
Black	Male	9	-.019	.440	-.317	.280	.233	.385	-.066	.532	.150	.473	-.148	.449	.122	.430	.252	.786	.467
Indian	Female	12	.398	.202	.258	.539	.372	.249	.231	.512	.319	.262	.178	.459	.363	.235	.079	.343	.712
Indian	Male	5	.417	.310	.019	.815	.508	.356	.110	.906	.429	.440	.030	.827	.451	.381	.091	.074	.929
White	Female	15	.417	.406	.232	.603	.464	.356	.278	.650	.475	.299	.289	.661	.452	.349	.058	.110	.896
White	Male	9	.560	.233	.380	.741	.579	.371	.398	.759	.488	.301	.308	.668	.542	.255	.091	.303	.741

Table C-1 Experiment 1: Detailed between-test ANOVA results: including means, standard deviations and confidence intervals for each IAT

Sample Race	Sample Gender	n	IAT 1: Male / Female				IAT 2: Boy / Girl				IAT 3: He / She				Battery			ANOVA	
			Mean	SD	95% Conf Int		Mean	SD	95% Conf Int		Mean	SD	95% Conf Int		Mean	SD	Range	F	p
All	Both	59	.266	.418	.167	.364	.290	.376	.191	.388	.237	.354	.138	.336	.264	.382	.053	.279	.757
All	Female	40	.440	.328	.343	.538	.439	.315	.341	.536	.387	.292	.290	.485	.422	.310	.053	.374	.689
All	Male	19	-.102	.347	-.241	.036	-.024	.297	-.162	.115	-.079	.251	-.218	.059	-.068	.298	.078	.340	.713
Black	Female	14	.480	.379	.299	.660	.488	.278	.307	.669	.382	.339	.201	.563	.450	.330	.106	.433	.651
Black	Male	5	-.340	.333	-.653	.027	-.011	.347	-.324	.302	-.077	.279	-.390	.236	-.143	.331	.329	1.471	.268
Indian	Female	9	.324	.210	.173	.475	.478	.228	.327	.629	.434	.220	.283	.585	.412	.221	.154	1.178	.325
Indian	Male	6	-.182	.282	-.453	.089	-.100	.327	-.372	.171	-.086	.324	-.357	.186	-.123	.296	.096	.166	.849
White	Female	17	.469	.336	.303	.635	.377	.382	.212	.543	.367	.297	.201	.533	.405	.337	.102	.466	.630
White	Male	8	.106	.307	-.089	.302	.026	.272	-.169	.221	-.076	.208	-.271	.120	.019	.265	.182	.943	.405

Table C-2 Experiment 2: Detailed between-test ANOVA results: including means, standard deviations and confidence intervals for each IAT

Sample Race	Sample Gender		IAT 1: White / Black				IAT 2: Male / Female				IAT 3: W. Male / B. Female				Battery			ANOVA	
		n	Mean	SD	95% Conf Int		Mean	SD	95% Conf Int		Mean	SD	95% Conf Int		Mean	SD	Range	F	p
All	Both	47	.385	.363	.271	.498	.344	.421	.231	.458	.064	.396	-.050	.178	.264	.417	.321	9.235	.000
Black	Both	18	.242	.385	.051	.433	.077	.390	-.114	.268	.054	.434	-.137	.245	.124	.405	.188	1.158	.322
Indian	Both	10	.321	.287	.118	.524	.523	.320	.320	.726	.138	.330	-.065	.341	.327	.341	.385	3.78	.036
White	Both	19	.553	.322	.383	.723	.504	.375	.334	.674	.034	.406	-.136	.204	.364	.433	.519	11.443	.000
All	Female	38	.390	.353	.276	.503	.447	.358	.334	.560	.030	.347	-.084	.143	.289	.396	.417	15.634	.000
All	Male	9	.364	.425	.041	.688	-.088	.405	-.411	.235	.210	.564	-.114	.533	.162	.490	.452	2.152	.138
Black	Female	13	.235	.416	.005	.465	.186	.386	-.044	.416	.026	.425	-.204	.256	.149	.408	.209	.931	.403
Black	Male	5	.259	.333	-.108	.625	-.207	.249	-.537	. 159	.128	.501	-.238	.494	.060	.403	.466	2.040	.173
Indian	Female	10	.321	.287	.118	.524	.523	.320	.210	.445	.138	.330	-.065	.341	.327	.341	.385	3.78	.036
White	Female	15	.569	.263	.436	.701	.622	.213	.490	.755	-.040	.282	-.172	.093	.384	.392	.662	31.389	.000
White	Male	4	.496	.540	-.183	1.176	.060	.549	-.619	.740	.312	.700	-.368	.991	.289	.574	.436	.530	.606

Table C-3 Experiment 3: Detailed between-test ANOVA results: including means, standard deviations and confidence intervals for each IAT

Appendix D: IAT Results for Experiment 3 IATs on converting the Female/Male IAT results to Male/Female IAT results through a reversal of sign.

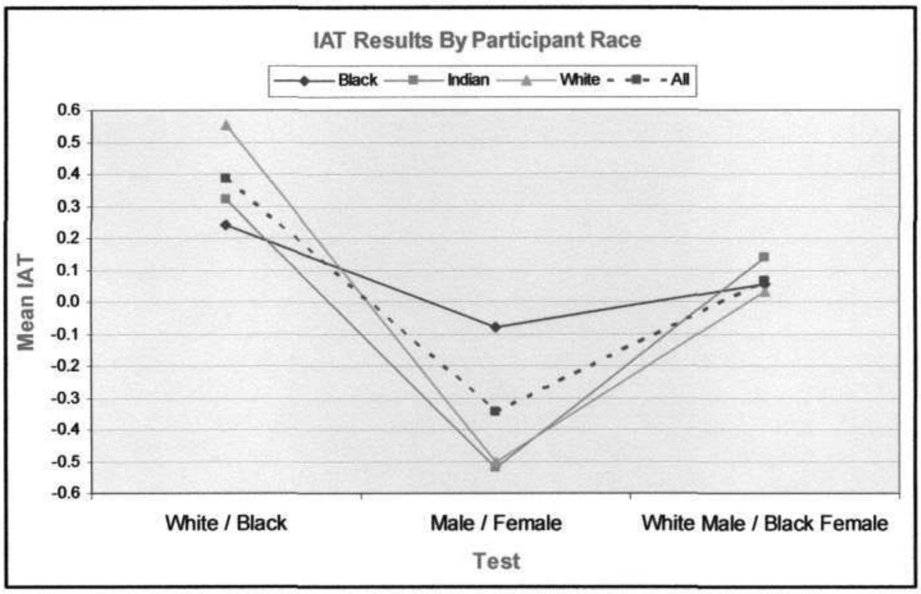

Figure D-1 IAT results by participant race. Composite IAT results fall between race and gender IAT results.

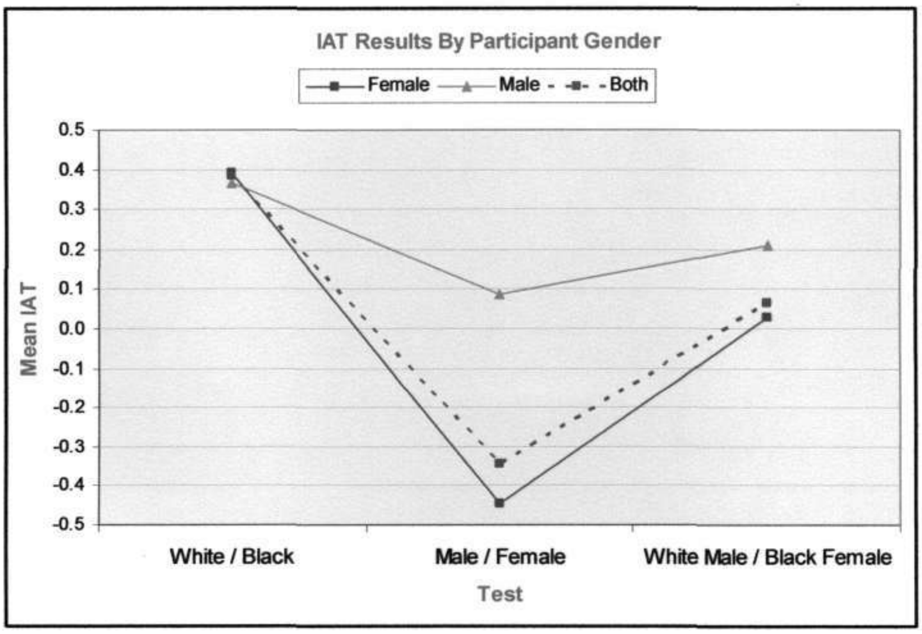

Figure D-2 IAT results by participant gender. Composite IAT results fall between race and gender IAT results.

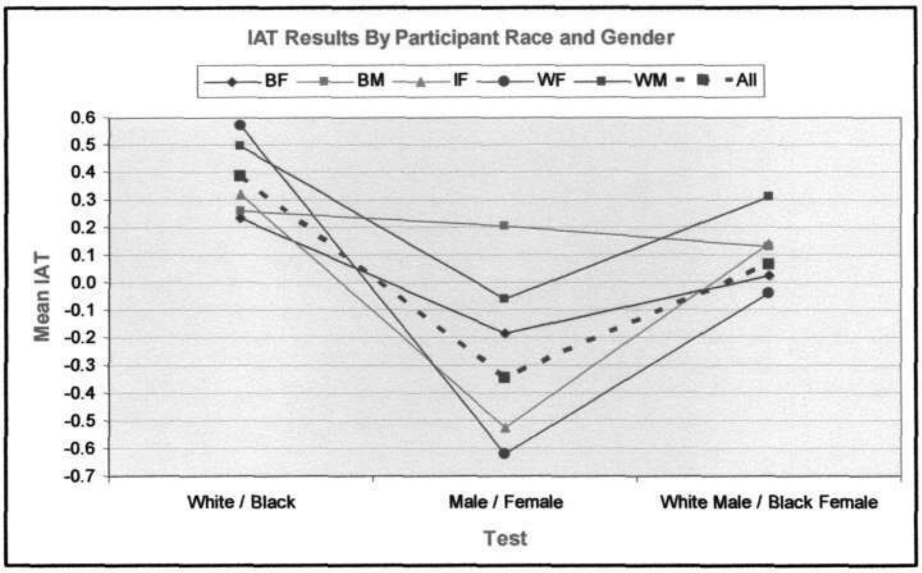

Figure D-3 IAT results by participant race and gender. Composite IAT results fall between race and gender IAT results except for the black male sub-sample.

Appendix E: Supplementary Between-Group Results

Test	Sample Group	n	Mean	SD	F	p
IAT 1: White/Black	Black	22	.011	0.348		
	Indian	17	.404	0.228		
	White	24	.471	0.353		
	ANOVA		.292	0.380	13.097	.000
IAT 2: Western/African	Black	22	.128	0.368		
	Indian	17	.412	0.316		
	White	24	.507	0.319		
	ANOVA	63	.349	0.371	7.711	.001
IAT 3: Prev Adv/Prev Disadv	Black	22	.136	0.419		
	Indian	17	.351	0.314		
	White	24	.480	0.293		
	ANOVA	63	.325	0.373	5.690	.005

Table D-1 Experiment 1: Supplementary between-group results by participant race

Test	Sample Group	n	Mean	SD	F	p
IAT 1: White/Black	Black Female	13	.031	.286		
	Black Male	9	-.019	.440		
	Indian Female	12	.398	.202		
	Indian Male	5	.417	.310		
	White Female	15	.417	.406		
	White Male	9	.560	.233		
	ANOVA		.292	.380	5.324	.000
IAT 2: Western/African	Black Female	13	.055	.352		
	Black Male	9	.233	.385		
	Indian Female	12	.372	.249		
	Indian Male	5	.508	.460		
	White Female	15	.464	.356		
	White Male	9	.579	.247		
	ANOVA		.349	.371	3.614	.007
IAT 3: Prev Adv/Prev Disadv	Black Female	13	.126	.398		
	Black Male	9	.150	.473		
	Indian Female	12	.319	.262		
	Indian Male	5	.429	.440		
	White Female	15	.475	.299		
	White Male	9	.488	.301		
	ANOVA		.325	.373	2.250	.062

Table D-2 Experiment 1: Supplementary between-group results by participant race and gender

Test	Grouped By	n	Mean	SD	t	p
IAT 1: Female/Male	Female	40	.440	.328		
	Male	19	-.102	.347		
	t-test	59	.266	.418	5.827	.000
IAT 2: Girl/Boy	Female	40	.439	.315		
	Male	19	-.024	.297		
	t-test	59	.290	.376	5.367	.000
IAT 3: She/He	Female	40	.387	.292		
	Male	19	-.079	.251		
	t-test	59	.237	.354	5.982	.000

Table D-3 Experiment 2: Supplementary between-group results by participant gender

Test	Grouped By	n	Mean	SD	F	p
IAT 1: Female/Male	Black Female	14	.480	.379		
	Black Male	5	-.340	.333		
	Indian Female	9	.324	.210		
	Indian Male	6	-.182	.282		
	White Female	17	.469	.336		
	White Male	8	.106	.307		
	ANOVA		.266	.418	8.841	.000
IAT 2: Girl/Boy	Black Female	14	.488	.278		
	Black Male	5	-.011	.347		
	Indian Female	9	.478	.228		
	Indian Male	6	-.100	.327		
	White Female	17	.377	.382		
	White Male	8	.026	.272		
	ANOVA		.290	.376	5.861	.000
IAT 3: She/He	Black Female	14	.382	.339		
	Black Male	5	-.077	.279		
	Indian Female	9	.434	.220		
	Indian Male	6	-.086	.324		
	White Female	17	.367	.297		
	White Male	8	-.076	.208		
	ANOVA		.237	.354	6.761	.000

Table D-4 Experiment 2: Supplementary between-group results by participant race and gender