Examining the Utility of the Function Acquisition Speed Test (FAST) for Assessing Social Biases

Thesis submitted to the Department of Psychology, Faculty of Science and Engineering, in fulfilment of the requirements for the degree of Master of Science, Maynooth University.

Matthew Wall

October 2021

Research Supervisor: Dr. Bryan Roche

Head of Department: Dr. Michael Cooke

Table of Contents

Title **Page Number**

List of Tables

List of Figures

List of Equations

Acknowledgements

I would like to express my thanks to my supervisor Dr. Bryan Roche, who was always very generous with his time for me. Your advice and encouragement were integral to the completion of this thesis. It would be impossible for me to fully express my thanks towards you, but for your guidance I am sincerely grateful.

I am deeply thankful for my family, who supported me throughout this endeavour, you went above and beyond in helping me through this process. In particular to my Dad, who taught me the value of perseverance, your advice did not go unheard. To my brother Andrew who first inspired my interest in academia and has indulged me in many a conversation about all things philosophical, thank you.

I would also like to extend my gratitude to the Department of Psychology Faculty, who were always kind to me and expressed interest in my work. In particular, I would like to thank Anne Dooley, who provided me with some wise words of advice and encouragement.

Finally, I would like to thank my friends, who have patiently tolerated many a rant about one subject or another. Without your love and support I would not be where I am today.

Abstract

The current research was focused on assessing the utility of a new behaviour-analytic implicit Function Acquisition Speed Test (FAST), for the proxy measurement of real-world social attitudes. In Experiment 1, the FAST was administered to a sample of men and women to assess the strength of verbal relations (attitudes) regarding gender biases. An explicit measure of attitudes towards gender was also administered as part of a strategy to establish preliminary convergent validity for the FAST. In the domain of gender attitudes, the FAST scores for gender bias converged with those of the explicit measure. However, while male participants self-reported a greater level of gender bias than the females, the cohort as a whole was not found to be gender-biased using the implicit measure, nor were the females when considered alone. This finding was interpreted in terms of System Justification Theory (SJT) as part of a conceptual bridge building exercise between behaviour analysis and mainstream social psychology. The predictions of this theory were also employed to rationalise the need for Experiment 2. Experiment 2 was a replication of Experiment 1 within the context of racial bias amongst a sample of White and Non-White adult participants. The results showed that the cohort as a whole showed a significant implicit pro-White bias, in line with the predictions of SJT, as did the Non-White cohort when considered alone. This provided the FAST with a degree of predictive validity against conceptual frameworks within the literature. In addition, divergent validity was established through the expected lack of correlation between self-reports of racial bias and FAST scores. It was concluded that the FAST may represent an acceptable behaviour-analytic alternative to social cognitive implicit test methods and may be useful in sensitive research contexts in which self-reports are likely to be unreliable.

Chapter 1

General Introduction

1.1 Introduction

With recent advances in the analysis of verbal behaviour it has become possible to examine socially relevant areas of behavior that have long been considered outside of the domain of the behavior-analytic approach (See Dymond & Roche, 2013 for a review). Specifically, in recent years there has been a marked focus on the development of implicit-style tests of verbal behavior within our field (e.g., Barnes-Holmes et al., 2006). These tests provide us with the means to assess the historical verbal practices of individuals, and in so doing, reveal something about their social histories. These tests achieve this, not in a survey style self-report, but by drawing upon processes that index the fluency of particular verbal relations in the individuals' repertoire (Roche et al., 2003). These types of tests hold the promise of allowing us to assess social history without relying on the accuracy of discriminations made by an individual in a self-report, often regarding sensitive subjects such as racism, sexism and so on. For want of a better word these types of tests may be referred to as "implicit" tests, or implicit attitude tests, but as I will outline in this chapter, they are best understood when unhampered by social-cognitive concepts such as attitude, unconsciousness, and other poorly defined mental processes. When the term implicit test is used in this thesis, it is being used purely for conventional reasons, and is not indicative of a particular process or procedure that is different to or requires special explanation not formerly possible in behaviour analysis. Indeed, in the course of exploring how behaviourists can develop better implicit tests than already exist, the concept of implicitness will be unpacked in functional terms.

This thesis will begin by outlining the most popular of the implicit tests – the Implicit Association Test (IAT; Greenwald et al., 1998). This will consist of both an illustration of its uses, and an unpacking of its apparent processes; an exercise which will allow a highlighting

of obvious areas for improvement and translational research. This thesis will then go on to consider the behavior analytic effort to build these types of tests for a broader purpose than measuring "attitudes", but which nevertheless can be used in social research within a functional-behavioral tradition. Perhaps most interestingly, this thesis will outline a research history within behaviour analysis that illustrates that the relevant processes underlying the IAT had already been identified, before the IAT had even been conceived. This will show how, for several years, a preliminary method was explored in the literature that served as precursor to the IAT that was better understood, if cumbersome in its execution (See Watt et al., 1991). Having illustrated conceptual, methodological, and data analytic concerns regarding the IAT, many of which also apply to the more recently developed Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes et al., 2006). This thesis will then describe the application of behavioural methodology in the development of an improved test that is functionally transparent, non-mentalistic, non-proprietary, and atheoretical. Specifically, the Function Acquisition Speed Test (FAST; O'Reilly et al., 2012; Cummins & Roche, 2020) will be outlined as a product of a ground-up research program to develop an "implicit" style test to the functional standards we usually expect within the field of behaviour analysis.

1.2 The Implicit Association Test

1.2.1 Background to the IAT

Before delving into the IAT itself, it might be worthwhile to examine the social-cognitive position that led to the need for an implicit test in the first instance. In this social-cognitive model, "attitudes" are approached as "evaluative judgements" arising from prior experiences, that mediate present behaviour outside of conscious awareness (Greenwald & Banaji, 1995). For example, previously formed evaluative judgements, such as negative

experiences with White people, will influence later behaviour towards White people, even if these mental associations are no longer accessible consciously. Greenwald and Banaji (1995) were keen to distinguish their concept of implicit attitudes from other priming and context effects, with which they had much in common. They clarified that priming and context effects are similar to implicit attitudes, insofar as they both relate to how prior events influence a subject's response to a current stimulus. The key difference being that priming and context effects are specific and operationally defined, while implicit effects are more theory-defined, and necessitate the subject's unawareness of the effect of their prior experience on their present behaviour. To aid in the comprehension of this differentiation, they noted some priming and context effects that *would* fall under the umbrella of implicit cognition. They drew on Thorndike's (1920) halo effect as an example of how a judgement of one aspect of a person can influence later unrelated judgements of that person (e.g., attractiveness positively influences character judgements; Greenwald & Banaji, 1995). They also subsumed Zajonc's (1968) mere exposure effect into the implicit domain (i.e., when an individual is repeatedly exposed to a person, place, or thing they tend to develop a favourable attitude towards it). To put this simply, they proposed that we need not be aware of an attitude for it to influence our behaviour. Though indirect measures of attitude have been used in the past, they were employed to minimize the demand characteristics of an experimental situation, such as evaluation apprehension (a subject's desire to be rated favourably or healthily by the experimenter, thus skewing their response; Rosenberg, 1969). In contrast, Greenwald and Banaji declared that if attitudes can be non-conscious then "indirect measures are theoretically essential" (Greenwald & Banaji, 1995 p. 5)

In keeping with their account of implicit attitudes, what followed was a seminal paper on an indirect measure of their own creation, the IAT (Greenwald et al., 1998). With this, Greenwald et al. sought to overcome their field's dependence on direct measures of attitudes,

and to pioneer a method of circumventing response bias in attitude studies. The need for a test of this sort arose from a crisis of confidence concerning the validity of self-report measures in the prediction of behaviour. Indeed, Ajzen and Fishbein (1977) noted that the lack of correlation between attitude measures and observed behaviour could be seen across a range of studies. The problem with attitude measures, they reasoned, was a lack of specificity. A single instance of an observed behaviour always involves four elements, an *action*, towards a *target*, in a specific *context*, at a specific *time* (Ajzen & Fishbein, 1977). Ajzen and Fishbein's (1977) central thesis was that the strength of an attitude measure in the prediction of behaviour relies on the degree of correspondence between the measure and these four elements. More specifically, consider these two questions asked the day before an election, "What is your opinion of the Fine Gael party", and "Will you be voting for Leo Varadkar in the election tomorrow?". The latter will obviously have greater predictive capacity because it specifies a target (Leo Varadkar), an action (voting), a context (the election) and a time (tomorrow). Intuitively of course, a subject's attitude might have *some* predictive capabilities in that it provides us with broad behavioural probabilities, but it will fail to predict specific actions in specific contexts. In this instance the person from the example may have a positive opinion of the party but would prefer to vote for a different candidate. Such a high level of specificity can be difficult to achieve in an attitude questionnaire, as such, something with more utility was needed.

In pursuit of a measure with more utility, Greenwald et al. (1998) approached attitudes from an entirely different perspective. They proposed that attitudes have different components, and what an individual consciously reports they believe may mask their unconscious "associations". These "associations" between stimuli they claimed, may be a better predictor of behaviour in some situations (e.g., Brunel et al., 2004). To be specific, Greenwald and colleagues (Greenwald & Banaji,1995; Greenwald et al., 1998) delineated the

concept of attitudes into discrete entities. That which an individual reports as their conscious belief or affective state was to be considered an *explicit attitude*. Beliefs outside of conscious awareness, were now to be reconsidered as *implicit attitudes*. Each may therefore bear a different relationship to overt future behaviour because they are not synonymous. This reformulation, it was hoped, would rescue the concept of attitudes from obsolescence due to its lack of utility in the prediction of behaviour. In addition, because they had proposed that implicit attitudes would not be amenable to traditional testing methods, a new test was needed to accompany this conceptual development. The IAT fit this purpose in that it was purported to measure the "unconscious" factor central to the implicit attitude concept. The IAT, as we will soon examine, involves rapid reflexive responding to stimuli, and so should be capable of measuring the implicit attitude construct they defined as, "actions or judgements that are under the control of automatically activated evaluation, without the performer's awareness of that causation" (Greenwald et al., 1998. p. 1464). Intuitively, this test would seem to be a more precise proxy of the target-action elements needed for the prediction of behaviour, while retaining some generalizability across context and time.

In plain terms then, Greenwald and colleagues claimed to have invented a test that represented a sort of Holy Grail for psychology, in that, it could literally read the unconscious minds of test takers. They were quite explicit in this, going so far as to say that it could unearth discriminatory practises even in those who would have no knowledge of their own biases. Greenwald and Banaji (2013) argued that the Race IAT could reveal biases in individuals who honestly described themselves as racially egalitarian, and even predict discriminatory practises from those same individuals. Of course, this entire notion of attitudes as mental associations mediating current behaviour outside of conscious awareness is heavily laden with social-cognitive assumptions. This position is untenable within the behavior analytic tradition in which we eschew behavior-behavior relations as explanations, while

acknowledging that behavior can sometimes function as a stimulus for further responses (Hayes & Brownstein, 1986). To the behaviour analyst, the conception of implicit attitudes given by the IAT creators is problematic, as it suggests that an implicit attitude is an unseen force which can only be measured indirectly. Therefore, functional control over an implicit attitude cannot be attained by researchers attempting to study this construct. As the following quote suggests, this conception of implicit attitudes would, by definition, fall outside of the domain of behaviour analysis.

> Finally, there are times when direct control is impossible in principle. A claim that we have unconscious thoughts and that these produce stimuli might be an example. Here we are using the term "response-produced stimuli" solely to provide a consistent account, but at a considerable cost. We have disguised an analysis that cannot in principle meet all the goals of science from a behavior-analytic viewpoint in the cloak of terminology that suggests these goals can be met. (Hayes & Brownstein, 1986, p. 189)

While the idea that an internal response can function as a stimulus for further responses is consistent with the principles of radical behaviourism, this does not provide a carte blanche for speculation as to the nature of these private responses. In effect, Greenwald and colleagues have invented a private behaviour (an implicit bias) to explain outward visible behaviour. A behavioural explanation for the concept of an implicit bias would find a sturdier foundation by first generating an implicit bias in a laboratory setting, and then testing for its presence, as opposed to mere inference about its properties. Indeed, this was precisely the approach taken by Roche et al., (2003) in their functional account of the IAT, an account which shall be explored in detail later (see p. 38). As we shall explore later in this chapter, it would seem that the cognitive assumptions underlying the concept of an implicit bias can be

safely jettisoned without compromising the utility of implicit tests, thus rendering the domain more amenable to behaviour analysis.

1.2.2 IAT Methodology

Following the case for implicit attitudes made by Greenwald and Banaji (1995), Greenwald et al. (1998) proposed the Implicit Association Test. The novel IAT purported to measure the unconscious associations between a "target-concept" (e.g., African American or European American names), and evaluative stimuli, such as positive and negative words. The IAT works by measuring the speed at which participants can follow a rule to respond in common ways (i.e., a positional computer keyboard press) to stimuli appearing on a computer screen. These stimuli are typically exemplars of either socially compatible or social incompatible concept and attribute classes (e.g., African American names and positive words are considered socially incompatible in the Race IAT) depending on the test phase.

The original IAT (Greenwald et al., 1998) was administered across five blocks. The IAT will be illustrated here using the assessment of racial biases, purely for explanatory reasons. In block 1, subjects complete a target-concept discrimination task by distinguishing between Black names and White names presented sequentially on screen. Subjects discriminate the stimuli by pressing a left-hand key (E) for exemplars from one class, while pressing a right-hand key (I) for exemplars of the other. The rules for how to discriminate between stimuli are presented in advance of the task, and continually at the top of the screen as the task proceeds. Corrective feedback on performance is provided in the form of a red X that appears on screen following an error. Correct responses are not consequated. This procedure simply serves to ensure that the target stimulus classes are already formed in the history of the participant, and that they are not foreign to them when the critical test blocks are presented. In the second block, subjects must discriminate between exemplars from the

attribute stimulus classes, in this case positive and negative evaluative stimulus classes (e.g., happy, evil). This is done in the same manner and for the same reasons. The third block is the first critical test block and uses stimuli from both prior blocks. Specifically, block three combines the previous two blocks in presenting either attribute or target-concept discriminations on alternating trials. In this example, Black names and unpleasant words would share a common key-positional response, as would White names and pleasant words. The fourth block is identical to the first, except that the response requirements for the target attribute are flipped, so that if Black names previously required a left-hand key press, they now require a right-hand key press and vice versa. The final block five is a reversal of block three, so that now Black names and pleasant words share a common key-positional response, as would White names and unpleasant words. The IAT effect is calculated as the difference in average response latency across blocks three and five. This difference is indicative of a pre-existing bias to respond in a common way to exemplars from a particular pair of target and evaluative stimulus classes (e.g., Black names and negative words, and White names and positive words, rather than Black names and positive words, and White names and negative words). Of course, there are a few other procedural details that should be taken note of.

Importantly, if the researcher suspects that participants will have greater ease in assigning some sets of target and attribute concepts together (e.g., in a Racial IAT White-Good, Black-Bad), then the block with these requirements is dubbed the compatible condition. The other condition, where it is expected a participant will have greater difficult in assigning a common response to sets of target and attribute concepts is dubbed the incompatible condition (e.g., White-Bad, Black-Good). The ordering of these blocks is varied between subjects to prevent procedural effects. The IAT format and scoring method has evolved slightly since this initial paper. The IAT now makes use of short practise blocks before the critical test blocks. The use of time-based penalties for incorrect responses has also

been removed. This method and the accompanying changes to how the raw data is analysed were outlined in a later paper (Greenwald et al., 2003). In essence however, the IAT remains substantially similar to this core methodology. While there is no doubt that the IAT has generated substantial research since its initial conception (See Greenwald et al., 2009), that is not to say that the methodology has gone without its critics (See Hofmann et al., 2005; Oswald et al., 2013).

1.3 Critiquing the IAT

1.3.1 Social-Cognitive Critiques

There has been considerable criticism of the IAT even from within its own field and outlining all of these critiques is beyond the scope of this chapter. However, in what follows, a sample of the most noteworthy critiques will be examined, with a particular focus on those that raise concerns to the behaviour analyst.

1.3.2 Implicit or Indirect?

Perhaps of most concern to the behaviour-analyst is the issue of opaqueness that arises from the mentalistic terminology involved in the usual description of the IAT. More specifically, a technically precise definition of the construct IAT researchers refer to as "implicit bias" remains elusive. Indeed, this was the question of interest to one of the more prominent critics of the IAT, Jan De Houwer, who has argued that there is considerable ambiguity around the functional properties of the term "implicit" (De Houwer, 2006). De Houwer (2006) first sought to delineate the difference between a direct and an indirect measure. In his view, a direct measure involves asking the participant to self-assess the property being measured i.e., self-report their feelings or thoughts on a subject. An indirect measure on the other hand, involves inferring an attitude from a behaviour other than the

participants self-assessment. On this basis, the IAT is clearly an indirect measure in the sense that it infers an attitude from task reaction times.

An important clarification should be made at this point, indirect measures can be based on self-report, e.g., on the Minnesota Multiphasic Personality Inventory, the degree to which a participant endorses the statement "I have a good appetite" is used to index their depression (De Houwer, 2006). The participant is not being asked directly to assess their own depression; therefore, the measure remains indirect. However, a measure being indirect does not necessitate that it also be an implicit measure. According to De Houwer (2006) an implicit measure has the functional properties of being uncontrollable, unconscious and automatic. From the aforesaid, we can deduce that the direct/indirect distinction refers to properties of the test procedure, while whether a test is implicit/explicit refers to the outcome of the measurement procedure (De Houwer, 2006). That is to say, a methodologically implicit test would leave the subject unaware of the property of their behaviour under investigation. On the other hand, an indirect test might be methodologically transparent, but nonetheless produce a result not easily altered by a subject. The latter, he argued, is the case for the IAT. That is, while not all subjects discriminate what is being measured in the IAT a sizable portion do (De Houwer, 2006). If then, subjects correctly deduce the purpose of the test, and can reason that their results are a product of their own attitudes, then an implicit bias may not be the unconscious entity that IAT researchers claim it to be. It may very well be the case (and the preponderance of evidence suggests so) that subjects cannot alter or "cheat" an IAT, but nonetheless retain an awareness of their own attitude, and the effect it has on their test score (De Houwer, 2006). If this is the case, the IAT may be better viewed as measuring a more automatic, fluid or context dependent construct, rather than a dynamic unseen "attitude".

1.3.3 Meta-Analysis of IAT Correlates

If the implicit bias construct does not refer to a separate unconscious entity that exists independently of the more "regular" conscious attitude construct, then what is it that the test is measuring? A closer look at the IAT's predictive abilities across a range of criterion measures of discrimination may shed some light on this question. Researchers proposed that that the IAT's superiority over explicit measures had been overstated, and its rates of correlation with criterion variables are actually inferior to explicit measures in several domains (Oswald et al., 2013). Oswald et al. noted several criticisms of the criterion measures of discrimination used by IAT researchers, as well as inconsistent choices in the analysis of data. They argued that IAT researchers treated data from nearly identical studies differently, and used unclear, or possibly ad hoc criteria when selecting effects for inclusion in analysis. Indeed, Oswald et al. noted the cumulative effect of these choices was to obfuscate variance in performance on explicit measures, and emphasise absolute judgements of majority and minority groups, or both. Absolute or binary judgements of the target groups being more conducive to finding correlates with the IAT, as opposed to results that are more ambiguous, or lie in the middle ground between positive/negative evaluations. For this reason, Oswald et al. performed a new meta-analysis of IAT criterion studies. The criterion variables examined were categorised in the following way: brain activity, response time, microbehaviour, interpersonal behaviour, person perception, and policy/political preferences (Oswald et al., 2013). Their meta-analysis focused on the domains of racial/ethnic discrimination. Oswald et al. found that the mean correlations between IAT and criterion scores were .15 and .12 for racial and interethnic behaviour respectively. This was lower in comparison to an earlier analysis performed by IAT researchers which found correlations of .24 and .20 for racial and other intergroup behaviour respectively (Oswald et al., 2013). Additionally, it was concluded that the IAT was inferior to explicit measures across all

criterion variables examined except for neuroimaging studies. Increased brain activity however, in the absence of observable differences in behaviour, has very little bearing on discrimination in any way that might be considered socially meaningful (Oswald et al., 2013). These findings are critical of the idea that the IAT measures an entirely different attitude construct, at least in any observable or significant sense.

1.3.4 Methodological Issues

In addition to questions surrounding the nature of the construct the IAT purports to measure, there remain methodological issues of concern regarding its format. Such issues have been brought to light even by those within the social cognitive community. As will now be examined, perhaps the most concerning of these methodological issues relates to confusion concerning the core process underlying the IAT effect.

In an early but prominent critique of the IAT, Rothermund and Wentura (2004) outlined the fact that despite a high number of studies with good face validity, the core underlying processes of the IAT had not been delineated. They proposed that, the assumption that the IAT measures "associations" between target and attribute categories is less straightforward than originally thought. Specifically, it was their position that at least a portion of the IAT effect could be explained through a different model that rested on an account of salience asymmetry (Rothermund & Wentura, 2004). In their view, it is not an "association" between current experimental stimuli, but the salience of stimuli employed in the test that causes the stimuli to be more easily related via a common response. Rothermund and Wentura (2004) used the classic example of a flower-insect IAT, this version employs insect and flower related words as the target concepts, and pleasant and unpleasant words as attribute concepts. They proposed, that if the insect and the unpleasant word categories are more salient than the flower and pleasant word categories, the former categories will be more

easily related by virtue of their greater salience. In effect then, the degree of semantic "association" may not be the primary process underlying the IAT effect.

To further bulwark the above-stated point, Rothermund and Wentura (2004) pointed to the work of Brendl et al. (2001), who employed a modified insect-flower IAT, which used nonwords in lieu of flowers as one of the target categories. Brendl et al. found a reversed IAT effect. That is, responses were faster when insects and pleasant words required a common response, and neutral nonwords and flowers were assigned to a different common response. This finding is incompatible with the idea that the results of regular flower-insect IATs are a result of insect-negative associations. Of course, this salience asymmetry critique did not go entirely unnoticed by the creators of the IAT.

Greenwald et al. (2005) responded that there was a difference in the definition of the word "association" between their account and that of Rothermund and Wentura's (2004). Greenwald et al. (2005) maintained that they were uncommitted to any particular account of "association" between stimuli. Furthermore, they suggested that while there was disagreement between their account and that of Rothermund and Wentura's, it was not of any empirical concern. The reasoning Greenwald et al. (2005) provided was that an account of the IAT effect as a result of salience asymmetry could not explain the myriad of correlations with other measures that IAT studies have demonstrated. However, given the critiques of Oswald et al. (2013) outlined earlier (i.e., that correlations between the IAT effect and criterion variables were often inferior to explicit measures), this assertion made by Greenwald et al. has lost some credibility. Nonetheless, whichever account of the IAT effect may garnish the most empirical support, the fact remains that confusion over a term as central to social cognition as "association" greatly compromises the IAT researchers account of their own measure, and suggests the need for a more functionally understood account of the effect.

1.3.5 Behavioural Critiques

Criticisms of the IAT have not been limited to researchers within the social-cognitive domain, a number of criticisms from behaviour-analysts have been levelled against the test. The first behavioural model of the IAT (Roche et al., 2005), removed much of the mentalistic jargon shrouding the test. They reconceived it as a measure of a participant's fluencies with the relevant verbal categories employed in the test, and their degree of experience at juxtaposing members of those verbal categories. As part of this reconception, a host of critiques were offered concerning the methodology of the IAT, as well as its statistical analytic method. In the following sections, the main criticisms of the IAT from the behaviour-analytic community will be briefly reviewed.

1.3.6 History of the Use of Reaction Time Measures in Social-Cognitivism

Before we look at these behavioural critiques of the IAT, it may be worthwhile to take a brief look at the history of reaction times within social-cognitivism to better understand some of the core issues with the IAT. In doing so, we should first look at the Stroop task, a prominent test within the social cognitive domain. Although there have been many variations of the test, the original version as conducted by Stroop is the one that shall be described here. In brief, the Stroop task is a test wherein a participant is presented with a series of written colour words and solid colour squares (MacLeod, 1991). This written colour word is depicted in a colour incongruent with the semantic meaning of the word (e.g., the word red depicted in green). The Stroop test relies upon the fact that it takes a greater amount of time to name this colour aloud than it does to simply name the colour of a solid colour square. MacLeod (1991) traces the history of the Stroop task to the work of James Cattell over 100 years ago, a researcher that Stroop considered an influence on his own work. Cattell showed that objects

and colours took longer for a subject to name aloud than the corresponding written word. His reasoning for why this is the case is as follows.

"*This is because, in the case of words and letters, the association between the idea and name has taken place so often that the process has become automatic, whereas in the case of colors and pictures we must by a voluntary effort choose the name*" (Cattell, 1886, p. 65, as cited in Macleod, 1991).

Indeed, Stroop argued a similar point to the above, that the difference in time taken to name the stimulus was a result of differential practise (MacLeod, 1991). On the basis of this logic, in his original experiments with the task, Stroop incorporated a time penalty equal to twice the average response time for every uncorrected error on the task. Stroop himself admitted that this was procedure was entirely arbitrary (MacLeod, 1991). In fact, error-rates on Stroop tasks are rarely examined, except in niche cases reaction times have been preserved as the primary metric of interest. This is done under the presumption that reaction times in some way index mental effort, or at the very least they measure the degree to which something has been overlearned to the point of automaticity. While this assumption implicitly or explicitly is nigh universal within social-cognitivism, it is a less than ideal measure to the behavioural researcher. As we shall see in the following sections, this assumption has carried over to the IAT and perhaps has led to some of its most core problems.

1.3.7 Reaction Time Vs. Accuracy

The IAT's use of reaction times as its chief metric is unquestionably mentalistic, but as pointed out previously it is fitting within the social-cognitive tradition. In behaviour analysis however, reactions times in and of themselves in the absence of response accuracy, are rarely used as an indication of the strength or stability of a behaviour (O' Reilly et al., 2012; but see Binder, 1996). Within the behavioural tradition, fluency is usually indexed in

terms of both speed and accuracy, emphasis is given to speed only after a high level of response accuracy has been achieved. Despite the IAT's apparent fixation on reaction times as its primary metric of interest, it has been argued that it is in fact accuracy that indirectly underpins the scoring system and effect size calculations, thereby leading to obfuscation of core processes even further (O' Reilly et al., 2012). Specifically, consider how reaction time is measured for an incorrect response in the most recent incarnation of the IAT procedure; time is measured from the point of stimulus onset until the correct response is provided. Correct responses are ensured by the provision of negative feedback on screen and the endurance of the trial until the response is correct. Therefore, reaction time measures are inflated by the time taken to correct one's erroneous response, which in Greenwald's research approximates 400ms (Greenwald et al., 2003). In effect, higher response times will be more likely to be found on whichever block generates the most errors. The response times recorded for correct responses only, if considered in isolation from error responses, will reveal a smaller effect. While some researchers have opted for this more conservative measure, the most widely used IAT scoring algorithm does not require it (Greenwald et al., 2003). This state of affairs is far from fortuitous, insofar as the use of the response time inflation method was designed consciously to reflect the assumed extra mental effort involved in a task that results in an error response. This method replaces an older one that involved the addition of an arbitrary time penalty to error responses; a less opaque but equally conceptually questionable procedure (see Gavin et al., 2012; Ridgeway et al., 2010).

1.3.8 The Negative Impact of Punishment, Errors Beget Errors

O'Reilly et al., (2012) speculated that one corollary of the negative feedback procedure may be to generally slow down responses and increase errors on the inconsistent block of the test. Specifically, these authors suggested that the negative feedback presented

following error responses might function as a sort of generalised punisher and is by definition more likely to be encountered during the inconsistent block. The imbalanced nature of this feedback (i.e., no positive feedback is ever presented) may actually serve to exaggerate the IAT effect to the extent that it actually increases error rates on the block in which it is most often encountered. More specifically, rapid responding in the consistent block is negatively reinforced by the absence of feedback interruptions and the punishing effect of negative feedback. By contrast, rapid responding in the inconsistent block is punished, leading to slower responding and potentially more errors as response fluency is disrupted (O' Reilly et al., 2012).

The foregoing idea concerning punishment is based on well understood processes from learning theory. More specifically, response suppression as caused by intermittent punishment in a learning task. One illustration of this from the classic learning literature was developed by Camp et al. (1967), who showed, using rats, that the magnitude of the response suppression effect increases with shorter temporal spacing between response and punishment. They showed that response speed is a function of the interval between response and punishment on previous learning trials, with shorter delays creating larger response pauses on subsequent trials (Camp et al., 1967). The presentation of a red X on screen immediately (no pause) following error responses on an IAT is intended to function ostensibly as a punisher and response correction method (whilst also functioning as a response time elongation method). Given this it is fair to assume that as a punisher it may increase the response times and overall error rate on that block, and this effect will increase exponentially with the frequency of errors. This issue remains to be examined empirically within the context of the IAT, but some intriguing research from the Stroop, suggests that this is an accurate interpretation.

Rabbit and Rodgers (1977) found that on a Stroop task, subjects were more likely to make errors on trials subsequent to an error. This increased error probability lasted for several trials. In short, Rabbit and Rodgers (1977) found that the first trial after an error was likely to produce an 'involuntary error correcting response' (ECR). That is to say, the subject responds in a way that would have been correct on the previous trial. If this ECR happens to be the correct response to the next trial, then this will result in a very rapid response. However, if the ECR is incorrect on the subsequent trial, then this will likely result in an additional error, followed by response hesitancy on subsequent trials. In effect, this provides empirical evidence for the argument made by O' Reilly et al. (2012), that within these types of procedures, errors beget further errors. Insofar as the red X corrective feedback method acts as a punisher (which has yet to be empirically examined), it may well serve to simultaneously increase both error rates and response times.

The extent to which published IAT effects have been inflated by the previously outlined process concerning error correcting responses is not yet known. Interestingly, however, the issue was of sufficient concern to Heathcote et al. (1991) that they proposed that a modification to the Stroop task, namely that response times on trials following an error should not be included in the subsequent analyses. These authors made use of a *dummy trial* method within the Stroop task, whereby responses following an error are discarded, and continue to be, until a correct response is made. The first error trial is then repeated later in the block. Noting that errors produce slower reaction times on the following trial, Heathcote et al. (1991) argued that the dummy trial method is an effective means to eliminate the effect of punishment on Stroop effect sizes. This method improves procedural transparency and preserves the integrity of the measure. Unfortunately, no such effort has been made by the developers of the IAT, although some researchers do voluntarily omit error trials in their

calculation of the D score. Nevertheless, there is no consistency in this practise across research laboratories.

1.3.9 The Use and Abuse of D-Algorithm

A separate but related point is the issue of how reaction times are used to compute the IAT effect. The D-algorithm is the scoring system used for the IAT and it has received little change since its inception (Greenwald et al., 2003). The D-algorithm is a standardised reaction times measure based on mean response times across the whole block. This is contrary to the more traditional behaviour analytic emphasis on response accuracy. Generally speaking, response times are only discriminative of stimulus control when response accuracy has peaked at close to 100% (Binder, 1996). In effect, any changes in response rate across a block of tasks in the IAT are irrelevant to the measure and affect the score only insofar as these changes affect average response times (Greenwald et al., 2003; O' Reilly et al., 2012). Therefore, accelerations in the rate of learning across the block are not taken into account. Thus, the nature of the IAT performance *qua behaviour* is not itself understood nor apparently of interest.

Another curious feature of the procedure employed in the IAT is the fact that the IAT, as a measure of "automatic" responding, requires participants to respond quickly in order to tap into implicit, as opposed to explicit attitudes. However, as important as rapid responding is, it is not controlled in any effective way, and researchers rely merely on pre-experimental instructions to ensure that it occurs. In effect, participants can respond relatively slowly on multiple trials, despite the instructions, so long as their responses do not exceed 10,000ms in latency. Individual trials greater than 10,000ms in length are discarded before analysis (Greenwald et al., 2003). The interaction between reduced response time and the likely resulting increased accuracy is an important dynamic effect not studied in the literature.

Presumably, response latency varies as a direct function of response accuracy, with error rates increasing with speed. Despite its over reliance on mere instruction over trial-by-trial contingencies, the IAT seems to have proved its sensitivity to pre-experimental social contingencies quite well; however, it does not seem like good practise to leave this dynamic unexplored. Understanding this dynamic is key to understanding the behavioural processes at work, and therefore what the IAT is actually measuring. Post-hoc data truncation in which responses over 3000 milliseconds are reduced to 3000 milliseconds (Greenwald et al., 2003), obscure the relationship between response accuracy and latency even further. Of course, these are not the only features of the IAT scoring algorithm which may be of concern to the behaviour analyst.

The IAT scoring algorithm has several quirks which serve to enhance inferential statistical significance levels, rather than enhance stimulus control exerted by the learning tasks within each trial. For example, by eliminating participants for whom 10% of responses are below 300 milliseconds, and eliminating individual trials greater than 10,000ms, data is smoothed and regressed towards the mean. This practice would be more provocative if it were reported in a data cleaning procedure in the results section of each study. By embedding the data cleaning into a scoring algorithm, many participants responses do not achieve the status of "data" in order to be eliminated post-hoc during the analysis stage. While data smoothing is a common practice within experimental psychology, it beholds the researchers to prove that the data is not under stimulus control, rather than eliminate it because it is not supporting experimental hypotheses. In other words, these highly varied responses are produced by the IAT procedure itself, not factors outside of the control of the experimenter. Within experimental psychology, the assumption should be that all responses are within the control of the experimenter. Achieving high levels of stimulus control should be the aim of

the endeavour, rather than the elimination of inconvenient data. Put simply, behavioural variance should be our subject matter, not an inconvenience (Sidman, 1960).

All of the foregoing concerns point to means by which we could improve methods like the IAT so that they are procedurally and conceptually transparent, and the data-analytic methods do not compromise the integrity of the effect underlying the scores. Before we review the development of such a test within the behaviour-analytic tradition (i.e., the Function Acquisition Speed Test: FAST), we should examine some parallel developments that were occurring in behaviour analysis during the development of the IAT that laid the foundation for the FAST itself.

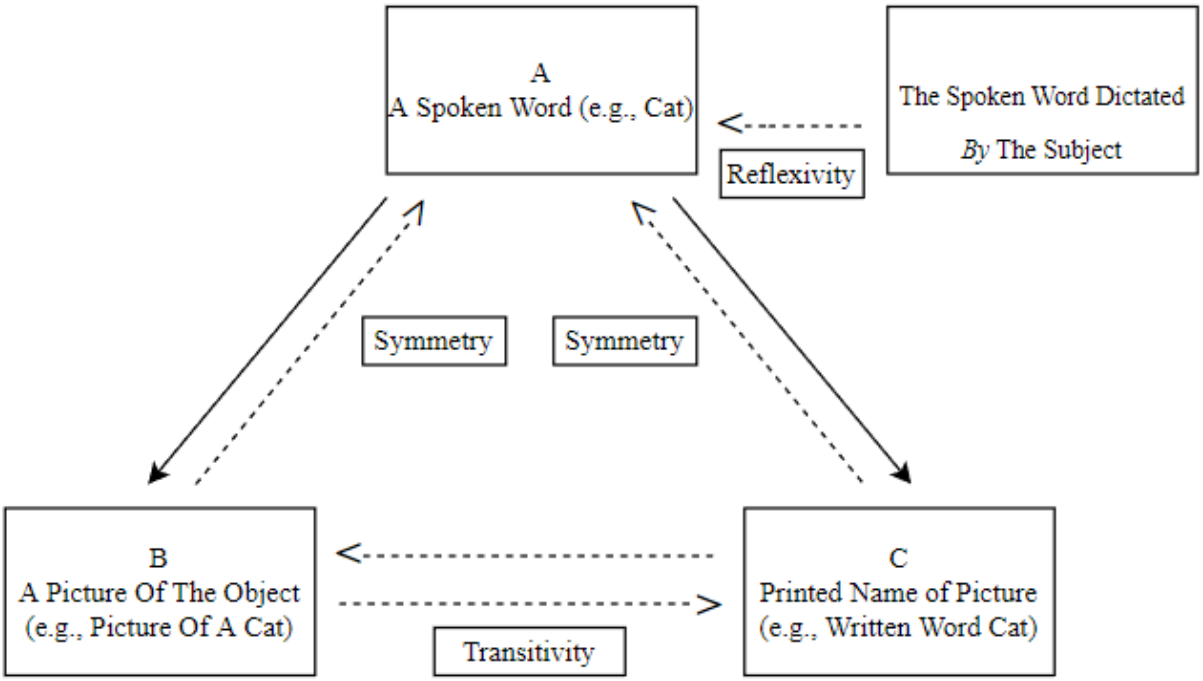
1.4 The Analysis of Verbal Relations

Research into the stimulus equivalence phenomenon has had important implications in educational contexts, particularly for developmentally delayed children, but it also has implications for how we might build more robust implicit tests within behavioural analysis. The stimulus equivalence effect was first outlined by Murray Sidman (1971, 1994), who was conducting research on individuals with deficiencies in reading, writing and speaking. To illustrate, a typical stimulus equivalence procedure will now be outlined. For the purpose of this example let us say that the stimulus items are cat related i.e., a picture of a cat, the written word cat, and the word cat verbally dictated to the subjects. Subjects are first trained using a Matching-to-Sample procedure (MTS) in which subjects are required to respond by selecting one of two comparison stimuli (i.e., B1 a cat picture or B2 a different animal picture) contingent on a sample stimulus (i.e., A1 the word cat dictated to the subject). On other trials subjects are required to respond by selecting one of two further comparison stimuli (i.e., C1 the written word cat or C2 a different written word) contingent on the same sample stimulus (i.e., A1). Using this type of procedure, Sidman was able to show that

verbally-able humans are capable of matching stimuli on a novel task without reinforcement, based on their conditional discrimination training history (See Figure 1). More specifically, when presented with a sample C1, and comparison stimuli B1 or B2, an individual given the training above should be able to pick B1 and not B2 without this relation ever having been explicitly reinforced. More fully, a verbally able human is capable of responding with A1 in the presence of A1 (reflexivity, i.e., verbally repeating the word cat), with A1 in the presence of B1 (symmetry, i.e., verbally saying the word cat in the presence of a cat picture), and with C1 in the presence of B1 (transitivity, i.e., selecting the written word cat in the presence of a cat picture in the absence of this relation ever having been directly taught). Together, these three features, when present, confirm the emergence of a stimulus equivalence relation among the stimuli.

Figure 1

A Basic Equivalence Relation



Note. Solid lines indicate trained relations, the hashed lines indicate the various properties of stimulus equivalence.

Figure 1 demonstrates a practical application of the equivalence paradigm. From this diagram it can be seen that with minimal training many other relations can arise. Despite never being explicitly trained, the B-C and C-B relations can arise merely from a common relation to A. As outlined in Sidman (1994), Sidman conducted multiple experiments into the utility of stimulus equivalence in teaching language skills to children with and without verbal deficiencies. While stimulus equivalence is certainly an important phenomenon in this area, it is not restricted to this domain. The ramifications of Sidman's research were widespread within behaviour analysis, but of particular importance to the current thesis, they laid the groundwork for a behavioural approach to implicit testing. The finding that stimuli can be related to one another in indirect ways is an obvious parallel to the social cognitive notion of implicit attitudes. While in those early days the term implicit was not used, the way in which Watt et al., (1991) employed the stimulus equivalence phenomenon is a clear foundation for all future behavioural approaches to implicit testing.

1.5 The Watt et al. Paradigm

Key findings within the stimulus equivalence paradigm, concerned with the effects of previously established relations on the acquisition of novel equivalence relations, began to emerge concurrent with the explosion of research into the subject of implicit attitudes within the social-cognitive domain. These findings provided key insights into the process that likely underpins the IAT, and lead to the development of similar methods within behavioural research. Indeed, as we will see, several years before the presentation of the IAT, behavioural researchers were hailing these developments as suggesting potential procedures for measuring what social psychologists might refer to as attitudes. More specifically, in what

can now only be considered a seminal study Watt et al. (1991), demonstrated that pre-experimental social learning could disrupt/impede the formation of equivalence classes in the laboratory. As such, the disrupting effect of prior social learning could be indexed quantitatively, and used as a measure of the strength of verbal relations in the history of the participant.

The Watt et al. (1991) experiment took advantage of the evident divisions between social groups in Northern Ireland (Catholics and Protestants), to demonstrate the aforementioned interference effect. A sample of Northern Irish Protestants, Northern Irish Catholics, and English Protestant subjects were recruited for this study. Using an MTS procedure, participants were first trained to select one of three novel nonsense syllables in the presence of Northern Irish Catholic names (A-B relations). Following this, on further trials participants were in turn trained to respond with Protestant symbols in the presence of the previous nonsense syllables (B-C relations). Participants were then tested for the properties of stimulus equivalence (i.e., had the Catholic names and Protestant symbols become related through symmetry and transitivity). Here a Protestant symbol served as the sample stimulus, and Catholic names as comparison choice stimuli. During these test probe trials, novel Protestant names were included among the comparison choice stimuli to function as potential distractors. All five English participants correctly chose a Catholic name when presented with the Protestant symbol, as did seven Northern Irish Catholic participants. However, six Northern Irish Protestants and five Northern Irish Catholics tended to choose a Protestant name, despite the availability of comparison stimuli (i.e., Catholic names) that were transitively related to the sample stimuli via stimulus equivalence. That is, according to the stimulus equivalence paradigm, Catholic names and Protestant Symbols should have become transitively related through the symmetry training of Protestant symbols/Catholic names with the nonsense stimuli. In effect, the social learning of the Northern Irish participants overrode

the experimental contingencies that would normally predict the derivation of an equivalence relation amongst the stimuli. With the publication of these findings, Watt et al. (1991) had created the first rudimentary behaviour-analytic implicit test. They had managed to divine the social histories of participants through indexing the degree to which equivalence performances, using socially sensitive stimuli, were impaired.

At first glance it may appear that the Watt et al. procedure merely measures the existence of previous verbal relations, but not necessarily attitudes per se. However, from a behaviour-analytic perspective, not much more may be required in the functional analysis of attitudes than an account of the stimulus-stimulus relations at work in the verbal repertoire of the individual, and the evaluative response functions established for those stimuli. Evaluative in this sense referring broadly to the type of response the stimulus elicits, i.e., if the stimulus elicits an avoidance or disgust response it could be deemed evaluatively negative. Alternatively, the stimulus may be verbally related to a class of words easily recognised as "bad" or "good". In this sense then, a negative attitude towards vegetables might be conceived simply as a negative evaluative response (e.g., avoidance), of not only vegetables as a category label, but of exemplars from that verbal class (e.g., carrots). Insofar as an individual responds to the category of vegetables in this way, we can say they have a negative attitude towards it.

Of course, it is entirely possible that responses to individual exemplars of a class of stimuli may have functions independent of other members of that same class. In short, a negative attitude can be conceived of as a general pattern of negative evaluative responding to members of a class of stimuli, while acknowledging there may be exceptions to the rule. In the above example, while a subject may display a negative pattern of responding to the vegetable stimulus class, independent exemplars such as carrots may have functions not

consistent with its class membership i.e., the subject may dislike vegetables generally but enjoy carrots. Indeed, this is the approach taken by behaviour-analytic researchers from the outset of this particular foray into attitude research (see Grey & Barnes, 1996; Roche et al., 2002; Gavin et al., 2008). In short, while the term "attitude" may be conceptually problematic, it is sufficient for the behaviour-analyst to think of an attitude as a pattern of positive or negative evaluative responses towards a particular category of stimuli. As seen in the Watt et al. (1991) study, prior learning in the social environment overrode the experimental contingencies, thus it can be said that at a minimum for these participants, Catholic stimuli were incompatible with Protestant stimuli. Given this, it is not too far of a conceptual leap to state that the social contingencies present in Northern Ireland at the time taught participants to categorise Catholics and Protestants in binary ways (at least for this sample). Attitudes then, are not as far out of reach for behaviour-analysts as they may have once been considered. Indeed, it is quite possible to conceive a process of how attitudes form from a behavioural perspective.

According to O'Reilly et al. (2015) attitudes are a product of our culture via a series of verbal events (interactions with others), that are either explicitly reinforced, or a product of untrained derived relations between stimuli. The culture a person is raised in forms a vast web of verbal contingencies that specify relations between stimuli (e.g., rules, norms, mores and taboos). Put simply our culture teaches us a set of relations, what objects are strong or weak, pleasant or unpleasant. Stimulus equivalence allows this small set of relations to expand outwards. This means that even if a person is never explicitly taught a prejudice such as "Black people are inferior to White people" then they could derive this prejudice if they are taught by their environment (verbally, through media etc.) that "Black people are criminals", "criminals are worse than ordinary people" "White people are not criminals" therefore "White people are superior to Black people". The functions of the stimulus

"criminal" were transferred to the related object "Black people", this is essentially the RFT account of attitudes that the FAST is designed around. With this in mind, the Watt et al. (1991) procedure can safely be called a measure of "attitudes" so long as the behaviour-analytic researcher is comfortable with the above-stated account.

It is worth noting at this point however, that racist or sexist behaviour (or "attitudes" more generally) need not arise via verbal stimulus relations. It could arise as a result of classical conditioning e.g., a person is attacked by a member of a particular ethnic group. Alternatively racist behaviour could be a result of operant shaping of behaviour e.g., a child receiving praise for bullying a black person. In these instances, the resulting behaviour is a product of direct contingencies rather than derived relations. Furthermore, the learned response to these direct contingencies may not lead to racist verbal behaviour, and at this stage it is unknown whether the Watt et al. method or other behaviourally oriented implicit tests are sensitive to conditioned or operantly learned behaviour. There is at least some evidence that the IAT is sensitive to classical conditioning as shown by Olson and Fazio (2001). They first exposed participants to pairings between Pokémon characters and positively/negatively valanced words (e.g., excellent, terrible) and positive or negative images (e.g., puppies, cockroaches). On a subsequent IAT a positive bias towards the Pokémon paired with positive stimuli was found. While Olson and Fazio (2001) suggested that classical conditioning is one possible avenue for attitude formation, whether this could be considered an "attitude" is up for debate, as it involves inferring attitude from behaviour rather than a purely descriptive approach. Regardless, the foregoing account suggests that "attitudes" as measured by implicit tests can arise in multiple ways aside from verbal mediation of stimulus relations. Now that we have an acceptable account of attitudes from a behavioural perspective, we can examine some of the applications of the Watt et al. method in light of this reconceptualization.

Approaching it precisely as a promising measure to develop a behaviour-analytic account of social attitudes, the Watt et al. procedure was harnessed in several subsequent studies. For instance, Moxon et al. (1993) showed the procedure's utility in analysing gender stereotypes. Specifically, these researchers found that equivalence relations were relatively more difficult to establish for participants when those predicted emergent relations contained both female names and stereotypically male occupations. In another study, Merwin & Wilson (2005) found the same effect, whereby it was relatively more difficult to establish equivalence relations between self-referential terms and positive terms for individuals scoring low on self-reported measures of self-esteem.

Using the method devised by Watt and colleagues, Roche et al. (2005) conducted preliminary research on its utility in discriminating child sex-offenders and child pornographers, from adult sex-offenders and non-offenders. Although this study was not peer reviewed, it should still demonstrate the utility of the Watt et al. method in principle. Subjects were exposed to series of conditional discriminations on a computer screen designed to teach the following trained stimulus pairs "Child-Tree", "Tree-Lamp", "Cloud-Insect", "Insect-Lollipop". A block of training was composed of 32 trials, wherein each relation (e.g., Child-Tree) was presented eight times. All choices were followed by corrective feedback. Training continued until a subject could produce consistent and correct responding across a block of 32 trials. Control subjects were expected to learn the derived equivalence relations "Child-Lamp" and "Cloud-Lollipop". However, deriving this relation was expected to be significantly more difficult for sex-offenders, due to the socially loaded connotations of the word "lollipop". This word is frequently used among child pornographers and offenders more generally to refer to sexually available children (Roche et al., 2005). Following satisfaction of the learning criterion on the training blocks, subjects were exposed to a block of 32 testing

trials involving four tasks designed to test for the derived equivalence relations (i.e., Child-Lamp, Lamp-Child, Cloud-Lollipop, Lollipop-Cloud). The ordering of the tasks was quasi-random, and each task was presented eight times. Blocks of 32 trials would continue until the subject correctly derived the relation of "Child-Lamp" and "Cloud-Lollipop", or until 12 blocks had been administered, whichever occurred first.

For the non-offender this derived relation should have been trivial to learn, but for the sex-offender, it was likely that they would mistake the equivalence task as a simple matching procedure, in the same way as the Northern Irish subjects did in the Watt et al. (1991) study. That is to say, subjects who are involved in a verbal environment which sexualizes children, are likely to pick "Lollipop" in the presence of "Child" and vice versa. Indeed, none of the four contact offenders against children managed to derive the correct equivalence relation, and three out of four convicted of child-pornography offences failed to derive the relation within the allotted 12 blocks. Of the adult sex-offenders only one out of four failed to derive the relation, similarly only one out of four non-offenders failed to derive the relation over the 12 blocks (Roche et al., 2005).

At this point the reader may be questioning why the Watt et al, method works, is it some special property of the particular attitudes examined in the studies outlined above? Perhaps outlining two studies which looked at a similar process to Watt et al. at a more basic level will make the answer to this question clearer. The work of Tyndall et al. (2004) demonstrates that there is a far more fundamental process at work in the studies using the Watt et al. paradigm than any property of the particular attitudes they examined. This was an expansion on phobia related research undertaken by Plaud, which strongly suggested that equivalence class formation was slower when the stimuli being used were negative/aversive (See Plaud 1995, 1997). Tyndall et al. (2004) first established two functional classes of

stimuli with nonsense syllables. The first functional class was composed of six S+ stimuli (i.e., discriminative stimuli for response emission). The second was composed of six S- stimuli (i.e., responding towards these stimuli was punished, responding away was reinforced). Participants were then exposed to matching to sample training, wherein they were trained to form two three-member equivalence classes. These equivalence classes were composed of four different combinations of the S+/S- stimuli. Subjects required more training when they were required to form two discrete equivalence classes from the six S+ stimuli (i.e., the functionally similar stimuli), than when they were required to do the same for the six S- stimuli (i.e., they did not share a function). It was also more difficult for subjects to form classes which mixed the S+ and S- stimuli, than when the stimuli were separated (i.e., it was easier to form classes containing only either S+ or S- stimuli).

In a related study by Hall et al. (2003), an associative conditioning procedure was used to establish associations between each of two monochromatic geometric shapes (e.g., square, triangle, circle, or star) and one coloured rectangle. In effect, they established two shape-colour associations. Following this training, participants were trained to produce a common key press (left/right) for each of the stimuli when presented alone on screen. Participants in the "consistent" condition were required to make the same key press for stimuli from the same associative relation, whereas in the "inconsistent" condition they were required to produce different responses to members of the same associative pair. Results showed that participants in the consistent condition made fewer errors than those in the inconsistent condition under test conditions without feedback. In effect, these data show that it is difficult to form distinct functional response classes for stimuli from associated pairs of stimuli. This study also points to the same core process identified by Watt et al. (1991) within the stimulus equivalence literature. It can be deduced from the two foregoing studies that the

reason participants in the Watt et al. study often failed to derive a relation between Catholic names and Protestant symbols was because they were from discrete classes of stimuli for those participants. Choosing a Protestant name in the presence of a Protestant symbol was thus easier than deriving the relation that should have occurred. With the Tyndall et al. (2004) and the Hall et al. (2003) accounts in mind, it seemed quite plausible that a more refined "implicit" test could be developed, which would bear more recent advances in behaviour analysis in mind.

1.6 The Functionally Modified IAT

In 2003, Roche et al. proposed a new model of the IAT, based partly on a Relational Frame Theory approach to language in terms of a network of various relations among classes of stimuli. RFT is essentially an expansion of Sidman's account of stimulus equivalence outlined earlier. However, Sidman's account was limited to a purely descriptive account of the processes he observed, it was Sidman's view that stimulus equivalence was itself a basic stimulus function not reducible to more fundamental processes (Sidman, 1994). In contrast RFT deals with a much wider set of stimulus relations e.g., sameness, opposition, hierarchy, and temporality (Stewart & Roche, 2013, p. 63). While Sidman's account dealt purely with a description of stimulus relations, RFT accounts for the ways in which stimuli are framed relationally to one another. Though there are a myriad of possible relational frames, all can be understood in terms of three properties, namely mutual entailment, combinatorial entailment and transfer/transformation of functions.

For purposes of illustration, let us take a hypothetical situation where a child is being taught the value of some foreign currency (Example adapted from Stewart & Roche, 2013, pp. 62-63). If the child is taught a relation such as coin A is worth more than coin B, and they

subsequently derive that coin B is worth less than coin A, then this is the process of mutual entailment (i.e., the B-A relation can be derived from the A-B relation). The child is then presented with a third coin, coin C and taught that this coin is worth less than coin B. If she then demonstrates the derivation of a relations such as coin C is worth less than A, and A is worth more than C, then this is combinatorial entailment (i.e., the A-C/C-A relation can be derived purely from the A-B and B-C relation). Now let us say that the child is taught that coins can be used to buy things (i.e., they have an appetitive function). The child is then presented with a banknote and told that this is worth far more than the coins. Offered a choice between coin A and a banknote, the child will likely choose the banknote, and on the basis of this we can say that the functions of the banknote have been transformed by its relation to the coins. The banknote was previously a slip of paper, a neutral stimulus, but by framing it in relation to the coins, the appetitive functions of the coins were transferred to the banknote. This phenomenon is known as transformation of function. When taken together, these three concepts, mutual entailment, combinatorial entailment and transformation of function form the core of Relational Frame Theory (RFT).

Building on new findings from RFT, Roche et al. (2003) proposed that the IAT effect could be construed simply in terms of a participant's fluency with relating the relevant verbal categories, and their ability to juxtapose members of those categories. Thus, the IAT may not be measuring the extent to which a person endorses a relation between a concept stimulus and an evaluative stimulus, but rather the extent to which those relations have been responded to in their verbal history. Roche et al. (2003) first tested this new model of the IAT in an experiment with laboratory-controlled stimulus relations. To begin with we should look at the arbitrary equivalence relations they established to test with an IAT. It should be remembered that all of the following tasks were administered via computer. In their experiment four

separate three-member equivalence classes were established using 12 nonsense syllables as stimuli (i.e., A1-B1-C1, D1-E1-F1, A2-B2-C2, and D2-E2-F2). An MTS procedure was then administered to link the classes together into two separate superordinate six-member equivalence relations (i.e., A1-B1-C1-D1-E1-F1 and A2-B2-C2-D2-E2-F2). This was achieved by training subjects to match the C stimuli in each class with the D stimuli using a conditional discrimination format. A respondent conditioning procedure was then used to train a colour function to one member of each of the four original three-member equivalence classes. Specifically, a picture of a red blob was paired with presentations of A1 on a computer screen. Using the same process, a blue, green, and purple blob were then associated with D1, A2 and D2 respectively.

Roche et al. (2003) relied upon the phenomenon of transformation of functions to transfer the colour functions to all members of the original equivalence relations. As was outlined earlier, transformation of function refers to a process whereby the psychological functions of an object transfer to another related object, and thus transform the related objects functions. In this instance, it is only necessary to know that when a response function is established for one member of an equivalence relation, the function often spontaneously transfers to all members of the class (e.g., Roche and Barnes, 1997). To take an example from Roche et al. (2005), if a child salivates in the presence of the word “chocolate”, and they are then subsequently told that the Irish word for this is “Seacláid” (i.e., the words are equivalent), then stating this second word should invoke the same salivatory response (despite never being associated with actual chocolate). Effectively then, through transformation of functions, two colours became associated with all members of each six-member superordinate class. Red and blue became associated with A1-B1-C1-D1-E1-F1, and green and purple became associated with A2-B2-C2-D2-E2-F2. The subjects were then

administered a modified IAT which only incorporated two testing blocks. Subjects were required to rapidly respond with a left-hand (Z) or a right-hand (M) keypress to a series of nonsense syllables from the two established superordinate equivalence classes (Roche et al. 2003). Specifically, one of the nonsense syllables would appear on screen and the subject would have to respond with a press of the Z or M key depending on the rules of the block. One block, referred to as a within-class task required pressing the Z key for red or blue, and the M key for green or purple. It was referred to as such, because it required responding consistent with the colour functions trained to the superordinate equivalence classes. The block consisted of four trial types (i.e., C1, F1, C2 or F2 would appear on screen). Each trial was presented 20 times for a total of 80 trials, trials were presented in quasi-random order. A second block was referred to as the across-class task due to the fact that it required the subjects to produce the same functional response (e.g., press the Z key) for members of mutually exclusive superordinate equivalence classes (i.e., press Z for red or blue and press M for blue or green). The same stimuli as the previous block were employed, in the same way, for the same number of trials. Roche et al. (2003) found that subjects tended to respond more accurately on the within-class tasks, providing credence to the argument that the IAT merely measures the fluency of relations, and is thus dependent on a participant's history of relating the stimuli with one another. While this study was an interesting demonstration of how the IAT can be understood in a functional behaviour-analytic way, perhaps a practical application would better illustrate the relevant processes.

In an effort to test the practical utility of their functional model of the IAT, Roche et al. (2005) administered the procedure to a sample of incarcerated contact sex-offenders against children, and a sample of roughly matched non-contact sex-offender prisoners who were convicted of child pornography offences. A control group of non-sex-offender prisoners

were also employed as participants. Their IAT procedure was identical to the Roche et al., (2003) account, except of course for the stimuli employed. In this case the test was designed to ascertain subjects' fluencies in acquiring functional response classes involving exemplars from classes of both sexual verbal stimuli and cartoon images of children, as well as a between cartoon images of adult and negative verbal stimuli. One block of testing required a common response for children and sexual words, and a different response for adults and negative words. The other block flipped these requirements. For the sex-offender groups the former block type should be the easier task due to class consistency. The sex-offenders are already likely to relate children with sexual as opposed to negative words. This block therefore involves within-class responding for these groups. As expected, the general pattern of responding differed between groups. For both contact and non-contact offenders, the majority (3/4) of individuals in both groups produced more correct responses for the within-class task (i.e., relating children with sexual words, and adults with negative words). For the control group however, all four participants produced more correct responses on the "across-class" (for the control group this actually likely involves within-class responding) task i.e., relating adults with sexual words and children with negative words. This was not a peer-reviewed study. However, Roche et al. (2005) suggested that these preliminary results indicated that sex offenders against children can be broadly identified by the ease with which functional response classes can be established involving child-related and sexual terms.

In effect, the findings of Roche et al. (2003, 2005) demonstrated that the IAT is based on juxtaposing within-class responding with across-class responding. For example, on a Race IAT participants with a history of relating Black with bad and White with good will find a block requiring this type of responding easier. By comparing this with another block, where the rules of responding are juxtaposed, the participants verbal history can be deduced from

whichever block they found easier. The reader might recall at this point the argument that was made concerning whether the Watt et al. paradigm was measuring histories of verbal relations, or "attitudes". A similar argument could be made here, when reference is made to within-class responding being easier for a subject to perform (i.e., faster with less errors), what is meant is that the subject's history of verbal relations supports an equivalence relation between the relevant objects. Due to the fact that their history reinforced equivalence between the objects, performing the same functional response for both objects is by definition easier, as a relation of "sameness" has already been taught. An attitude in this sense is merely "a history of both explicitly reinforced relations and untrained derived relations between verbal stimuli" (O' Reilly, Roche and Cartwright, 2015 p.168). The experiments explained hitherto have demonstrated in principle that an attitudes "strength" can be indexed by juxtaposing within and across-class responding and comparing the number of correct responses under these different contingencies. The results of these studies were sufficiently promising to encourage larger scale experimentation with a functional model of the IAT in mind.

In 2008, Gavin et al. further examined the functional-analytic model of the IAT. Because this empirical test of the functional model of the IAT did not include many IAT features, such as the use of negative feedback, or the D-score algorithm, Gavin et al. (2008) referred to their procedure as the Implicit Relational Test (IRT). It did not represent the IAT in format precisely, but it did harness the same core process according to the model. To be exact, their IRT did not include feedback after responding, as a result, unlike the IAT, no time penalty was added for an incorrect response. In order to ensure correct responding, labels in the top left and top right indicated how to respond correctly according to the rules of the block. To guarantee rapid responding, a limited response window of 3000ms was imposed. If a subject did not respond within this window their response was recorded as incorrect, with a

latency of 3000ms, and the next trial began. An additional purpose of the limited response window was to circumvent the arbitrary statistical procedures devised by Greenwald et al. (2003), and thus make scoring less opaque. While Greenwald et al. (2003) advised truncating response times above 3000ms to 3000ms, and response times below 300ms to 300ms, such truncation was not employed in the scoring of the IRT. Instead, a purely descriptive approach was employed, wherein response accuracy and average response time were compared across the consistent and inconsistent block. As a product of the functionally understood, ground-up approach, Gavin et al. employed nonsense syllables as the stimuli in their IRT, in lieu of real words whose relations are not understood. Of course, researchers first had to train the relations in question before testing with the IRT could be conducted.

Specifically, Gavin et al. (2008) administered a respondent word-picture association procedure on a computer using 15 participants. This involved pairing two nonsense syllables "Ler" (A1: in blue font) and "Vek" (A2: in red font), with sexual and aversive imagery respectively. Using a mix of trace and simultaneous conditioning, subjects learned the associations over the course of 10 trials for each of the two word-picture associations. The next phase trained two three-member equivalence relations using a linear training protocol (i.e., A1-B1, B1-C1 and A2-B2, B2-C2). The "Ler" and "Vek" stimuli functioned as A stimuli in this equivalence relation (i.e., A1-B1-C1, A2-B2-C2), other nonsense syllables (e.g., Cug, Paf) were employed as the other stimuli. Response feedback was provided at this stage, and blocks of 16 training trials continued until the subject achieved at least 15 out of 16 correct responses on a block (no subjects failed to meet this criterion). Subjects were then tested for the properties of stimulus equivalence. A block of 16 testing tasks was presented in the following way A1-C1 (C2), A2-C2 (C1), C1-A1 (A2), and C2-A2 (A1), where the stimuli in parentheses indicate incorrect responses. No feedback was presented at this stage, and the

same correct response criterion as the previous stage was imposed (no subjects failed to meet this criterion). Transfer of functions was relied upon to transfer the functions of the "Ler" (A1: blue font, sexual imagery) and "Vek" (A2: red font, aversive imagery) stimuli to all members of the class.

The IRT was then administered. All nonsense syllables were now presented in black font in this test, so stimulus relations depended entirely upon shared and derived stimulus functions. This feature was intended to model natural language categories in which multiple stimulus functions exist across stimuli. Each block consisted of 90 trials involving the presentation of four separate stimulus types: sexual images, aversive images, class 1 stimuli (i.e., A1, B1, C1), or class 2 stimuli (i.e., A2, B2, C2). In the relationally consistent block, subjects were required to respond to sexual images and class 1 stimuli (blue function) with a press of the "Z" key, as well as aversive images and class 2 stimuli (red function) with a press of the "M" key. As mentioned previously, instructions in the top corners of the screen indicated correct responding (e.g., consistent block rules "Press left for Blue and Sexual", and "Press right for Red and Aversive"). Response requirements were juxtaposed in the inconsistent block, so that sexual images and members of the class 2 stimuli shared a common response, and aversive images and members of class 1 stimuli shared a response. The results supported the hypothesis generated by the model. That is 13 out of 15 participants responded with greater accuracy on the consistent task block, and the difference in responding across task blocks was both significant and large. Gavin et al. (2008) concluded that a behavioural model of the Implicit Association Test was a viable ground for proceeding with investigation into the core process underlying the test, and an exploration of different potential applications of these types of procedures. This would also free the IAT from the mentalistic terminology used within the social cognitive paradigm.

The preceding studies have delineated a likely candidate process underlying the IAT. This is based on several well controlled laboratory demonstrations of the IAT effect. However, a higher level of proof for this concept would come from a demonstration not only of the effect created using laboratory trained classes, but also the reversal of the effect resulting from the retraining and reordering of those class structures. This was the subject of analysis for Ridgeway et al. (2010), who sought to replicate and extend the Gavin et al. (2008) study using a similar methodology. In a sample of eight participants and using nonsense syllables as the stimuli, Ridgeway et al. trained and tested for two three-member equivalence classes with a linear training protocol (i.e., A1-B1-C1 and, A2-B2-C2). The next phase used a respondent conditioning procedure to simulate how natural relations acquire affective stimulus functions. Specifically, A1 (printed in blue) and A2 (printed in red), were paired contiguously and contingently with plant and animal photographs respectively. A matching-to-sample procedure was then used to test for acquisition of the respondently conditioned relations. This involved the presentation of the A stimuli as samples on individual trials, along with red and blue coloured "blobs" as the two comparison stimuli. Subjects had to select a blob by clicking on their choice with the computer mouse. Using the same procedure, the C stimuli (in black font) were presented in a test for derived transfer of functions (i.e., in which C1→ blue, and C2 → red). In effect, each equivalence relation was now composed of stimuli that shared a simple stimulus function.

At this point, an IAT with the same modifications employed by Gavin et al. (2008) (i.e., no corrective feedback, a limited response window) was administered to assess the relations between the stimuli within the two stimulus classes. However, the IAT administered by Ridgeway et al. (2010) featured only 40 trials per block. As usual, the purpose of each of the two blocks was to establish two functional response classes, each of which were either

consistent or inconsistent with the configuration of the established equivalence classes. The results were unsurprising, participants produced more correct responses on the consistent block. The novel aspect of the study however, related to the subsequent reorganization of those equivalence classes. Specifically, the subjects were exposed to a second matching-to-sample procedure in which the B-C relations were reversed. That is, emergent equivalence relations were now configured as A1-B1-C2, A2-B2-C1. While this training was counter to that administered previously, training continued to criterion, and until a novel equivalence performance was observed during testing. Only six of the eight participants showed a reversed transfer of function effect which confirmed the existence of new equivalence classes. These six participants were then readministered the modified IAT. Three of these six subjects produced a reversed IAT effect, in which the block formerly defined as consistent now yielded more errors than the block formerly defined as inconsistent. Two of the six participants responded with equal accuracy on both blocks, and one exhibited the same effect as on the initial IAT but to a lesser degree. This outcome further supported the idea that the core process underlying the IAT was one in which past social contingencies are brought into conflict with current reinforcement contingencies in the testing context. This is at least a sufficient explanation for the effect, and many other additional social-cognitive assumptions may therefore be unnecessary. While the prior studies are interesting in principle demonstrations of the behaviour-analytically modified IAT, it might be useful for us to look at some examples of its practical applications. Specifically, several studies have now employed this methodology in measuring subjects' history of sexual associations.

Roche et al. (2012) used the same functionally modified IAT as Gavin et al. (2008) in an experiment into the sexual interests of normal and sex-offender populations. They recruited 60 individuals and split them into 6 groups. This included, 10 female controls, 10

male controls, 10 contact sex-offenders against children, 10 internet offenders convicted of child pornography possession, 10 male sex-offenders against adults and 10 male prisoners convicted of a non-sexual crime. The stimuli used were cartoon images of children and adults, sexual words, and horrible words. In order to ensure familiarity with the stimuli, and that they were categorised correctly, prior to the critical test blocks a categorization test for all stimuli was employed. Thus, this was in effect a three-block modified IAT. In one test block, a common functional response (pressing the Z key) was required for cartoon images of children and sexual terms. A different, but common functional response (pressing the M key) was required for cartoon adults and horrible words. This block was presumed to be class-consistent for the child sex-offenders, as they were presumed to be more likely to categorize children as sexual than horrible. As ever, the response requirements were juxtaposed in the second test block. Exemplars from each of the four stimulus classes were presented 20 times each in quasi-random order across 80 trials in each test block.

Roche et al. (2012) found the largest positive difference score (indicative of a positive child-sexual association) for the child contact offenders, and the second largest for the internet offenders. Perhaps of interest, all groups composed of males showed at least a small positive difference score. The female control subjects however, displayed a resistance to relating the child images with sexual terms (i.e., they displayed higher response accuracy on the second test block). Despite this, a significant positive IAT effect was only observed for the child contact offenders. These results, taken together, indicate that the functionally modified IAT clearly has some utility in differentiating child sex-offenders from other populations.

In a study somewhat similar to the previous one Gavin et al. (2012) revisited the analysis of sexual categorization using the modified IAT. Specifically, 54 subjects from the

general population were exposed to a simple categorization task to test for their ability to differentiate between child and adult terms, and sexual and non-sexual terms. All subjects were then presented with the modified IAT. They were instructed that on each trial, they should press either the red or blue key on the computer keyboard in front of them. These coloured keys were positioned on the left and right of the keyboard respectively. An exemplar from each of the four categories were presented on each trial, and subjects were required to make the appropriate colour key response. On consistent block trials, sexual and adult terms shared a response key, and child and non-sexual terms shared a different response key. On the inconsistent block, adult and non-sexual terms shared a response key, whereas child and sexual terms shared a response key.

Only 12 participants (5 female, 7 male), responded with greater accuracy on the inconsistent block than on the consistent block, although the magnitude of the response accuracy difference between blocks was very different across males and females from this group (2.4 on average for these females, 13.7 for males). This suggests that for these males, there was a history of prior responding that facilitated the formation of a functional response class containing child and sexual terms. No comparable performances were observed for the female participants. Of particular interest was the difference in responding between the remaining male and female participants. Specifically, females demonstrated a significant "IAT effect" insofar as they responded with significantly greater accuracy on the consistent compared to the inconsistent block. However, for the male participants no such effect was observed, with performances across the two blocks being more undifferentiated. The authors concluded that these results illuminate a difference in the way males and females categorize children. The females showed a clear resistance to categorizing children sexually, the males by contrast showed relative ambivalence, and a flexibility across blocks under different

reinforcement contingencies. Importantly, while the males did not as a group show a preference for the formation of child-sexual and adult-non-sexual functional response classes, neither did they show a resistance to doing so on the inconsistent block. This finding has important implications for the identification of child sex-offenders. If the average male shows little to no resistance to the formation of child-sexual relations, then the task of identifying individuals with a history of child-sexual relations using implicit style tests becomes all the more difficult. It is possible, that the employment of words specific to the child sex-offender community as stimuli would have greater utility in the differentiation of these groups (e.g., the use of the term "lollipop" as employed by Roche et al. 2005). This would circumvent the issue of male ambivalence in the formation of child-sexual functional response classes somewhat, as offenders would more readily match terms they have previously used in this context to child related stimuli.

At this point in the research, the IAT had been modified in several ways to be more in line with behaviour analytic reasoning, and given its clear viability as a measure of social history, was beginning to look like an alternative test format in its own right. We can see from the aforementioned studies utilising the functionally modified IAT that the underlying processes of the IAT were now well understood. At this stage, considering the extent of the modifications to the procedure, and the clearer understanding of its underpinnings, it seemed worthwhile to endeavour to produce a new implicit test that would shed the skin of the IAT completely. Before we outline this new Function Acquisition Speed Test, it would seem prudent to look at a related behavioural test which similarly endeavoured to develop a more behaviourally oriented approach to implicit testing. The reason for this being, that the FAST has several notable improvements over the IRAP, which will be clearer if the IRAP's procedure is outlined first.

1.7 The Implicit Relational Assessment Procedure (IRAP)

1.7.1 Background and Procedure

Concurrent to the research developments outlined to this point, a parallel but unrelated procedure was being developed with the same ultimate agenda of providing a behaviour analytically oriented approach to the field of implicit testing. This procedure was referred to as the Implicit Relation Assessment Procedure (IRAP; Barnes-Holmes D. et al., 2006), and also traced its lineages back to the seminal Watt et al. (1991) study. This test, like the FAST, was more consciously inspired by a RFT approach as opposed to a social-cognitivist one. Since its inception, the IRAP has been designed to detect not only the degree of relatedness between stimuli, but also the nature of the relations between them. This allows researchers to identify the direction of a participant's particular preferences, unlike the FAST or IAT which are limited to the identification of binary relationships between verbal classes (i.e., they either exist or do not exist). For example, while the IAT and its variants allows researchers to confirm a preference for White faces over Black faces, it does not allow us to ascertain whether the valence response to White faces was itself positive, or whether, it was instead simply less negative than the response to Black faces. The IRAP is able to achieve a level of specificity beyond that of the IAT/FAST through particular aspects of its testing procedure.

The IRAP involves the presentation of two stimuli simultaneously on screen, with the positive or negative evaluative words on top, the target stimuli in the middle, and the words true and false (or other relational evaluator stimuli) on the bottom left and right. Participants respond with a key press (e.g., E-True, or I-False, with this configuration varied across trials) dependent on the rules of the particular block. For example, Barnes-Holmes D., Barnes-

Holmes Y., Stewart & Boles, (2010) used the IRAP to assess attitudes towards meat and vegetables in a sample of vegetarians and meat eaters. For each trial of the IRAP, the evaluative stimulus “pleasant” or “unpleasant” appeared at the top of the screen, a picture of either a piece of meat or a vegetable appeared in the middle, and the responses true or false were used as relational evaluators. The IRAP was split into eight blocks, two practise blocks and six test blocks, with each block containing 40 trials. Participants are required to achieve a median response latency of 3000ms, and 80% correct responses during practise in order to proceed to the test blocks. Participants who fail to meet these criteria were allowed to repeat the practise blocks up to four times. Blocks alternated between pro-meat and pro-vegetable. More specifically, if, during a pro-vegetable block, a participant responded with “true” in the presence of a picture of meat and the word “pleasant” then a red X appears on screen until the correct “false” response is made. The response time is calculated as the interval between the stimulus presentation and the final correct response, the same method employed in a typical IAT procedure. The next trial then proceeds after a 400ms inter-trial interval (ITI). If, in contrast, a subject responds with “false” on this trial, the trial ends immediately without feedback, and the next trial is presented after the 400ms ITI.

1.7.2 Examples of IRAP Studies

As the first implicit test with an explicitly behavioural background the IRAP has generated considerable interest in the behaviour-analytic community. Its utility has been demonstrated across a range of studies, and its ability to distinguish the directionality of a bias has allowed it to expand over and beyond the limitations of the IAT’s methodology. Cullen et al. (2009) demonstrated this in a study on the malleability of ageist attitudes. Their first experiment used the IRAP to demonstrate that there is a general pro-young bias in the absence of any exemplar training. The trials in this IRAP were structured as follows, a

sample stimulus was presented in the middle top of the screen, a target word was presented in the middle and the response options “similar” or “opposite” were presented in the bottom left or right, each corresponding to a response key. The target words were either positive or negative in valence. The sample stimuli were the terms “young people” and “old people”. The consistent block rules required responding with young-good and old-bad, the inconsistent block flipped these contingencies. The results indicated that mean response time latencies were slower in the inconsistent condition, supporting the hypothesis that there would be a general pro-young bias. This experiment was followed by an investigation into how this general pro-young bias could be altered or even reversed through exemplar training.

In their second experiment, using a different sample, Cullen et al. (2009) divided their subjects into pro-old/anti-young and pro-young/anti-old groups. The former group being exposed to pictures and descriptions of positive older figures and negative younger figures, and the latter group the inverse. Subsequently, as expected, participants IRAP performance was dependent on the type of exemplar training they received. The pro-young group displayed broadly the same pattern of scores as the no-training group. However, the pro-old exemplar condition differentially affected attitudes to old and young, by weakening pro-young relations and strengthening pro-old relations. This was shown by subtracting the mean response latencies for consistent trials from inconsistent trials for each test group. The pro-young group exhibited larger mean latencies (meaning average responding was slower) for inconsistent trials, relative to consistent ones. By contrast the pro-old group exhibited larger mean latencies on consistent block trials relating to old people compared to inconsistent trials of the same type. Despite their training, the pro-old group still had larger mean latencies for inconsistent block trials relating to young people, however the difference between trial types was diminished in comparison to the pro-young group. That is to say, by examining the

latencies of individual trial types it is possible to discern not only a preference towards young or old people but also the nature of this preference. For example, whether a pro-old bias is a result of simply a positive attitude towards old people and a neutral attitude towards young people, or a neutral attitude towards old people but a dislike of young people. This demonstrates that the IRAP permits attitude assessments towards individual concepts, an allowance absent in the IAT methodology which is limited to binary evaluations i.e., the IAT can only discern whether a subject prefers young or old people.

The IRAP has also been used in a similar way to IAT studies, such as the fear of spider's experiment conducted by Nicholson and Barnes-Holmes (2012). This study demonstrated the IRAP's ability to correctly distinguish between two groups of high and low spider fear, as indicated by self-reported fear. A significant correlation between an explicit spider fear measure and D-IRAP scores was found. Additionally, D-IRAP scores correlated with a Behavioural Approach Task (BAT) involving a live Tarantula sealed in a clear plastic container. That is to say, low spider fear according to their IRAP scores predicted subjects' willingness to approach an actual spider.

Like the functionally modified IAT studies outlined earlier, the IRAP has also been employed in an attempt to distinguish sex offenders from the general population. Dawson et al. (2009) used the IRAP in a comparison study of a sample of child sex-offenders and a more general university sample. As an early use of the IRAP in a forensic format, the results were not perfect, a high rate of false positives (43.7%) and negatives (31.2%) was found when the data were examined according to D-IRAP scores. However, in the "child-sexual" trial type, the non-offender group demonstrated a significant IRAP effect (i.e., child-not sexual) while there was a complete absence of this effect in the offender group. This indicated that offenders classified children as both sexual and non-sexual with equal speed. The data

suggests then, that the IRAP has at least some moderate discriminative validity in predicting group membership in the domain of sex-offending.

1.7.3 Meta-Analysis and Reliability

Vahey et al. (2015) conducted a meta-analysis of 15 IRAP studies in order to quantify the degree to which IRAP effects co-vary with clinically relevant criterion variables (e.g., known group differences, self-reports and BAT's). This was done by extracting correlation data from a variety of IRAP studies, where the data was not framed in terms of Pearson's r it was converted to this value. It was the conclusion of these researchers that the IRAP compares favourably with IAT in the clinical domain. Their meta-analysis produced a meta-effect of $\bar{r}$ = .45 (range of .23 → .67; meaning that 95% of IRAP studies in a clinical domain will co-vary with a criterion variable with an effect size within this range). This was in comparison to IAT psychopathology meta-effects of $\bar{r}$ = .22 and $\bar{r}$ = .3, analyses that were conducted with a similar number of contributing effects (Vahey et al., 2015). In summary, according to the studies published up to the point of this analysis, the IRAP is at least as useful a tool as the IAT. It also has clinically relevant applications which the authors suspect may improve further with continuing refinements. We can conclude from this meta-analysis, and the aforementioned studies, that the IRAP has generated considerable data of interest and clearly has applications in some contexts. Furthermore, as the first implicit test designed from an explicitly behavioural perspective it has been received with great intrigue from the behaviour-analytic community.

While the results of the meta-analysis conducted by Vahey et al. seem promising, they do not provide the whole picture on the utility of the IRAP. Notably, issues concerning its reliability as a metric have been raised. Golijani-Moghaddam et al. (2013) examined 9 IRAP

samples from various studies in terms of reliability. The researchers acknowledged that the concept of reliability arises from classical test theory, and thus is inherently based on the idea of underlying attributes (i.e., unobservable constructs), and therefore is not in line with the philosophical assumptions of the IRAP. Ontological assumptions aside, it was their view that an analysis of reliability could nonetheless be fruitful. Golijani-Moghaddam et al. found that across 9 IRAP samples, the internal reliability reached just 0.653, which is short of the recommended minimum of 0.7 suggested by Nunally (1978). They noted however that the two IRAP samples in their analysis which elected to impose a shorter response window of just 2000ms (as opposed to the regular 3000ms) were of acceptable reliability.

In a more general analysis of a variety of implicit measures Greenwald and Lai (2020) concluded that the reliability of the IRAP was 0.6, providing further evidence that the IRAP does not reach acceptable internal consistency. While this is a concerning finding for IRAP researchers, it should be noted that its reliability is comparable with most other measures of implicit bias e.g., Go/No-Go Association Task (0.66) or the Extrinsic Affective Simon Task (0.38) (Greenwald & Lai, 2020). The IAT on the other hand, according to Greenwald & Lai (2020), has a reliability of 0.8. The consensus across these two analyses seems to conclude that the IRAP has an unacceptable reliability, though of a comparable level to other implicit measures. It would seem that at least in terms of reliability the IAT remains the gold standard in implicit testing. With that said, many of the same issues encountered with the IAT have risen once more in the development, application and analysis of the IRAP, including the issue of whether or not an IRAP performance can be faked.

1.7.4 Fakability, Fact or Fiction?

Drake et al. (2016) elected to examine the reliability and fakability of an idiographic (i.e., personalised for the test-taker) variation of the IRAP. Specifically, the IRAP employed here was loaded with the names of two people, one that participants felt had a very positive effect on their life and one who had a very negative effect. The evaluative stimuli employed were positive and negative words. Before being presented with the IRAP participants judged these words on their valence according to a Likert scale, the resulting Cronbach's alpha indicated that the words in each set were closely related. This idiographic IRAP was administered according to the rationale that due to the strong personal relevance of the task it would likely be more difficult for participants to alter their performance. The standard IRAP procedure was employed and was composed of two blocks, a pro-friend/anti-enemy block, and a pro-enemy/anti-friend block. All participants completed three IRAP's, however some participants were given faking instructions before one or more of their IRAP's depending on which condition they were assigned to. Participants were randomly assigned to one of three conditions, the real condition in which no faking instructions were given, the real-real-fake condition and the real-fake-fake condition. Faking instructions were given to participants before the blocks they were tasked to fake. The results of these IRAP's indicated that the D scores for IRAP's administered without faking instructions were substantially biased towards pro-friend/anti-enemy responding. By contrast, there was a dramatic reversal in scores for IRAP's administered after faking instructions, with D scores now skewed strongly in the pro-enemy/anti-friend direction. These results indicated that with the provision of relatively simplistic instructions the IRAP is susceptible to being faked. The authors suggested that the IRAP's fakability may be reduced by the shortening of response windows. It could be argued however, that because the IRAP already suffers from high participant attrition, any further

shortening of the response window could lead to difficulties in general administration of the task. Nonetheless, it remains unknown to what extent its fakability may have influenced results in other experiments.

While not a direct response to Drake et al. (2016), the issue of faking has been examined by IRAP researchers, specifically Hughes et al. (2016) sought to answer precisely what instructions were necessary for faking to succeed. To this end Hughes et al. conducted four experiments, each involving the presentation of two IRAP's. In all experiments, the first IRAP presented was a baseline control. In the first experiment, before the second IRAP, half of the participants were instructed to fake their responses, however no detailed instructions as to how to do so were provided. These participants did not manage to alter their IRAP scores. Experiment 2 was quite similar to the first, bar the fact that for the non-control participants more detailed faking instructions were provided before the second IRAP. These instructions were still relatively vague however, only alerting subjects to pay attention to how they respond in order to develop their own strategy. These participants successfully managed to alter their scores, but only to a more neutral point than their initial IRAP, rather than a complete reversal. The third experiment provided faking instructions for all participants before their second IRAP. Half of these received the same instructions as experiment two (i.e., asked to generate their own faking strategy) and the other half were provided with far more detailed instructions. Specifically, they were told which blocks to respond quickly on and which blocks to respond slowly on, while still maintaining the necessary latency/accuracy criteria. Before each block the relevant instruction for that block was reiterated. As in the previous experiment, participants instructed to generate their own strategy managed to neutralise their scores. The participants provided with detailed instructions however managed to reverse their scores entirely. The final experiment involved

a roughly equal sample of homosexual and heterosexual men, and the IRAP was coded to measure sexual preferences for men vs. women. Before their second IRAP, each group was instructed to behave as if they were their counterparts e.g., heterosexual men should try to respond as if they were homosexual. Specific instructions were provided on how to respond depending on trial type, therefore these instructions were even more nuanced than in experiment 3. The heterosexual men were instructed to respond slowly on Men-Attractive and Women-attractive trial types in one block, and in the other block to respond slowly to Women-Attractive trials and quickly to Men-Attractive trials. Subsequent analysis revealed that the homosexual participants attenuated their scores on the Men-Attractive trial types and increased their scores on the Women-Attractive trials, the heterosexual participants also successfully reversed their pattern of responding.

It was the opinion of Hughes et al. (2016) that the faking instructions provided in experiment 3 and 4 were non-naturalistic (i.e., highly unlikely to occur without instruction or prior knowledge). The faking performed in experiment 2 could arise naturalistically then, but as noted previously, those participants only achieved a neutralisation of their scores, rather than a complete reversal. It was the conclusion of these researchers therefore, that while the IRAP was fakable in principle, it was only possible with detailed instructions, and was therefore no different in terms of fakability to the IAT. It seems reasonable to suggest that the degree of instruction provided by Drake et al. (2016) to their participants likely falls into the same category as experiment 3. Specifically, Drake et al. informed participants of the IRAP's rationale, instructed them to respond slower on one block, and encouraged them to rehearse their answers before another block. If we accept the suggestion of Hughes et al., then it would seem reasonable to disregard the results of Drake et al., on the basis that though the IRAP can be faked in theory, such detailed instructions will never arise naturalistically. Even if this

point is conceded however, the results of Hughes et al. experiment 2 indicate that with minimal instruction it is quite possible to neutralise an IRAP score. Furthermore, it is interesting that the debate has been framed in this way, that a complete neutralisation of the IRAP effect with minimal instruction is considered almost irrelevant. Even if the assertion of Hughes et al. is correct, that reversal of IRAP effects is impossible without detailed instruction, why has that been deemed the ideal criterion for fakability? Despite not being reversed completely, their scores still changed significantly with only the vague instruction to develop their own strategy, and to "Pay attention to how you respond during the task in order to figure out how you can fake it" (Hughes et al., 2016 p. 638). Functionally speaking, their scores were successfully faked, to the extent that they "… successfully eliminated all traces of their beliefs…" (p.639). While, as ever, further research should be conducted on the subject, the results of Hughes et al. should be concerning to IRAP researchers, particularly as the format has not changed since this study and IRAP continues to proliferate.

1.7.5 IRAP Specificity

Before we examine the apparent methodological similarities between the IAT and IRAP it seems beneficial to first address the core difference between the two procedures, and whether the IRAP is in fact an improvement over the traditional procedure. Hitherto it has been argued that the IRAP is likely as susceptible to faking as the IAT, and that broadly speaking it seems to measure the same construct. However, the claim that the IRAP can measure implicit attitudes with greater specificity has gone unexamined. In the following section it will be argued that the IRAP's block structure which purportedly improves specificity may actually be susceptible to a framing bias.

O’ Shea et al. (2016) were interested in the claim that the IRAP can measure absolute as opposed to relative implicit attitudes. That is to say, an IAT can only assess a preference for one stimulus category over the other (relative). An IRAP on the other hand can purportedly assess whether this preference is a result of a neutral disposition towards one category and a dislike of the other, a dislike of one category and a liking of the other, or any variation thereof. O’ Shea et al. (2016) noted that past IRAP studies had turned up unusual findings such as implicit assessments of people with guns as “safe”, or perhaps even more strangely an implicit pro-death bias among a normative sample of participants. O’Shea et al. hypothesised that such counterintuitive findings may have arisen as a results of a positive framing bias (PFB), which could have implications for the interpretation of all IRAPs. In support of this argument, the researchers pointed to psycholinguistic research, which indicates there is a bias inherent in the English language towards describing things in increasing as opposed to decreasing terms e.g., describing a person who has gained weight as “fatter” rather than “less thin”. On this basis, O’Shea et al. argued that it may be easier for participants to select true when presented with positive descriptions of a stimulus than to respond with false.

In order to test their hypothesis O’ Shea et al. (2016) split their participants into three conditions, a standard, positive and a negative framing condition. Each participant across all conditions was administered four IRAPS on different stimulus categories (Nature, Weight, Social System, Nonword). It was expected that participants would have varying degrees of familiarity with each stimulus category e.g., the weight IRAP was expected to have stronger prior associations relative to the social systems IRAP. In the standard condition (using the weight IRAP as an example) participants would be presented with the following instructions before a thin positive block “On this block please respond as if Thin Person is positive and

Fat Person is negative" (O' Shea et al. 2016 p. 162). On a fat positive block, they would be presented with these instructions "On this block please respond as if Thin person is negative and Fat Person is positive" (O' Shea et al. 2016 p. 162). Importantly instructions were counterbalanced across participants so that roughly half received instructions with "Fat Person" as the first stimulus category in each sentence. In the positive condition however, the stimulus instructions were alternated across each block, so that the positive frame was emphasised rather than being counterbalanced between participants (i.e., whichever stimulus category was to be paired with positive words would appear first in the sentence). The negative framing condition was structured in the same way, except that the instructions were reversed so that the negative frame was emphasised, as in the second quote above.

In the standard condition, it was shown that participants were faster to respond with true when pairing any category with positive words. O' Shea et al. administered multiple IRAP's and found that this was particularly true of the IRAPs where participants were expected to have less familiarity with the relevant categories i.e., nonwords and social systems. It was hypothesised that this was the result of a cognitive heuristic employed by participants, wherein they selected only one of the possible stimulus pairings they had to perform in a block (usually the positive) and based all further responding on that pairing. The cumulative effect this has on scoring is to inflate positive evaluations of stimulus items (particularly those unfamiliar to the participant), to the point where it would appear they have a positive evaluation, even where their true attitudes might be neutral or even negative. O' Shea et al. also demonstrated a robust framing effect for the instructions presented to participants, positive framing instructions elevated absolute attitudes towards the stimulus item, and negative instructions decreased absolute attitudes. This was particularly true for the Nonwords and Social System IRAP's indicating framing effects are more profound in

estimates of items unfamiliar to the subject. O' Shea et al. (2016) note that to date, few IRAP studies have clearly delineated the way the task is presented to subjects, in spite of the fact that this could have a profound effect on their results. Finally, the researchers were curious as to whether these framing effects were preserved when relative attitudes were extracted from the data (i.e., the conventional IAT scoring method). The results of this method were much more in line with previous IAT studies, the expected pro-thin/anti-fat bias was found, as was the pro-flower/anti-insect, and the results for the stimulus categories expected to be unfamiliar to participants were relatively neutral. In short, O' Shea et al. suggest that participants will usually be faster to respond with *true* rather than *false* when asked to pair any category of stimulus items with positive words, this finding may have profound effects on the observed data. This and the other issues with the IRAP explored here are both important to the interpretation of IRAP data generally, and the factors necessary to build a superior implicit test.

1.7.6 IRAP Concluding Remarks

The development of the IRAP was a gallant effort to draw upon solid relational frame theory research into derived relational responding in the conception of a novel test format that would be more exhaustive and nuanced than the IAT. However, the IRAP did not sufficiently untether itself from many traditional procedural aspects inherent in the IAT, themselves inherited from the social cognitive tradition dating all the way back to Stroop (1935). Specifically, the IRAP suffers from unacceptable data attrition due to the elimination of participants who fail to satisfy the practice criteria, and further participants who failed to satisfy the response criteria identified post-hoc. This is a curious problem from a behaviour-analytic point of view, insofar as behavioural variation is our very subject matter, and is not viewed as a confound as it is in the social-cognitive tradition. In addition, the IRAP takes

considerable time to administer, upwards of 30 minutes, requiring enormous experimental effort and raising concerns about expediency in applied research. Let us now examine a few of the more noteworthy methodological concerns a behaviour analytic research might have with the IRAP procedure

With regard to data-analytic techniques, the IRAP adopted an almost identical scoring algorithm to the IAT, more typical of cognitive psychology and group design experiments. Specifically, the IRAP uses only response times rather than accuracy or fluency of performances on each block of the test as its dependent measure. The IRAP also normalises data by calculating z scores after the trimming of outlying data. Within the behaviour-analytic tradition, it would typically be more appropriate to achieve data stability through modification of the procedure rather than through post-hoc data cleaning methods. The use of data cleaning methods could indicate that insufficient control is being applied at the front end of the procedure. While it is always preferable to have as much control over participants behaviour as possible, it is acknowledged that data-cleaning may be necessary where significant lapses in participant attention is apparent. Furthermore, the IRAP uses negative feedback alone, in what can only be described as an imbalanced learning procedure, a factor which may also contribute to abnormalities in participant responding and thus further usage of data-cleaning methods.

It is curious that any behaviour-analytic oriented learning procedure would involve only the punishment of incorrect responses, and not involve the reinforcement of correct responses. But there is good reason for this within the social-cognitive tradition. Specifically, this method was adopted because it allows researchers to elongate the response times on error trials, which are typically on the more difficult “inconsistent” blocks, and thereby ensure sufficiently longer response times on those blocks. For a behaviour-analytically oriented test

however, such response time elongation on the inconsistent block task should be a function of improved stimulus control, rather than the addition of artificial time penalties embedded in a response correction procedure borrowed directly and in whole cloth from the IAT. While the use of practise blocks does serve to stabilize data before the critical test blocks, in which differences in response speed are analysed, it is then curious that the practise blocks are not themselves treated as a data source. Rather, these data are eliminated from analysis, whereas in fact the process of interest is at work in those very practise blocks. What is being observed in the key test blocks that are administered later is a reduced difference in response speed across the blocks of the test. Such reduced effects are desirable however, if the objective of the analysis is to reach critical alpha levels in inferential statistical tests, but it is detrimental to effect sizes. Again, this strategy is borrowed from the social-cognitive literature in which p-values are preferred over effect sizes, as outlined most clearly by Greenwald et al. (2003) in their exploration of optimal scoring algorithms.

While it is not in any way experimentally significant, it is interesting that the IRAP is the first computer-based learning procedure developed within modern behaviour analysis that has deviated from the use of the Z and M keys on computer keyboards as an operanda for positional responding, and instead employs the E and I keys as is typical in the IAT. One can only surmise from this and several other features common to the IAT, that the IRAP consists of a post-hoc functional-analytic interpretation, albeit extensive repackaging of the IAT, rather than a novel procedure developed in ground up research. Indeed, there is not a single study using the IRAP in which the effect is generated *ab initio* using laboratory-created stimulus classes to demonstrate the core process, without the confound of the use of stimuli created outside the laboratory in real world settings. It is fair to conclude, therefore, that it is premature to hail the IRAP as a behaviour-analytic alternative to the IAT, as coherent as its

underlying position is presented through the Relational Elaboration and Coherence (REC) model (Barnes-Holmes D et al., 2010) and the Multi-Dimensional Multi-Level (MDML) framework (Barnes-Holmes D et al., 2017). That is to say, despite its supposed basis in these theoretical models, there are enough deviations from standard behaviour-analytic procedure to question the claim that this is a behaviour-analytic test in its own right, and not a rebranding of the IAT with its own extensions and functional modifications.

While it is true that IRAP researchers have appealed to the Watt et al. method as the intellectual progenitor to their testing method (Barnes-Holmes et al. 2008), it is not clear that this is actually the case. As Barnes-Holmes et al. (2008) rightly point out, the Watt et al. method utilised the fact that where stimulus equivalence is expected to emerge from a series of conditional discriminations, it may not emerge if the expected equivalence is in contradiction with socially learned verbal relations. It is well and good to claim that this method is a precursor to something like the IRAP, as both methods operate on the assumption that it will be more difficult to learn a relation that goes against pre-experimental learning. However, it does not follow from this that the IRAP is methodologically distinct from the IAT. The IRAP did not build off the Watt et al. method incrementally as would be expected in the field of behaviour analysis, rather it lifted elements of its procedure directly from the IAT. This has been outlined in the previous sections, but it is worth briefly reiterating. Firstly, the IRAP employs a punisher when an incorrect response is given, but no reinforcer for a correct response. Secondly IRAP data is normalised before analysis, despite the more prudent approach being to achieve tighter experimental control in the procedure itself. Finally, D-IRAP scores are calculated using the same scoring methods suggested by Greenwald et al. (2003).

None of the above were facets of the Watt et al. method, that method employed both reinforcement and punishment, and did not analyse their data using the complex methods employed in a standard IRAP. Given that the IRAP methodology and data analytic methods are far more in line with a cognitive approach, it would be more intellectually honest to state that while it may claim behavioural theoretical underpinnings this is simply not the case. On the basis of these facts, it seems reasonable to assert that the IRAP is more akin to an IAT offshoot than a behavioural approach to implicit testing in its own right. While speculative, it could be the case that the methodology and accompanying data analytic methods employed in an IRAP are similar to the IAT for reasons of publication. Its scoring methods are readily understood by IAT researchers because they are so similar, thus allowing IRAP research to reach a much wider audience. It is at this point that the reasoning behind the development of the Function Acquisition Speed Test becomes clear. In contrast to the IRAP it does not borrow major elements from the IAT without reason. While bearing a superficial similarity to the IAT, all of its procedural and scoring methods have been developed within a behavioural paradigm, and as such are understood in a functional fashion.

1.8 The Function Acquisition Speed Test

1.8.1 The FAST Methodology and Underlying Principles

Building on the functional modifications to the IAT outlined earlier and aiming to avoid some of the pitfalls in the design of the IRAP, it was thought that the time was right to develop an "implicit" test based entirely on behavioural principles. The Function Acquisition Speed Test (FAST; O'Reilly et al., 2012; O'Reilly et al., 2013; Cummins & Roche, 2020) is a new "implicit" test format that has arisen out of several replications and extensions of the Watt et al. (1991) procedure. The FAST's theoretical underpinnings also draw from the

effects reported by Plaud et al., (1995, 1997, 1998), and extended by Tyndall et al. (2004) concerning factors which inhibit equivalence class formation. The FAST is the logical progression of earlier work conducted using functionally modified IAT's, which stripped the IAT of superfluous aspects of its procedure in an effort to make it compatible with the functional analytic approach. The latter effort also involved a conscious effort to avoid replicating mistakes inherited by the IRAP from the IAT format.

The FAST has been developed from the ground up to provide both a behavioural alternative to the IAT, and a procedure more easily administered than the IRAP. In format, the FAST bears considerable resemblance to the IAT but employs methodological features that have evolved across numerous iterations. In published research every aspect of the FAST procedure has at the very least been commented upon or deconstructed conceptually, and most have been studied experimentally. Not all of that work is in peer-reviewed journals, but much of it is available in publicly accessible PHD theses (See Cartwright, 2013; Cummins, 2017; Gavin, 2008; Lalor, 2019; O'Reilly, 2012). With that said, let us now examine the methodological features that compose the FAST test format.

In the first incarnation of this test format as it appeared under the FAST moniker, the FAST consisted of four blocks of training, two of which served as baseline blocks involving test-irrelevant stimuli, and the latter two of which consisted of the critical test blocks. The purpose of the baseline blocks will be returned to shortly, but it makes more sense to first outline the nature of the critical test blocks. Like the IAT, the FAST uses four categories of stimuli in the test blocks; two of which are categorical (e.g., racial) and two of which are evaluative (e.g., good, bad). Participants are required, under time pressure created by a 3s response window (rather than mere instruction), to respond quickly in a common way (e.g., a press of the Z key), to exemplar stimuli from two categories (e.g., images of Black faces and

positive words), and to respond in a different way (e.g., a press of the M key), to words and images from two further categories (e.g., image of White faces and negative words). In effect, the procedure aims to establish two functional response classes containing members that in one block are socially compatible, and in another block are socially incompatible, or are suspected to be such. The test is in fact a learning task rather than an assessment in the traditional sense, and the manner of its presentation and the index of all effects are all in line with this learning approach, rather than an approach based on classical test theory (e.g., which alludes to true underlying effects etc.). Stimuli are presented on screen for 50 trials in succession and appear in the absence of instructions on screen. A rule is not provided at any point regarding how participants should respond, therefore responding is shaped by feedback alone. Feedback (i.e., the words CORRECT or WRONG) appear on the screen briefly after every positional response made by the participant. A second block of 50 trials is then presented, in which the response requirements are juxtaposed (e.g., Black faces and bad words share a response, and White faces and good words share a response).

As mentioned previously, baseline blocks were commonly used in the first iteration of the FAST. These two baseline blocks were identical in structure to the critical blocks, and were identical to each other, but involved test irrelevant stimuli that were low in emotional valence and were presumed to be unrelated to each other in any way (e.g., mushrooms, cars, clouds, furniture). The purpose of these blocks was to provide a minimum amount of practise with the procedure, but more importantly to establish a baseline level of functional response class acquisition rate that could be compared to the acquisition rates on the critical test blocks (O'Reilly et al., 2013). The idea was that any differences in functional response class acquisition rates across the two critical blocks could be moderated by acquisition rates typical

for any random stimulus set. A baseline block was presented before and after the critical test blocks in order to assess the stability of baseline rates of function acquisition across time.

Unpublished research showed that acquisition rates for completely neutral stimuli are in fact slower than for salient and emotionally evocative stimuli. With hindsight, this is not a surprising finding, given that studies conducted in the development of the FAST procedure (e.g., Tyndall et al., 2004) had found that stimuli with recently established response functions, formed equivalence classes more readily than stimuli with no established response functions. In addition, parallel research had shown that nonsense word stimuli that are pronounceable form equivalence relations more readily than stimuli that are unpronounceable (Mandell & Sheen, 1994). Research has also shown that stimulus classes are formed more readily among salient and meaningful stimuli (Fields et al., 2012). In effect, it has become apparent that baseline blocks using neutral stimuli are not appropriate to use in the calculation of effects against which to index acquisition rates in the critical test blocks. As a result, the practice of including baseline blocks has ceased. Several of the studies outlined below employed baseline blocks, in general this should not factor into how they are interpreted, the reader need only know that later versions of the FAST no longer include them.

It may also interest the reader to know that the FAST draws on the concept of behavioural momentum (Nevin & Grace, 2000). Put simply, if a high-rate stable behaviour is disrupted by a change in contingency, it is likely that responding will continue along the original trajectory, rather than give way to the new contingency. To put this in context, if the contingencies of a FAST block are inconsistent with a subject's pre-experimental verbal history, then the subject is likely to make more errors and respond more slowly on that block. Inversely, if the contingencies are consistent with their verbal history, they will respond more rapidly with fewer errors. While the concept of behavioural momentum may or may not

appeal to various readers, we can also simply consider the effect in terms of resistance to change. Overlearned behaviour will display resistance to change as a function of the degree of overlearning. Now that we understand the theoretical principles underlying the FAST, some further words on its methodology and scoring and how it differs from the IAT and IRAP in this respect is warranted.

While the FAST owes some theoretical debt to the concept of behavioural momentum, the developers of the IRAP have sought to explain their test in a somewhat different way through the REC and later the MDML model. According to these theories, the reason for the divergence between the results of implicit and explicit tests in some domains is a result of the temporal factor. According to this conceptualisation "when a stimulus is encountered, a relational response may occur relatively quickly and be followed by additional relational responses. These additional relational responses may occur toward the stimulus itself or toward the initial response to that stimulus" (Hughes and Barnes-Holmes 2012, p 102). That is, under the time pressure of implicit tests subjects will emit a brief immediate relational response (BIRR), whereas given the extended amount of time of an explicit test they will emit extended and elaborated relational responses (EERR). For example, when encountering someone from a different ethnic group a person might have an automatic negative evaluation (BIRR), however this response may serve as an impetus for a further response "Only bad people judge others based on skin colour" (EERR). This individual likely does not consider themselves a bad person, and so on an explicit test would likely respond in an egalitarian way. Under time pressure however, their initial BIRR is more likely to be captured as there is no room to emit an EERR. While this is an interesting behavioural conceptualisation of implicit attitudes, it does not seem to add anything over and above a behavioural momentum conceptualization.

The concept of a BIRR seems does not appear to be significantly different from behavioural momentum, it can be understood as momentum applied to relational responses in terms of a test where responses are required to be brief and immediate. As for the EERR, in the absence of a time restraint verbal contingencies will operate in different ways where multiple sources of control are applied over the same response. If some of those sources of control take different amounts of time to generate responses, then different response windows will produce different responses. As resistance to change is an integral part of behavioural momentum (Nevin & Grace, 2000), it follows that where the required response contradicts a subject's reinforcement history, they are likely to respond slower and with more errors. In the case of implicit testing, the immediate response is likely the most fluent as a longer response window allows for more complex responses which are influenced by social desirability. Responses under multiple sources of control such as those influenced by social desirability are likely to be less fluent, and thus won't be produced under a short response window. The fact that a short response window is enforced in implicit testing is to ensure that the most fluent response is observed, and not the socially desirable response which is context dependent. In effect, the time limit serves as a sieve which prevents socially desirable responses from passing through. As it does not seem apparent that the models offered by IRAP researchers offer anything over and above the behavioural momentum conceptual case it would seem prudent to be more conservative in this aspect. In short, until it becomes clear that the REC and MDML models offer something that behavioural momentum cannot, FAST research shall continue to appeal to the latter rather than the former. It would now seem timely to explore other ways in which the FAST differs from more standard implicit tests.

As a methodology constructed in direct response to concerns over methodological features of previous implicit tests the FAST offers several notable improvements, both in

format and in its analysis of the data. Specifically, in lieu of response times as the chief metric of interest, the FAST uses the satisfaction of a learning criteria or response fluency measures as its primary index of task performance. Early versions of the FAST used the number of trials required to reach a response accuracy criterion on each block (e.g., ten consecutive correct responses) as the index of response class acquisition "speed". Another version of the FAST has relied on a slope score method. A slope score is calculated by first generating a cumulative record of correct responses for both the consistent and inconsistent block. A cumulative record is generated by plotting the number of correct responses in a block as a function of the time taken to complete that block. The slope of this function corresponds to a subject's rate of learning, the higher the slope of this function the faster their rate of learning was. The slope score is then calculated by subtracting the slope of the consistent block from the slope of the inconsistent block. A positive difference here indicates that responding was quicker and with fewer errors on the consistent block. A negative difference indicates the inverse, that responding on the inconsistent block was quicker and less error prone.

For both the response accuracy criterion and the slope score method the final index was the difference in learning rates across the blocks. These methods also provide an indication of the direction of bias towards learning one functional response class configuration over another. Reliance on scoring indices such as the correct response criterion and the slope score method means that the FAST uses metrics familiar to behaviour-analysts. These metrics also do not invoke cognitive concepts such as mental effort, as response time measures do in tasks such as the Stroop (1935) and the IAT (Greenwald et al., 1998). For this reason, it is curious that the IRAP relies on response times alone in calculating its index, but its score algorithm was perfectly in line with that of the IAT, presumably for reasons of

accessibility to the measure by social cognitivists. On the subject of response times, there is a key difference between how the IAT encourages rapid responding and how the FAST achieves this.

Rapid responding is a key feature of implicit tests. Response time constraints prevent subjects from engaging in a deliberative process of responding. After all, if a subject is granted too much time to deliberate on the correct response according to the contingencies of the block then there would be little difference between an implicit and an explicit test. The IAT achieves rapid responding through a process that is unacceptable to behaviour analysis. Namely, as was outlined earlier, participants are merely instructed to respond rapidly at the outset. In addition to this, if a subject fails to respond rapidly enough, there is nothing in the methodology of the IAT itself which prevents this. Rather, IAT researchers have ensured rapid responding through post-hoc data truncation (i.e., on any trial where the subjects response time exceeds 10,000ms, the data for that particular trial is deleted). A similar process of post-hoc data truncation is employed in the IRAP. The FAST by stark contrast has elected to impose response time constraints in vivo by limiting the window of responding to 3000ms. If the subject fails to respond within this timeframe, then the word “Wrong” in red text appears on screen, and their response is recorded as an error (Gavin et al., 2012). This is done due to the FAST’s behaviour analytic perspective wherein tighter experimental control is always preferable to post-hoc artificial alteration of the data. While it might be easy to point out that this is a relatively minor change, it is nonetheless an important one, and it is questionable why IAT researchers have never implemented a similar process.

The reason why the foregoing and other methodological changes have never been implemented might be to do with IAT researchers approach to scoring the IAT. Greenwald and colleagues have not updated the IAT’s scoring algorithms since their original paper on

the subject (Greenwald et al., 2003). They have instead allowed the format and calculation of scores to become reified. In their efforts to create an alternative to self-reports the developers of the IAT have fallen into the same trap by allowing the IAT to be treated in almost psychometric like fashion. That is to say, as a result of their elaborate scoring algorithms, it is now difficult to make alterations to an IAT without falling prey to problems of comparability with other IAT studies. In short, there is no incentive for researchers to alter an IAT dependent on circumstance, or because they might perceive a methodological problem in its format. To implement such changes would prevent the production of results which can be compared to the rest of the IAT literature. In effect, the IAT has become reified, and is not amenable to experimental alteration. To avoid this problem, the FAST has adopted a different philosophy entirely in its approach to experimentation. The FAST is entirely open source with the precise details of its methodology freely available in published research. Additionally, the reasoning behind its successive alterations is outlined in each study. Researchers who wish to utilise the FAST in their own studies are encouraged to make adaptations depending on experimental context, or other factors which they might think relevant. There is no danger that different FAST studies will not be comparable as from the behavioural perspective there is no underlying construct of implicit bias which may be missed if the test is tweaked in some small way. With that said, it seems time to examine how different experiments have employed the FAST, and what it has shown to be capable of thus far.

1.8.2 Using the FAST to Test for Laboratory-Controlled Stimulus Relations

O'Reilly et al. (2012) published the first study on the FAST as a test in its own right, and as such they wanted to emphasise the rationale underlying the test. Specifically, how a pre-existing stimulus relation can facilitate or inhibit the rate of functional response class

acquisition in a FAST block. As O'Reilly et al. (2012) was the first experimental analysis of the FAST researchers were careful to maintain a high level of experimental control through the use of laboratory-controlled stimuli relations, as opposed to social stimuli whose degree of relatedness cannot be known. Even if the FAST detected a relation between two natural verbal stimuli the accuracy of the FAST would remain unclear as the experimenter would still have no idea how strongly they were related previously. This experiment made use of the baseline blocks method outlined earlier (i.e., two practice blocks presented before and after the test blocks which provide a baseline level of response class acquisition). In addition, this FAST used a correct response criterion, wherein participants continued with a block until they had achieved 10 correct responses in a row (participants who did not achieve this within 100 trials were omitted). Upon the production of a response the feedback of "Correct" or "Wrong" was given. A response window of 3000ms was enforced, if the subject did not respond within this timeframe an incorrect response was recorded, but no feedback was given. The scoring method employed was a Strength of Relation (SoR) index, wherein the number of trials it took a subject to complete the inconsistent block was subtracted from the number needed to complete the consistent block. The block differential was then divided by the mean number of trials it took that subject to complete the two baseline blocks. Now that the FAST procedure in their study has been outlined, we should look at the laboratory-controlled stimulus relations that were trained ahead of the administration of the FAST.

O' Reilly et al. (2012) first exposed their subjects to MTS training, designed to establish two simple stimulus relations between nonsense syllables (A1-B1, B1-A1, A2-B2, B2-A2). The FAST was then administered, incorporating the two trained stimulus pairs and two novel stimuli (N1 and N2). The consistent block involved assigning a common functional response consistent with subjects MTS training (i.e., a left-hand key press for A1 and B1, a

right-hand key press for N1 and N2). In the inconsistent block, A1 and N1 required a common functional response and B1 and N2 shared a differed response (i.e., responding inconsistent with their MTS training). The order of the blocks was randomized across subjects. It was found that subjects formed functional responses more readily in the consistent block. Specifically, 13 out of 18 subjects showed a faster rate of response acquisition on the consistent relative to the inconsistent block. These results indicated that the FAST could be used to determine the pre-existence of stimulus-stimulus relations. Such a demonstration study was necessary as IAT research has never shown that it is sensitive to artificial stimulus-stimulus relations. Rather, its research process may have progressed too rapidly, and leapt straight to measuring relations researchers thought might exist, instead of relations that were known to exist, as they were created *ab initio*.

While an in-principle demonstration of the FAST procedure was welcome, it did not readily capture and index the relatedness of the types of relations of interest to social researchers. More specifically, if the FAST is to be used in an applied research context it must be sensitive to derived, as well as directly trained stimulus relations. This is because, derived relations more typically characterize naturalistic language than direct word associations (see Hayes et al., 2001). Given that the purpose of the FAST is to ascertain the relationships between real words used in the vernacular, it also needed to be tested for its suitability in indexing the strength of derived relations. To this end, O' Reilly et al. (2013) repeated and extended the 2012 study to train and test for equivalence relations, rather than a single conditional discrimination. This move was also important because neither the IAT nor the IRAP have been subject to analysis in terms of their ability to index derived relations specifically, as opposed to directly trained relations. In effect, when results are reported in the use of these tests in social research we have no way of knowing if the relations being indexed

were trained or derived naturalistically. It is reasonable to assume that the strength of derived relations may often be weaker than the strength of directly trained relations, although this is an empirical matter never examined in those research domains. From the perspective of the FAST literature (see O'Reilly et al., 2015) and consistent with an RFT perspective, the idea that implicit tests might index the strength of derived relations is a conceptual parallel to the idea that these tests index "implicit" relations. In other words, given that derived stimulus relations are by definition unreinforced and merely "implied" by the baseline relations that give rise to them, the concept of derived relations may make for an excellent technical description of what we might mean by the term "implicit". For example, an individual may show a racial bias on an implicit test because there are directly trained relationships between particular racial and negative terms in their vocabulary. Alternatively, the particular racial and negative terms are related only very indirectly in the complex, expansive, and subtly contextually controlled relational networks that represent their verbal repertoire. In other words, the relationship between particular racial terms and negative terms may never have been consciously discriminated, or reinforced, but is nevertheless supported by other practices within the verbal community in which the individual participates. In this sense, a racial prejudice is implied by the way in which the individual generally categorizes other stimuli, even if that particular relation between verbal relations has never been established.

O' Reilly et al. (2013) provided the following example that illustrates the above point. Imagine a parent who frequently refers to Irish people as "drunkards" in the presence of their child, and in a different context refers to all "drunkards" as "ignorant". Given these well-established verbal relations, a child may then derive an Irish-ignorant relation without reinforcement. The question that O' Reilly et al. had in mind when they conducted their 2013 study was whether or not it was possible, in such a case, for us to detect a bias against Irish

people as ignorant with such a child, even though their parents never explicitly told them that Irish people were ignorant. If the test is sensitive to this type of relation its utility is greatly improved in social research, and we can be confident that when we detect biases in social research that it might include very indirect or merely implied (i.e., implicit) relations among stimuli.

O' Reilly et al. (2013), employed the same FAST procedure as in the O'Reilly et al. (2012) study, but they slightly improved the SoR index calculation to reduce skew and kurtosis in the resulting data at the group level. They administered blocks of 32 training trials, across four training tasks designed to establish two related conditional discriminations (i.e., A1-B1, A2-B2, A1-C1 and A2-C2). This training was recycled until participants reached a criterion of 30 out of 32 correct responses. Testing proceeded in blocks of 16 trials, these trials probed for the derived B-C and C-B relations, without reinforcement, and for a limit of four cycles of the block. Participants then completed *three* consecutive FAST's and the associated baseline blocks in the following order: Baseline 1, Critical FAST blocks, Baseline 2, Critical Fast Blocks 2, Baseline 3, Critical FAST Blocks 3, Baseline 4. In all cases, a 10-consecutive-correct trial criterion was applied to complete a block. In this study, an SoR index was calculated by dividing the raw difference in trial requirements across the critical FAST blocks by the natural logarithm of the mean baseline trial requirement for those baseline blocks before and after each pair of critical FAST blocks. The advantage of this scoring system was that it creates a score of zero when a participant shows no difference in acquisition rates across the critical FAST blocks, is negative when their bias is in the unpredicted direction, and is positive when in the predicted direction.

Of the 16 participants who passed equivalence testing, 11 demonstrated effects in the predicted direction on the first FAST test block. Participants who did not pass equivalence

testing did not show any significant FAST effects at the level of means, or in the inferential analysis. O'Reilly et al. (2013) suggested that the weaker effects observed here compared to the O'Reilly et al. (2012) study were due to the derived nature of the relations being examined and were thus to be expected. This type of outcome may have interesting implications for the use of implicit tests in social research involving well established, as opposed to subtle and rarely explicitly derived verbal relations.

In a similar vein to the previous experiment, Cummins et al. (2018) sought to investigate whether the FAST was sensitive to the degree of relatedness between stimuli in laboratory-controlled stimulus relations. On a methodological note, this iteration of the FAST consisted of two blocks of 50 trials, one block designated the consistent, and the other the inconsistent. No baseline or practice blocks were employed. The response window for each trial was set at 3000ms, after which, if the participant had not responded, their response was recorded as an error and corrective feedback was presented. The FAST in this experiment employed the slope score method outlined earlier as its scoring system (i.e., the difference in the slope of the cumulative record between the consistent and inconsistent block).

With this explained, we can now examine the experimental manipulations employed by Cummins et al. (2018). IAT and IRAP studies have generally employed a "known-groups" approach wherein two groups who are thought to differ in relatedness between a set of stimuli are examined and compared in order to validate the respective test. What was truly needed however, was an examination of the variation in test effects between participants as a function of stimuli of either known relatedness or experimentally manipulated relatedness. A secondary question for this study was whether the FAST was sensitive to emerging derived stimulus relations at different stages of training. The reasoning behind this secondary purpose was that stimulus equivalence relations are generally only tested using an accuracy criterion.

As the FAST is also sensitive to response time, which may be an important variable in such an analysis, it could be a useful tool to researchers if it was indeed sensitive to derived relations at different stages of training.

In order to answer these questions Cummins et al. (2018) split their participants across six conditions. One group received a FAST with real words of strong relatedness, while the others received a FAST testing for arbitrary stimulus relations, which were experimentally manipulated with differential MTS iterations. The real words condition used words of known strong relation from a pre-existing index which lists the probability of a stimulus being discriminative of response with another stimulus (e.g., the word “Cheddar” is likely to be discriminative of a response with the word “Cheese”). Nonsense words were employed as the stimuli for all other groups. Any participant who did not reach criterion (i.e., did not display stimulus equivalence) within 40 minutes were excluded and thanked for their participation. One group received no MTS training whatsoever, along with the real words group these groups were administered the FAST immediately. Another group received 1 iteration of MTS training, and the FAST was administered immediately following this. A different group received a total of three iterations of MTS training sequentially, with approximately two-minute intervals between each (3-in-1 condition). The 3-in-1 condition were then required to complete the FAST. Participants who received two iterations of MTS training had their MTS sessions separated by 1 week. Following completion of the second MTS, they were administered a FAST. Finally, the last group received three iterations of MTS training, each session separated by a week (3-in-3 condition). Following their last session of MTS training, they were then administered the FAST. The consistent block for each FAST for all participants required responding consistent with their training, the inconsistent block

juxtaposed these requirements. Additionally, the consistent block for the real word condition involved responding consistent with the known relatedness between the stimuli.

The descriptive statistics were as expected, indicating that participants who received more than one iteration of MTS training produced larger FAST scores. The average block-slope scores and accuracy on the consistent block tended to increase as the purported relatedness between stimuli increased, and the average response times decreased. A mixed pattern of change for these three metrics was observed on the inconsistent block. Near-identical FAST scores were found between the real word condition and the 3-in-3 condition. A significant linear trend in FAST scores consistent with increases in relatedness between stimuli was observed across the conditions. A trend analysis revealed that greater stimulus relatedness resulted in a linear increase in differences between the consistent and inconsistent block. These findings suggest that the FAST is indeed indexing the degree of stimulus relatedness in an individual's learning history. Furthermore, these results indicate that the FAST may have some utility as metric of assessing the emergence and strength of stimulus relations. Cummins et al. (2018) was the first experiment of its kind to demonstrate that implicit measure effect sizes increase as a consequence of experimentally manipulated stimulus relatedness.

On a separate note, Cummins et al. (2018) also suggested an alternative method to the slope score employed in their study. Specifically, they speculated that a simpler measure employing the number of correct responses per minute, minus the number of incorrect responses per minute, could also index a participant's fluency on a FAST block. The difference in fluency between the consistent and inconsistent block could function as an alternate scoring method. This scoring method based on fluency would have the added advantage of ensuring that FAST scores are not subject to variation by bursts of rapid

responding. Bursts of rapid responding can increase the speed of responding and the number of correct responses per minute considerably, even when not under stimulus control. Because a simple response fluency differential system involves modulation by number of incorrect responses per minute, random responding will not overly increase a fluency score on a given block, as it may do using the slope score method. This response fluency differential system will be employed in the current thesis (See section 2.2.3.2).

The three foregoing studies were foundational in establishing the in-principle utility of the FAST method. However important procedural and conceptual questions remain, in answering these questions, basic processes can be illuminated further and beyond that achieved by research into other implicit test methods. More specifically, it was not yet known to what extent the FAST method indexed the relative degrees of relatedness between pairs of stimuli participating in derived relations. In other words, it was not yet known if the FAST score index was linearly related to the degree of relatedness of the stimuli in question. Put technically, the question being posed here is if the degree of nodal distance between stimuli in large derived relations is predictive of the FAST score recorded for the degree of relatedness of those stimuli. Previous research had shown that stimuli are more or less related based on their nodal distance from each other. For instance, Moss-Lourenco & Fields (2011) demonstrated that contrary to Sidman (1994) stimuli in an equivalence class are not functionally interchangeable or substitutable for each other. Rather, Moss-Lourenco & Fields (2011) showed that the degree of relatedness among stimuli in an equivalence class was an inverse function of the number of nodes separating the stimuli in the class.

Building on the above point, Cummins & Roche (2020) reasoned that if the FAST is to have utility in the assessment of relations, its score indices need to be linearly related to the nodal distance between stimuli under examination. In their study, 16 participants were trained

using a linear MTS procedure to form two four-member equivalence classes (i.e., A1-B1-C1-D1 and A2-B2-C2-D2) composed of nonsense syllables. After which, participants were immediately exposed to three FAST's, each of which tested for the relatedness of two pairs of stimuli. Naturally, the consistent block in these FAST's was consistent with the subjects MTS training. The FAST in this study was identical in methodology and scoring to Cummins et al. (2018) employing the slope score method once again. Each FAST differed in the nodal distance of the stimulus pairs under examination. Specifically, the three FAST's examined relations of zero (A1-B1 and A2-B2 relations), one (A1-C1 and A2-C2 relations) and two (A1-D1 and A2-D2) nodal distance, respectively. An MTS test for equivalence relations was administered only after the three FAST's were completed.

A decrease in FAST scores were observed as nodal distance between stimuli increased. This successive change in FAST scores proved to be significant using a trend analysis. In effect, the data showed that FAST effects are somewhat linearly related to the strength of relations being examined. Thus, we can now assume for the first time in the literature that a low score on these types of implicit tests does indeed indicate a weaker relation, rather than natural variance alone, or variances in the state of mental constructs such as mental associations.

Although the results of Cummins & Roche (2020) were promising, it remained the case that FAST effects were not reliable for each individual participant. More specifically, while 10 participants showed FAST effect variances broadly consistent with changes in the nodal distance parameter, six did not show evidence of such an effect, with five showing almost no variation in FAST scores across all three FAST's. Control of these effects at the individual level is still some way off due to the fact that the basic processes underlying theses effects are still being explored. However, at this point in time, it might be reasonable to

speculate that for those individuals who did not show variance in effects across the FAST's, the derived or implied relations may already have been well formed during the test procedure itself. In other words, while the derived relations were at no point reinforced explicitly, for reasons not yet known the FAST procedure may have been sufficient for the derived relations to emerge in these individuals and facilitate or retard learning on each block for pairs of different nodal distances. Put simply, for these individuals, nodal distance may not have much varied the strength of the relations across the various pairs. This possibility, combined with some variance in the FAST performance itself could obscure the expected effect using the current procedure. For the behaviour analyst, the solution to unwanted variance in behaviour is not to eliminate participants or to manipulate the data post hoc, but to alter the procedure to induce a narrower range of performances on each block of the test. Cummins & Roche (2020) suggested for example, that decreasing the length of the response window may reduce the range of fluency scores on each block and therefore potentially make differences across blocks more salient numerically. The current section has detailed the utility of the FAST in detecting laboratory-controlled stimulus relations. However, it may interest the reader to know that the FAST has been employed in the measurement of real-world social attitudes. Although, due to the nascency of the FAST method, few such studies have taken place as of yet. Nonetheless, it remains worthwhile to examine these studies in order to better understand the potential real-world applications of the FAST.

1.8.3 Applications of the FAST in Social Research

Only two published studies have examined the use of the FAST in real world social research. In one such study Cummins et al. (2019) employed the FAST to investigate "attitudes" towards condom use. Importantly, this study employed the SoR index used in the O' Reilly et al. (2013) study. As a reminder, this involved the inclusion of baseline blocks

and a correct response criterion, wherein a test block ceased once participants had achieved 10 correct responses in a row. A practice block involving entirely neutral stimulus categories was included in order to give participants some experience with the format and layout of the procedure, this block was fixed at 10 trials. The experimental manipulation in this study involved measuring the effects of the presentation of positively or negatively valanced message interventions in relation to the effects of condom use on the enjoyment of sexual behaviour.

Cummins et al. (2019) split participants into three conditions, a positive-message condition, a negative-message condition and a control no-message condition. The messages were presented on white laminated cards given to a participant for a duration of 30s prior to the practise block. In the no-message condition, the cards were blank, otherwise, dependent on the condition, the cards featured a positive or negative assessment of condoms. Before the first set of FAST testing blocks a baseline block employing novel unrelated stimuli was administered. The FAST testing blocks used images of condoms and sky images, the sky images functioned as neutral stimuli. In the first set of FAST test blocks, the consistent block involved condom images and positive words sharing a common response function, and sky images and number words sharing a different, but common, response function. The inconsistent block juxtaposed these requirements (i.e., positive words and sky images now required a common response function). The evaluative stimuli (i.e., positive and negative words) employed in this FAST were a subset of the stimuli employed by Greenwald et al. 1998.This first set of FAST blocks was dubbed the condom-positive phase. The order of the test blocks was varied between participants. The fact that the neutral stimuli were also images was important, this ensured that participants did not simply adopt a heuristic response pattern, wherein they would simply learn to press a specific key whenever they see a stimulus that is

not a word. In the next set of FAST test blocks (the condom-negative phase), the consistent block involved condom images and negative words sharing a common response function, and a different common response function for sky images and number words. The inconsistent block juxtaposed these requirements. Following this second set of test blocks an additional baseline block featuring different neutral stimuli to the first was presented.

The results indicated that the largest difference in FAST SoR scores was between the positive and negative-message conditions. Differences between the positive phase and negative phase FAST were only significant in the positive-message condition. There was no significant difference between phases in the negative-message condition, indicating no FAST effect. Despite this lack of difference between phases in the negative-message condition, there was a general qualitative shift in the expected direction (i.e., greater SoR scores on the negative phase FAST compared to the positive phase). Cummins et al. (2019) argued that the lack of a significant difference between phases for the negative-message participants may have been a result of their pre-experimental history in relating condoms to positive words. Some evidence of this argument was provided by the fact that participants in the control no-message condition, exhibited a slight skew towards the condom-positive phase FAST, compared to the condom-negative phase FAST. Taken together, the results were broadly in line with what was expected. That is, FAST scores were variable according to the specific message intervention provided, and therefore were serving as an index of the effectiveness of the message modalities.

Let us now examine the first use of the FAST in the assessment of naturalistic, non-experimentally controlled stimulus relations. Cartwright et al. (2016) used the FAST in an investigation of gender stereotypes in a sample of men and women. The version of the FAST used in this experiment was the same as that used in Cummins et al. (2018) and Cummins &

Roche (2020). Though Cartwright et al. was actually the first study chronologically to outline the use of the slope score, and the two block 50 trial method. In addition to this Cartwright et al. also employed a practice block of 16 trials (consisting of common everyday words) which was used to give subjects some familiarity with the FAST format. In order to assess gender stereotypes, Cartwright et al. (2016) incorporated stimuli relating to men, masculine traits (e.g., dominant, unemotional), women and feminine traits (e.g., nurturing, gentle). The consistent block required a common functional response for the male and masculine stimuli, and a different common response for women and feminine stimuli. The inconsistent block juxtaposed these requirements. For purposes of comparison, and to provide some convergent validity for the FAST, an IAT was also employed. This IAT adhered to the standard format (see Greenwald et al., 1998). The same stimuli as in the FAST were employed, in the same way, across both consistent and inconsistent blocks. Two self-report measures relating to gender stereotypes were also presented to participants. The FAST was presented first, followed by the two self-report measures, and the IAT was administered last.

Cartwright et al., found that 30 out of 30 participants who completed the FAST demonstrated significantly faster learning rates on the consistent block relative to the inconsistent block. Similar effects were observed on the IAT, where 27 out of 28 participants showed faster response latencies on the consistent block relative to the inconsistent block. Given this, it was somewhat unusual that while the two implicit measures co-varied, they did not correlate. The researchers suggested that this may have been a result of paradigmatic differences between the two measures, as the IAT measures speed of responding, while the FAST compares rates of learning across the two blocks. Somewhat interestingly, neither of the two implicit measures were linearly related to either of the explicit measures. Of course, this is not itself unusual, as the conditions under which implicit and explicit measures will

converge is an ongoing area of investigation. Taken together, these results broadly indicate that the FAST is indeed sensitive to naturalistic stimulus relations and can safely be employed as measure of bias or stereotyping in a real-world context.

With respect to the Cartwright et al., (2016) finding of a lack of correlation between the IAT and FAST it is worth noting that a direct comparison of the two tests using different scoring methods has been conducted but never published. Lalor (2019) conducted a study on attitudes towards abortion using an IAT, a FAST, and an explicit questionnaire. Both implicit tests were scored using a multitude of methods, including the D score method, the slope score method and the RFD method. The RFD method is the latest iteration of the FAST scoring method and will be employed in this paper (See section 2.2.3.2 for an outline). Lalor found a moderate correlation between the implicit measures across all scoring methods, but the strongest correlation was found when using the D score method. Additionally, a direct comparison with real world voting behaviour was possible due a referendum on the subject in Ireland. The explicit test demonstrated the highest accuracy, followed by the IAT, and then the FAST, though the difference between the implicit tests was not large. While this could suggest that the IAT is in some way superior to the FAST the results could just as easily be explained by various critiques of the D score method outlined earlier in this thesis. Namely that the increased diagnostic accuracy found using the D score method (a slight increase was also found when FAST scores were converted to D scores) can be attributed to the artificial data alteration embedded in the procedure (and thus not reflective of the participants actual performance). Whichever interpretation is taken, the results of Lalor (2019) confirm at least some convergence in findings between the FAST and IAT.

In relation to the stimuli employed by Cartwright et al., it could be argued that the functions of the verbal stimuli employed were assumed and indeed to a certain extent this is

true. While the particular meaning of certain words is certainly subjective to the extent that all language is socially constructed, the results of the FAST procedure itself seem to suggest that the prima facia meaning of the words was shared by all participants. Despite the fact that the stereotype words were assumed to be functionally related to either men or women, the results of the procedure itself vindicated this assumption. That is to say, Cartwright et al. demonstrated that for their participants the words 'dominant' and 'aggressive' were easier to establish as members of a common functional class with the word 'man', and the same held true for words the researchers judged to be feminine with the word 'woman'. Therefore, the assumed existence of these words as members of a common functional class was itself proved by the results of the FAST. Indeed, one of the main purposes of the FAST procedure is to confirm the existence of already existing functional classes, rather than to infer the precise semantic meaning of those words for each individual. If the stereotype words employed were not already in some way functionally related to either men or women, then the procedure would not have produced such clear and consistent results. However, it is worth noting that in principle it is possible to employ stimuli with known functions for each participant through a process of stimulus tailoring and this avenue of research has been given some theoretical thought (See section 4.6.4).

Further to the above point, there has been some research (albeit unpublished), which sought to ascertain the optimal stimulus set to employ in a FAST procedure. Specifically, Cartwright (2013) employed a series of homonegativity FAST's, each with procedural modifications, in an effort to assess the optimal stimuli to employ in testing this construct. In her first experiment, Cartwright employed a known groups paradigm wherein a pro and anti-gay group were compared in terms of their results on a single phase FAST using verbal stimuli. By single phase it is meant that there was only a single critical test block (in addition

to the practise blocks) which tasked participants with matching gay word exemplars with positive words. This FAST failed at distinguishing between the pro and anti-gay groups. Cartwright's second experiment was identical to the first, barring the fact that the stimuli representing the gay stimulus class were pictures of famous gay people. This iteration also failed to successfully discriminate between the groups. After further methodological tweaking (See experiment 3a; Cartwright, 2013) her last experiment (3b) involved a two phase FAST, a homopositive phase and a homonegative phase. In addition to this the 'gay' stimuli employed were more explicit, featuring images of men holding hands, kissing, or getting married. As a part of the shift to the two-phase FAST this experiment abandoned the known groups method in favour of assessing the correlation between this FAST and an explicit measure of attitudes towards homosexuals. It was Cartwright's view that the two-phase FAST would have broader applications, as it assessed contrasting histories of verbal relations in participants, and would allow for more precise discrimination among groups which may not be explicitly homophobic. This version of the FAST strongly correlated with the explicit questionnaire. Cartwright attributed this success mainly to the type of stimuli employed. It was her view that the verbal written stimuli employed in her first experiment likely did not evoke 'gay' functions due to their wide usage in society, conflicting usages in the vernacular, and their frequent pejorative use. Similarly, the use of famous individuals in her second experiment likely resulted in confounding cross-class evaluations in that they may have been viewed as a likable subtype, their sexual orientation may have been unknown to the participants, or their positive attributes were more salient than their homosexuality (Cartwright, 2013). The stimuli used in her last experiment however were clearly far more tailored to the target class (i.e., homosexual behaviour), and it was to this she attributed the success of the experiment. From this, we can see that the specific stimuli employed in a

FAST can have a pronounced effect on the results and is an area worthy of strong consideration in any study seeking to employ the FAST.

1.9 Using the FAST in Real-World Research

As the previous section outlined, there is a still growing body of research which demonstrates the utility of the FAST in assessing laboratory-controlled stimulus relations in a more ground up fashion than is apparent from the literature on the IAT. There has also been a limited number of studies employing the FAST in the assessment of real-world social attitudes. The IAT however, expanded quickly from proof-of-principle studies to application in areas of social relevance. The FAST, by contrast, has been developed in a slower, more methodical way. Owing to its behaviour-analytic underpinnings, it was the aim of early research to first understand the IAT in behaviour-analytic terms, and then to modify and improve it with respect to the findings of RFT and related behaviour-analytic work. Though this approach is far more empirically sound, it has come at the expense of studies applying the FAST method in social research. Moreover, while Cummins et al. (2019) demonstrated the FAST's ability to detect the effects of brief valanced messages on attitudes to condoms, this was not strictly an application of the FAST in the assessment of naturally occurring verbal relations. Therefore, only one published study, Cartwright et al. (2016) has employed the FAST for the analysis of real-world social histories.

The Cartwright et al. study was important as an in-principle demonstration of the FAST in a social research context. It also functioned to some extent as a conceptual bridge building study, insofar as it also employed the IAT for the same purpose, and demonstrated a high level of covariance between the two measures. Nevertheless, it likely would not qualify as a piece of translational research. This is because the examination of gender stereotypes in

Cartwright et al. (2016) was undertaken merely to test the utility of the FAST itself, rather than to model or explain phenomena as conceived within a different field from a behaviour-analytic perspective. In light of this, and because the FAST is leaving its nascent development phase, it would seem highly prudent to conduct research that not only assesses the utility of the FAST in a social research context, but simultaneously speaks to the theoretical frameworks of social-cognitive researchers. Such a strategy represents a bridge building exercise with neighbouring fields and may serve to make our research efforts more relevant to the concerns of psychologists outside of our immediate community (See Mace & Critchfield, 2010; Pilgrim, 2011).

One theoretical framework that has been used within social cognition to understand the sometimes unexpected effects resulting from IAT research is system justification theory (SJT; Jost & Banaji, 1994). This is a theory of high social relevance which has received some attention within the IAT literature and broader implicit theory more generally. It would appear therefore, to be an excellent choice of focus for research from within the behaviour-analytic domain. That is, rather than simply test the utility of the FAST in assessing expected biases within the community, we might test this in the broader context of SJT, which will now be outlined.

1.9.1 System Justification Theory

John T. Jost and Mahazarin Banaji first outlined system justification theory in 1994. They noted that social psychology hitherto had been unjustifiably replete with notions of ego and group justification in the understanding of the development of prejudices and stereotypes. More specifically, ego justification is a concept used to understand how stereotypes develop in order to protect self-interest (i.e., that people wish to hold favourable attitudes about

themselves). Group justification expands this idea to the level of the individual's social group, in terms of motivations to defend one's gender, ethnicity, class, and so forth (Jost & Banaji, 1994). These authors argued that such approaches to stereotype and prejudice development fail to consider the wider social influences. That is to say, individuals are not only motivated to defend themselves and their group, but are also motivated to defend and justify the society they live in. Therefore, a motivation to justify the system can sometimes supersede individual or group interests, and in turn incur negative evaluations of the self or the social group the individual belongs to.

Jost & Banaji (1994) argued against the prevailing social psychological theories of social identity and social dominance theory. A full outline of these social psychological theories is beyond the scope of this thesis. However, suffice it to say that from the system justification perspective these theories are too limited in their accounts of stereotype and prejudice. Specifically, social identity theory, they argued, expanded the notion of ego justification to the level of inter-group relations, and poses that stereotypes serve to rationalise how an individual's ingroup treats the outgroup (Jost & Banaji, 1994). Quite similarly, Jost and Banaji (1994), pointed out the fact that social dominance theory views stereotypes as "legitimising myths" which dominant groups hold in order to justify the oppression of other groups. Furthermore Jost et al. (2004) argued that the social dominance theory of stereotypes had arisen from wider evolutionary theory and relied too heavily on assumptions of genetic self-interest. In short, in the view of SJT theorists, the present social order cannot be explained as something that dominant groups impose and lower status groups resist. System justification holds that a motivation to defend and bolster the status quo is not found purely among dominant groups, but also supported by groups of lower social status, despite this being against their self-interest. As Jost (2019) puts it…

> "Why do some women feel they are entitled to lower salaries than men, why do people stay in harmful relationships, and why do some African American children come to believe that white dolls are more attractive and desirable than black dolls?" (p. 266)

It seemed possible that the type of questions SJT posed could be answered in terms of implicit cognition theory. SJT holds that groups who face discrimination display an outgroup bias towards the dominant group. Such a view cannot easily be explained in terms of either social identity or social dominance theory. While this is not the sole claim of SJT, it was a claim most easily addressed with the IAT. In fact, Jost et al. (2004) held that an outgroup bias towards the dominant group would be most easily measured at the implicit level. Indeed, for a variety of factors, social desirability among them, it would seem unlikely that a discriminated group would explicitly state their preference for the dominant group on a self-report measure. Before we go on to outline how such a claim might be tested with the FAST, we should examine some of the studies that Jost and colleagues point to as evidence of their theory.

Jost et al. (2002) performed an experiment relating to the claim of outgroup favouritism among lower status groups. In this experiment Jost et al. had students of Stanford University (high-status group) and San Jose University (low-status group) complete three separate IAT's, relating to self-concept, ingroup evaluation and implicit stereotyping. The self-concept IAT involved classifying words that were either self-relevant, or not self-relevant, with positively or negatively valanced words. The ingroup evaluation IAT featured stimuli relating to the ingroup and the outgroup. Finally, the stereotyping IAT involved ingroup and outgroup related terms, along with terms relating to academic or extracurricular activities. They found that double the amount of San Jose students relative to Stanford

students displayed an implicit outgroup bias. They sought to provide further corroboration for SJT by examining the correlations between these three IATS. It was their conclusion that, endorsing stereotypes of Stanford as academic, and San Jose as extracurricular, was associated with lower implicit self-esteem among the low-status San Jose students, but not the high-status Stanford students. Additionally, implicit ingroup bias was associated with higher implicit self-esteem for the Stanford students but not for the San Jose students. While this certainly provides some evidence for the claims of system justification theorists, there are clearer examples of this effect.

In another study, Rudman et al. (2002) demonstrated that the degree of outgroup bias was a function of the status of the minority group in question. The minority groups they examined included Jews and Asians (high status), overweight people (medium status) and poor people (low status). High-status minorities showed greater ingroup bias on an IAT than did low-status groups. In fact, the overweight and poor groups showed an implicit outgroup bias towards the dominant group. The greatest difference between explicit and implicit measures was found for poor people, the lowest status group. On the explicit measures they reported a strong ingroup bias, a result which was the reverse of their IAT scores, where a "dramatic tendency" to favour rich people over poor people was observed (Rudman et al., 2002). While the two foregoing studies are good general examples of the claims of SJT (i.e., that groups of low-status are more likely to display outgroup favouritism on implicit measures of bias), it would be worthwhile to take a closer look at some specific domains of interest.

In the context of race, Uhlman et al. (2002) used an IAT to investigate implicit skin colour biases among American and Chilean Hispanics. The groups of relevance to that study were the Blancos and Morenos, with Blancos being lighter skinned Hispanics than Morenos.

Uhlman et al. argued that Blancos are generally seen as a higher status social group than Morenos, and that Hispanic culture in general is dominated by Blancos, both politically, and in terms of socioeconomic status. It was found that Blancos had a significant Blanco-positive bias, as did Morenos but to a lesser degree than did the Blancos. That is, the discriminated minority expressed implicit outgroup favouritism towards a more dominant group. Of course, it is highly unlikely that shade of skin colour alone produced the finding of outgroup favouritism in the Uhlman et al. (2002) study. The findings are more likely to be a product of the social status of the groups in question, as the two studies above have already shown. While it could be argued that the Uhlman et al. study in the context of race was of limited size and scope, there is good reason to believe that the phenomenon of outgroup favouritism is widely spread.

One large study (Nosek et al. 2002) used data taken from projectimplicit.com (the publicly accessible IAT website offering a variety of tests) to compare European American (n = 103,316) and African American (n = 17,510) respondents in terms of implicit racial bias. It was found that on explicit measures African Americans showed an explicit ingroup bias, over and above the degree of ingroup bias found in the sample of European Americans. Furthermore, on the IAT it was found that African Americans demonstrated an outgroup bias, while European Americans showed an ingroup bias. Jost et al. (2004) conducted a secondary analysis on data curated from projectimplicit.com. They found that African Americans (n = 2,048) in their sample showed stronger explicit ingroup favouritism than the European American (n = 15,229) respondents (Jost et al., 2004). On the implicit measures however, European Americans showed ingroup favouritism, whereas the African Americans did not show the same pattern of responding. Overall then, Jost et al. (2004) found that 51.1% of European Americans showed ingroup favouritism on explicit measures, a figure that

increased to 78.4% in the implicit domain. Of the African Americans however, 65.4% demonstrated ingroup favouritism on explicit measure, inversely, 39.3% showed outgroup favouritism when measured implicitly. The two foregoing studies conform to the SJT prediction that disadvantaged groups will display an outgroup favouritism on implicit measures, that explicit measures are unable to reveal

In the context of gender, and informed by SJT, Jost (1997) examined the phenomenon of "depressed-entitlement" in a sample of women. That is to say, they wished to test whether women are likely to consider their equal work to be of lower value than that of male counterparts. While not an experiment into outgroup favouritism *per se*, nor did it utilise the IAT, this study was a part of a broader effort to show that oppressed groups will "internalise" their inferiority. Jost (1997) had a sample of women and men record their thoughts and opinions on several prompts (hereafter referred to as "thought lists"). He then had them rate the quality of other individuals thought lists. Participants were then asked to return to their own thought lists and rate them along several dimensions (e.g., logicality, sophistication, insight etc.) on a 15-point scale from "not at all" to "extremely". Finally, they were asked how much they would hypothetically pay an individual who had produced their work on a scale between 1 to 15 dollars (i.e., appraising their original thought list monetarily). The participant's thought lists were then evaluated by two external independent judges, one man and one woman, who were both unaware of the participants gender and of the hypothesis of the study. The judges appraised the thought lists in the same way as the participants. The results of which demonstrated that the independent judges did not perceive any differences between the thought lists of men and women. However, in their own self-ratings significant differences were found between men and women. Women rated themselves significantly lower in the dimensions of insight and self-payment. In the dimension of self-payment men

valued their work 18% more than women. In conclusion, despite there being little difference in how men and women rated the quality of their work, and no difference being found by external judges, women still exhibited a depressed-entitlement in terms of how much they thought their work was worth monetarily.

While these studies provide some empirical support for SJT, the theory has been formulated in social-cognitive terms that are not entirely amenable to the behaviour-analytic approach. It would seem prudent therefore to provide a tentative sketch of how some aspects of SJT can be understood in more conventional behavioural terms. As SJT is essentially a macro theory of society it cannot be conceptualised in terms of the behaviour of a single individual, therefore a behavioural conception of society is required. Fortunately, this is more feasible than it might appear at first glance. Despite a historic focus on the analysis of the contingencies which control an individual's behaviour, behaviour analysts have at least speculated about group behaviour. In Skinner's (1957) conception of verbal behaviour he described social interaction as a process whereby the verbal behaviour produced by an individual functions as the environmental cue for another individual's verbal behaviour. This process is cyclic in that the verbal response elicited by the second individual then functions as a stimulus for the production of more verbal behaviour from the first individual. Physical behaviour can also be substituted for verbal at any stage in this process. This conceptualization however, remains quite limited in size and scope.

If we expand our level of analysis out yet further to larger groups a complex web of interrelations is observed. Essentially an individual's response to a stimulus will function as the stimulus for a different person's response and vice versa, while the aggregate result of their combined behaviour might induce yet further responding from other individuals nigh ad infinitum. Such processes are referred to as a series of interlocking behavioural contingencies

(IBC; Glenn, 1988). Building on this the outcome of a series of IBC's will then function as a new contingency controlling the behaviour of the individuals who make up the group, a positive outcome will reinforce the current IBCs, and a negative one may lead to changes in the group's behaviour. This new contingency therefore is a metacontingency (Glenn, 1988). In the example of a business the aggregate product of the group's behaviour is a commodity to be sold on the market. If this commodity is profitable then it is likely that the current behavioural patterns will be preserved, while a failed product will likely induce institutional change. Now that we have the basis for our societal analysis laid out, we can begin to speculate as to how SJT could be conceptualised from this point of view.

Perhaps the most critical component of the concept of metacontingencies to this translational account is the idea of 'system-maintaining negative feedback' (Glenn, 1988). Unpacking this concept will be expediated by returning to our business example. If this business is a large national company with many regional sub offices, then each office is subject to its own IBC's. However, it cannot be reasonably expected that each and every individual will follow the exact same pattern of behaviour as their counterparts in distant offices, rather much individual behavioural variation is to be expected. These regional offices are still subject to the needs of head office (i.e., under control of the same metacontingencies), therefore only variations which nonetheless produce the same outcome will occur. Changes/variations which are drastically different will likely be extinguished unless they produce something better. This is essentially the metacontingent account of system-maintaining negative feedback, and lines up quite nicely with the definition of system-justification i.e., "psychological processes contributing to the preservation of existing social arrangements even at the expense of personal and group interest" (Jost, 1994 pg.1).

From this, we can see that broadly speaking SJT can be understood in behavioural terms, but what about the specific claim of outgroup favouritism?

Before we can account for the phenomenon of outgroup favouritism there is another important feature of the theory of metacontingencies which must be mentioned. Namely, the fact that much like a football team the members of the substructures which make up society can be substituted/replaced while preserving their function, subject to the metacontingencies controlling their behaviour e.g., scoring goals. As a country which for the majority of its history has been in the hands of white men, the influx of women and ethnic minorities into the Irish workplace and other institutions may, contrary to expectations, not constitute as much of a change to the overall structure of society as may be expected. Entering into these structures may not change the outcome of individual behaviour, it seems more likely that women and ethnic minorities would instead come under the control of the prevailing metacontingencies of society. That is to say, rather than changing the system, the system changes them. If the patterns of behaviour present in the IBCs of society and the metacontingencies which control them were already essentially racist/sexist in nature, then the behaviour of these women and ethnic minorities may in fact lead to self-prejudice. This is admittedly a speculative account, a full translation of SJT into behavioural terms is entirely beyond the scope of this thesis, but it does demonstrate in principle that the concepts arising from SJT are more amenable to behaviourism than one would think.

Taking from these studies and the translational account above, we can conclude that there is sufficient empirical evidence of the claims of SJT to allow it to guide us in our examination of social attitudes (Jost, 2019; Jost et al., 1994; Jost et al., 2004; see for an overview of SJT literature). If the current research using the FAST method finds outgroup favouritism biases, then it would provide some convergent validity for the method in terms of

its aligning with theoretical expectations in the field. To examine this idea in the current research context, it will be required to compare known groups with each other on both explicit and implicit measures along the lines of the research designs outlined up to this point. This method is known as the "known-groups paradigm" and will now be briefly discussed.

1.9.2 The Known-Groups Paradigm

If similar results to the findings of the SJT studies can be found using the FAST method, then this would provide some convergent validity for the FAST method. Of course, in order to assess such a claim, it would not be sufficient to examine any one social group in isolation. Rather, two groups of known differing social status would need to be examined in relation to one another.

It was perhaps Cronbach and Meehl (1955) who argued most strongly for the use of the known-groups paradigm when ascertaining the validity of a test. It was their view that when two groups are theoretically expected to differ on a metric, then testing this can provide validation for the metric in question. The example Cronbach and Meehl provided, was an experiment on the validation of scale thought to measure attitudes towards the church. Validation for this scale was provided by showing score differences between those who attended church and those who did not. In effect, if a test can be shown to successfully discriminate between known groups, then this provides support for a broader scale generalization of the test to different samples where potential differences are not known (Hattie & Cooksey, 1984). With this in mind, we can now outline how the FAST can be applied in the context of system justification.

1.9.3 The Current Study

The FAST will be employed in the current experimental context to first of all measure attitudes in a social research context and examine both the sensitivity of the FAST to the existence of suspected stereotypes, and its convergence with explicit measures of the same construct. In both of the studies reported here, known social groups will also be employed and compared on their performance on the FAST, and on the explicit self-report measures. As both groups in the case of both studies will be in power relationships with one another, this study will broadly speak to the type of ideas that have arisen from SJT literature. According to SJT, a low-status group should display an implicit outgroup favouritism, not captured by an explicit measure. Therefore, SJT might expect groups in power relations with each other to score similarly in terms of their implicit favouritism towards the more powerful group. To examine this idea in the context of the FAST method, the current thesis will first report on the use of the FAST in measuring gender biases amongst a sample of male and female participants. Participants will also be exposed to an explicit self-report measure of gender bias. Interestingly, while SJT has examined the phenomenon of outgroup bias across several social groups, no dedicated empirical study has been conducted to examine this phenomenon in female participants using an implicit test. Although it has been shown that women exhibit depressed self-entitlement in comparison to men, and this finding has been used by system justification theorists to support their position, it is not yet clear whether or not this expectation will be upheld using an implicit measure. However, the assumption that this should be the case is clearly implied by the literature. This research is therefore exploratory in nature and is relatively novel in its attempt to apply the FAST method in the context of SJT. However, it should be noted, that the primary goal of this research remains an assessment of

the FAST in a real-world context. This research does not intend to make a definitive claim on the veracity of SJT one way or the other.

The groups to be recruited for Experiment 1 will be males and female adults residing in the Republic of Ireland. A demographic survey will be administered to assess gender, age and residency. The groups will then be administered an explicit self-report measure of sexism. This will be followed by a FAST designed to assess attitude to gender. This FAST will use images of men and women as the two target stimulus categories, along with simple valanced evaluative terms (i.e., words relating to "good" and "bad") to function as the two evaluative word categories. In other words, the current research will use a hybrid picture/word stimulus method as used in the original IAT research.

The images of male, female, Black and White faces, as well as the evaluative terms were all taken from the race IAT used on the projectimplicit.com website (Greenwald et al., 2003; Nosek et al., 2002). This was done in order to manipulate as few variables as possible across the IAT and FAST procedures except the test parameters of importance. As the reader will see, the same stimulus evaluative terms were used in Experiment 2 in the context of examining race, but configured in such a way that functional response classes consisted of classes of ethnically identifiable similar stimuli. This is, as opposed to Experiment 1, wherein stimulus classes were discriminable in the context of gender. Once again, this allowed for greater consistency across the studies in which only key parameters are to be manipulated. Another advantage of employing the generically evaluative verbal stimuli as employed in the early IAT research, as opposed to specific stereotyped gender traits (e.g., professions, personality traits), is that it would allow the researcher to examine general hierarchical preferences for males over females. This is as opposed to simply measuring the existence of very specific stereotypes regarding masculinity and femininity which would not necessarily

reflect power relations or overall general preferences or evaluations. It could be argued that due to the age of these stimulus items their function may have changed over time, as it has been 20 years since their original usage. Factors such as the clothing worn in the facial images or the expressions on each face could influence the test in unintended ways e.g., clothing could indicate social class, an angry expression carries negative connotations. However, the images selected were all close-up shots of the individuals rendered in colourless black and white. Furthermore, all images featured neutral expressions and therefore should be absent of any unintended connotations. In regard to the written word stimuli, all of the words are universally positive (e.g., Love, Pleasure, Happy), or negative (e.g., Filthy, Rotten, Evil). It is unlikely therefore that, the functions of any the stimuli employed in FAST could have changed since their original use in IAT studies.

The first experiment supported the utility of the FAST in the context of the assessment of gender bias. However, some interesting data patterns emerged which alerted the researcher to potential social confounds that made it unlikely that SJT would have been supported using the current method in a sample of men and women. Additionally, a return to the SJT literature suggested the phenomenon of outgroup bias might be more likely to be found in different social groups, and indeed may not be applicable in the context of gender for a variety of reasons which will be discussed. Experiment 2 was therefore conducted in the context of racial bias, in an attempt to more fairly assess SJT in line with predictions which can be reasonably made, and to further extend the appraisal of the FAST as a viable method for measuring attitudes in the context of social research.

Chapter 2

Assessing the FAST as a Means of Detecting Gender Bias in a Sample of Men and Women

2.1 Introduction

As outlined in Chapter 1, the body of evidence arising from SJT suggests that low-status groups may display an outgroup favouritism towards dominant groups (Rudman et al., 2002). This outgroup favouritism may be more readily captured by implicit measures of bias (Jost et al., 2004). While the literature surrounding SJT has indicated the presence of the outgroup bias phenomenon in several social groups, no empirical study has ever examined this phenomenon in female subjects using an implicit measure. The purpose of the present experiment therefore, is to explore gender biases in a sample of men and women. The implicit FAST will be compared to the Modern Sexism scale (Swim et al., 1995; See Appendix A), in order to test whether the two metrics diverge. The Modern Sexism scale is split into two subscales, one measuring old-fashioned overt sexist beliefs and the other measuring more modern forms of sexism, only the latter will be employed in data analysis. The MS scale is designed to assess participants attitudes towards women in society, and should be an apt measure of a participant's explicit beliefs/perceptions about the role of women in modern society. It is expected that as women are a low status group relative to men they may display an outgroup bias towards men. Alternatively, perhaps women will display a diminished ingroup favouritism, relative to the male participants' ingroup favouritism. Both groups are expected to display similar scores on the explicit measure.

In format, Experiment 1 will proceed in the following way. A roughly equal sample of men and women will be recruited to complete an explicit and implicit measure of sexism. Prior to these measures, a demographic survey (See Appendix B) will be administered to assess age, sex, ethnicity and residency in the Republic of Ireland. The performances of the participants on the FAST will be correlated against their scores on the Modern Sexism scale (See Appendix A) to assess convergent validity. The stimulus classes employed in the FAST

will be images of men and women (of both White and Black ethnicities) as target stimulus classes, and positively and negatively valanced words as the evaluative stimulus classes. Female and male faces will always be assigned to different keyboard responses. Because a male positive verbal history is presumed, the consistent block will require learning a common functional response for male and positive verbal evaluative stimuli. A different, but common functional response will be required upon presentation of female and negative verbal evaluative stimuli. The inconsistent block will juxtapose these requirements, such that, male face stimuli and negative evaluative verbal stimuli will require the same functional response, and a different common functional response will be required for female face stimuli and positive evaluative verbal stimuli.

Under the contingencies of the FAST it is assumed that a common topographical response to different stimuli is a functional response class. However, this is not to say that topographically similar responses always denote common membership to the same functional response class. It is perfectly possible that even two behaviours with very different topographies are members of the same functional response class. To take an example from Hayes and Long (2013)

> "…consider a person who has gone to a restaurant for a lunch meeting with a friend. When he enters she gets his attention by waving her hand. The action is not merely one of raising the hand—the action is one of getting attention. If the person's arms were too tired to raise, the same functional action may have been instantiated by calling out, or standing up, or pushing back a chair. Hand raising is a participant in that whole event, but the event is not an assemblage—it is a functional whole. Without understanding the history, situation, and purpose of the act, the act itself cannot be appreciated. Getting the attention of a friend for a lunch meeting includes

> such contextual features as the history with this person, the circumstances that led to calling the meeting, and the agenda that will be covered." (Hayes & Long, 2013 p.6)

In the above example it can be seen that despite quite different topographies the behaviours in question remained members of the same functional response class. Approaching this from a different angle and on the basis of the same logic, it is entirely possible that responses of similar topographies are members of mutually distinct response classes. This would be dependent on the discriminative stimuli present, and how they are modulated by specific contextual cues, potentially leading to topographically identical responses that are nonetheless distinct from each other. To return this theoretical question to the context of the FAST, it remains possible that for some subjects the stimuli employed may have participated in histories unknown to the experimenter. This could compromise the degree to which they were established as affective discriminative stimuli for the intended response. With this in mind, such a process may threaten the integrity of the formation of the intended functional class. Of course, it would be impossible to completely control for this possibility without having a nigh infinite knowledge of the specific histories of each participant. As Skinner (1974) argues, analysis need not take into account every single factor in the history of a subject, it is satisfactory for analyses to proceed only so far as "effective action can be taken" (p. 210). Furthermore, in the context of the FAST, topographically identical responses within classes (e.g, press Z for Men and "good") and topographically distinct responses across classes (e.g, press Z for Men and "good" and M for Women and "bad") were generated under laboratory conditions. There were no contextual differences across the reinforcement conditions that were not controlled for. In short, though it can be dangerous to assume common membership in a functional class on the basis of topography alone, in the context of the FAST the behavioural topography was artificially generated and therefore under tight experimental control. Knowing this, employing the FAST here to test

for class compatibility between male/female stimuli and positive /negative stimuli should be a safe endeavour, unhampered by issues of differing behavioural topography.

2.2 Method

2.2.1 Participants

Ninety-eight participants (32 identified as male, 66 as female) aged between 18 and 60 participated in this study (Mean age = 23.77, SD = 8.73). The vast majority identified as White Caucasian, with the remainder identifying as Non-White of various ethnicities. Participants were predominantly recruited from the student cohort of Maynooth University. However, a snowballing recruitment method was employed, wherein participants were encouraged to share the study information (See Appendix C) and advertisement link (See Appendix D) with others. An open-ended response format for reporting gender was employed so as not to impose a binarized forced-choice. Participation was voluntary. However, a course credit was available to some participants who were currently enrolled in one module of an undergraduate Degree in Psychology at Maynooth University. Inclusion criteria included fluent English, normal or corrected to normal vision, full use of both hands and residence in the Republic of Ireland.

2.2.2 Ethical Considerations

This study was approved by the research ethics committee of Maynooth University. Participants were made aware that they could withdraw at any time during the experiment, and that their data could be excluded from the study by contacting the researchers via email. To this end, all participants were provided with a code (generated randomly by the Inquisit software) after completion of the consent form (See Appendix E), which they were encouraged to record. This code could then be sent to the researchers via email, allowing them to identify the participant's data. These data could then be removed from analysis or

provided to the participant in raw (uninterpreted) form depending on their request. However, no participants availed of this option. Following completion of the FAST, participants were debriefed with information about implicit testing and the full hypothesis of the study was provided (See Appendix F)

The FAST and MS scale are subclinical in nature. That is, they do not allow diagnosis of any condition or trait. Thus, there was no possibility of sensitive information being gathered about individuals. In addition, the data gathered was totally anonymous, so it was not possible to link test results to individuals without them providing their code. All data gathered was stored on the Millisecond Inc. Dublin server in a fully GDPR compliant way. Data was not transferred outside of the European Union at any point. Millisecond Inc. protects data from unauthorized use and complies with EU standards on data modification.

Due to the online nature of this study, some additional precautions were taken. Notably, best practise as dictated by Barchard & Williams (2008) on the exclusion of children from online research was observed. Anyone who indicated they were below the age of 18 on the demographic questionnaire was unable to progress any further. In addition, the study was only advertised in places where children were not expected to be found. To avoid drawing attention from children, the recruitment calls avoided the use of colourful imagery or any cartoon-like design. Given these precautions and the somewhat tedious nature of the tasks involved, it is likely that minors were successfully excluded from the study.

2.2.3 Apparatus

The Modern Sexism Scale, and the Function Acquisition Speed Test were delivered online using Inquisit software hosted on the millisecond.com European server

2.2.3.1 *Modern Sexism Scale*

The Modern Sexism Scale (Swim et al., 1995) is a 13-item scale, designed to appraise participants attitudes towards women and gender in society. The MS is split into two subscales, a 5-item scale intended to gauge more old-fashioned overt sexist beliefs (e.g., "Women are generally not as smart as men") and an 8-item scale focusing on modern forms of sexism such as the denial of continued discrimination against women (e.g., "On average, people in our society treat husbands and wives equally"). All items are scored using a Likert-scale from 1 (strongly disagree) to 5 (strongly agree). Though the entire scale was administered, for the purposes of this experiment, only the modern sexism-subscale was employed in data analysis. Some questions were modified slightly to account for the Irish context in which it was employed (e.g., "It is easy to understand the anger of women's groups in Ireland"). Conventionally, the scale is scored in such a way that, higher scores indicate lower degrees of bias against females. However, as the FAST is scored inversely (i.e., higher scores indicate more bias against females) the MS scale was reversed to be in accordance with this, for ease of data analysis.

2.2.3.2 *Function Acquisition Speed Test*

The FAST at its most rudimentary is simply a learning task, wherein subjects learn how to respond in one of two ways (i.e., a press of the "Z" or "M" key on a computer keyboard) upon the onscreen presentation of particular stimulus (i.e., a picture or word) on the basis of trial-by-trial feedback provided on screen (i.e., either "CORRECT" or "WRONG"). Unlike other implicit tests, the FAST does not provide instruction on the correct pattern of responding at the outset of the task. Instead, subjects learn as a result of the reinforcement contingency program for each of the two blocks of training that constitutes the FAST. These contingencies produce two functional response classes in each block (i.e., four

in total). Each stimulus item, whether word or picture, serves as a discriminative stimulus for a particular response. All stimuli fall into one of four categories, established either in the laboratory, or in this case, in natural language outside the laboratory (e.g., good words, male faces). The same stimulus items are employed in each block, but responses are reinforced according to a different contingency. That is, in one block, members of two verbal categories share one positional keyboard response, and members of two verbal categories share an alternative positional keyboard response. In this case, the two categories whose exemplars are discriminative for a common response are compatible in the history of the participant either through laboratory training, or through socialization and the use of the vernacular. In the other block of the FAST this contingency is juxtaposed such that exemplars from incompatible categories share a common response. In effect, the FAST indexes the degree of relatedness between categories (classes) of stimuli by examining the rate of acquisition of these functional response classes for each subject on each block of the test/training, and comparing these to identify which set of contingencies appears to be most compatible with the learning history of the participant.

The FAST consists of two phases/blocks, one which is expected to be in accordance with pre-experimental learning (consistent block) and one which is expected to diverge from pre-experimental learning (inconsistent block). In this instance, a male positive bias was presumed, therefore the “consistent” block required common positional keyboard responses to exemplars from a set of male facial images and positive words, and a second common positional keyboard response to images of female faces and negative words was reinforced. The order in which these blocks are presented to each subject was randomised by the Inquisit software. Before the first block, and then again in the interval between blocks, the following text was presented to subjects:

“In this task, you will need to use the 'Z' and 'M' keys on your keyboard. When you next press the spacebar, positive and negative words, and pictures of male and female faces of different ethnicities, will begin to appear on the screen, one at a time. You must learn to press either the 'Z' or the 'M' key, depending on what word or image appears on the screen, and based on the feedback that you are given after each response. Try to respond AS QUICKLY AND AS ACCURATELY AS POSSIBLE. When you're ready, press the spacebar to begin.”

After the participant presses the spacebar, the first intertrial interval (ITI) of 500ms was presented (i.e., a blank white screen). After the initial ITI the first stimulus was presented on screen, centred and in size 32-point font. The subjects were required to respond with a press of the “Z” or “M” key, if they did not respond within the 3000ms time limit, then the feedback for an erroneous response would be presented. If the subject successfully responded within the time limit, then the screen instantly cleared, and appropriate feedback according to the block contingencies would be presented (e.g., Consistent block: Female Image – Z key - “CORRECT”, “Rotten” – Z Key - “CORRECT”). Each block consisted of 50 trials of this type, involving stimulus exemplars taken from four categories of stimuli (i.e., four facial images of women, four facial images of men, four generic positive words, four generic negative words). Stimuli were selected for the study at random from a larger set of stimuli from previous IAT studies (Greenwald et al., 2003; Nosek et al., 2002). Stimuli were presented in a quasi-random order, with one stimulus from each of the four categories being chosen for presentation by the Inquisit software for each successive cycle of 4 trials. The FAST consisted of 12.5 cycles of these four-stimulus sets, amounting to 50 trials in total. The FAST has been scored in different ways in the past, but the scoring system employed in the current study is based on a suggestion by Cummins et al. (2018). In this scoring system, the rate of learning during each block (i.e., two functional response classes established

simultaneously), is calculated in terms of the differences between the rates of correct (CRPM) and incorrect (IRPM) responses per minute, divided by the total time taken to complete the 50 learning trials in the block. This calculation provides a response-rate differential score (RRD) for the block. An overall FAST score is calculated as a rate-fluency differential (RFD) score across the blocks. That is, the FAST score is calculated by subtracting the inconsistent block RRD from the consistent block RRD. This is expressed in equation 1 below. Importantly, inter-trial intervals and feedback presentation time are included in the total time metric. In the case of this experiment, a positive RFD score would indicate a male positive bias, and a negative score would indicate a female positive bias, insofar as this would indicate faster learning under contingencies reinforcing common responses to male and positive stimuli as well as to female and negative stimuli. In contrast, a negative score would indicate a female positive bias, insofar as the juxtaposed contingencies of the inconsistent block would, in this case, have been associated with faster functional response class acquisition and therefore a compatibility between male and negative stimuli, as well as between female and positive stimuli. Some of the analyses in the following sections will use individual fluency scores for each block as opposed to overall RFD scores. A blocks fluency score is calculated by subtracting the total incorrect responses on a block from the total correct and dividing the result by the total time taken for that block.

$$RFD = \left(\left(\frac{TC_C - TI_C}{TT_C}\right)\right) - \left(\left(\frac{TC_I - TI_I}{TT_I}\right)\right) \text{ x } 60000$$

Equation 1: Formula to calculate the rate-fluency differential score. TC denotes total correct, TI denotes total incorrect, and TT denotes total time taken in that block, the lower-case C or I indicate consistent or inconsistent block respectively. The result is multiplied by 60,000 so as to convert the value to ‘per minute’

2.3 Procedure

All experimental sessions were conducted remotely using the online Inquisit tool (from Millisecond Inc). Participants accessed the study using a link which led them to the first online page of the study hosted on the Millisecond Inc. server. Thus, participants completed the study on their own computer, in a place and time of their choosing. However, the initial study information section (See Appendix C) emphasised the importance of conducting the session in a quiet relaxed environment free of distractions. Some information about the study was provided alongside the link (See Appendix D). Because of the snowballing recruitment method employed, some participants who were acquaintances of the researchers were sent a link directly, others may have seen it advertised online or heard about it through a friend. Upon clicking the link participants were brought to a consent form which had to be read and acknowledged before proceeding. The experimental tasks were completed in the following order: 1) demographic questionnaire, 2) Modern Sexism scale and 3) FAST. The FAST was completed following the self-report measure (MS), because it was reasoned that any demand characteristics created by the experimental setting, or emerging participant awareness of the purpose and hypotheses of the study, would have greater effect on the self-report measure than on the implicit measure. Thus, the self-report measure was delivered before the implicit measure. Upon completing these tasks, participants were brought to a debriefing screen where the full purpose of the experiments aims, and central thesis was outlined. The contact information for the Maynooth research ethics committee and both researchers involved were provided alongside the debriefing information. This was to provide participants who may have felt there was an ethical violation in the experimental proceedings an outlet to contact with these concerns

2.4 Results

2.4.1 Missing Data and Excluded Cases

A small number of participants (N = 4) missed or did not answer two or fewer questions on the Modern Sexism scale. As this represented a very small portion of the dataset, a complex data replacement system was not required. Instead, a neutral score of 3 on the 1-5 Likert scale, was used in place of these missing scores in each case.

A portion of the participants were excluded from the analysis due to not completing the MS scale or the FAST. Additionally, if a participant scored 0 on any block of the FAST (i.e., indicating no responding at all) their data was excluded. One participant's dataset was excluded because the participants stopped responding during the FAST procedure and then restarted their participation from the beginning shortly afterwards. On the bases of these criteria a total of 6 participants were excluded from the analysis.

2.4.2 Descriptive Statistics

The means and standard deviations for males and females on the FAST blocks, RFD score and MS scale are provided in Table 1 below. Females generally achieved lower RFD scores than males (mean difference = 3.93, 95% CI: 1.43 to 6.43), indicating a female-positive bias. Males generally produced higher scores on the MS scale (mean difference = 4.73, 95% CI: 2.05 to 7.40), indicating a bias against women.

Table 1. *Means and Standard Deviations for Blocks scores, RFD scores and MS scale*

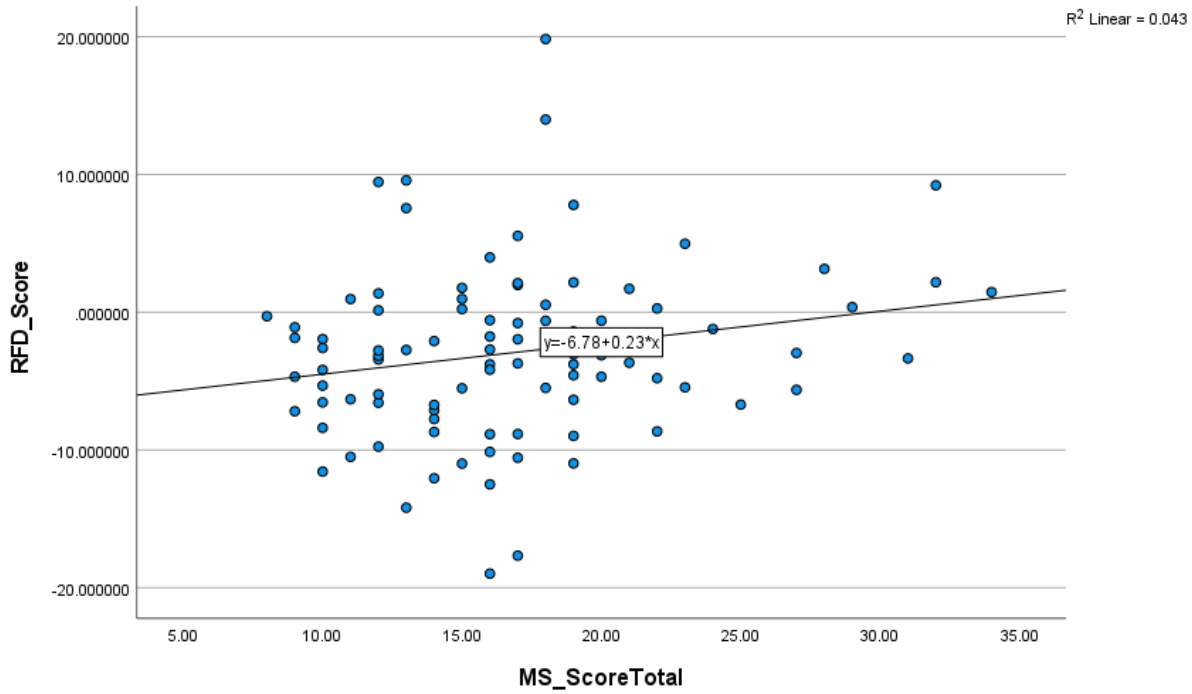
	Male			Female		
	N	M	SD	N	M	SD
RFD Score	32	-.26	5.12	66	-4.19	6.16
Consistent Fluency	32	16.27	7.21	66	17.79	6.46
Inconsistent Fluency	32	16.53	6.84	66	21.99	7.01
MS Scale	32	20.13	6.98	66	15.39	3.88

2.4.3 Correlations

The relationship between RFD scores, and the Modern Sexism (MS) scale scores was investigated using a Pearson product-moment correlation coefficient. Preliminary analyses were performed to ensure no violation of the assumptions of normality linearity and homoscedasticity. These analyses revealed, that in the case of both the RFD score and the MS score, the assumption of normality was violated, as indicated by a Kolmogorov-Smirnov test (MS, $p = .001$; RFD, $p = .048$). An examination of the histogram reveals that the data are relatively symmetric for RFD score but in the case of MS scores the data peaks in the middle but trails off to the right side. The histograms suggest that the data are relatively normally distributed for RFD scores but unevenly distributed for MS scores. The scatterplot (Figure 2) suggests a relatively linear relationship, and an even clustering of points suggesting that homoscedasticity was not violated. As none of these violations were particularly egregious, the correlation analysis proceeded as planned. There was a small, positive correlation between RFD score and MS score, $r = .207$, $n = 98$, $p = .041$, suggesting that higher RFD

scores (indicating more gender bias against females) corresponded with higher MS scores (indicating more gender bias against females; see Figure 2 below).

Figure 2: *A scatterplot representing the relationship between RFD scores and MS scores*.



Note. Higher scores on both measures indicates a bias against females.

Table 2: *Pearson Product-moment correlations between RFD scores and Modern Sexism scores for the sample as whole*

Scale	**1**	**2**	**3**	**4**
1. RFD Score	-			
2. Modern Sexism Scale	.207*	-		

Statistical significance, *p < .05; **p < .01; ***p < .001

2.4.4 Correlations By Gender

In order to examine the correlation between MS and RFD scores more closely the sample was split by gender between males and females and investigated again using a Pearson product-moment correlation coefficient. Preliminary analyses were performed to ensure no violation of the assumptions of normality, linearity and homoscedasticity. These analyses revealed that for both men and women in the case of MS scores, the assumption of normality was not violated as indicated by the Kolmogorov-Smirnov tests. In the case of RFD score however, the assumption of normality was violated for the women in this sample ($p = .021$) but not for the men, again according to a Kolmogorov-Smirnov test. The histograms seemed to support these findings, with the distribution of scores being relatively symmetric for both groups in the case of MS scores (though the data did somewhat cluster more to the left of centre for women). For RFD scores, again the data was relatively symmetric for men, but clustered more to the left for women. The scatterplots (See Figures 3 & 4) indicate that the assumption of homoscedasticity was not violated, however the distribution is indicative of no relationship between the two variables. Despite this, for the sake of clarity on the nature of the relationship between the two variables the analysis proceeded as planned. In the male sample there was no correlation between RFD and MS scores, $r = .145$, $n = 32$, $p = .428$. A similar finding was found in the sample of women, $r = .075$, $n = 66$, $p = .547$. These findings indicate that the small correlation found when the groups are examined together disappears when the sample is split between men and women. This finding could be attributed to the lower sample size present when the groups are analysed individually, as a result, the tests may not have had sufficient power.

Figure 3: *A scatterplot representing the relationship between RFD scores and MS scores for Males*

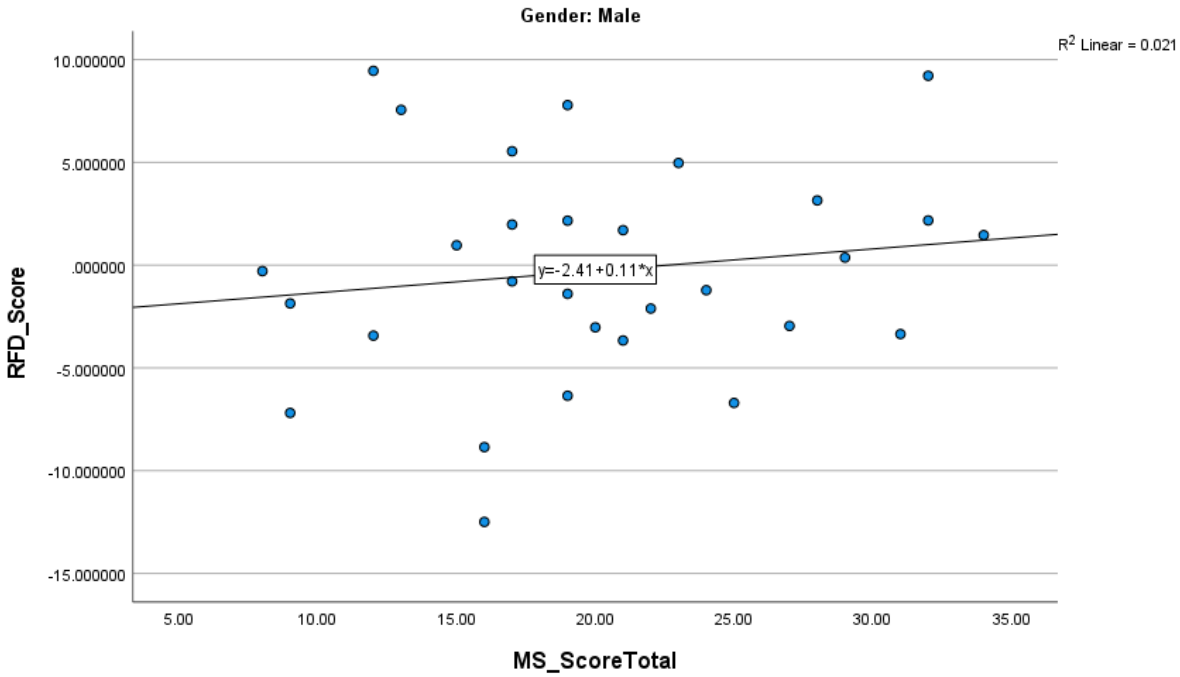


Table 3: *Pearson Product-moment correlations between RFD scores and Modern Sexism scores for male sample*

Scale	**1**	**2**
1. RFD Score	-	
2. Modern Sexism Scale	.145	-

Statistical significance, *p < .05; **p < .01; ***p < .001

Figure 4: *A scatterplot representing the relationship between RFD scores and MS scores for Females*

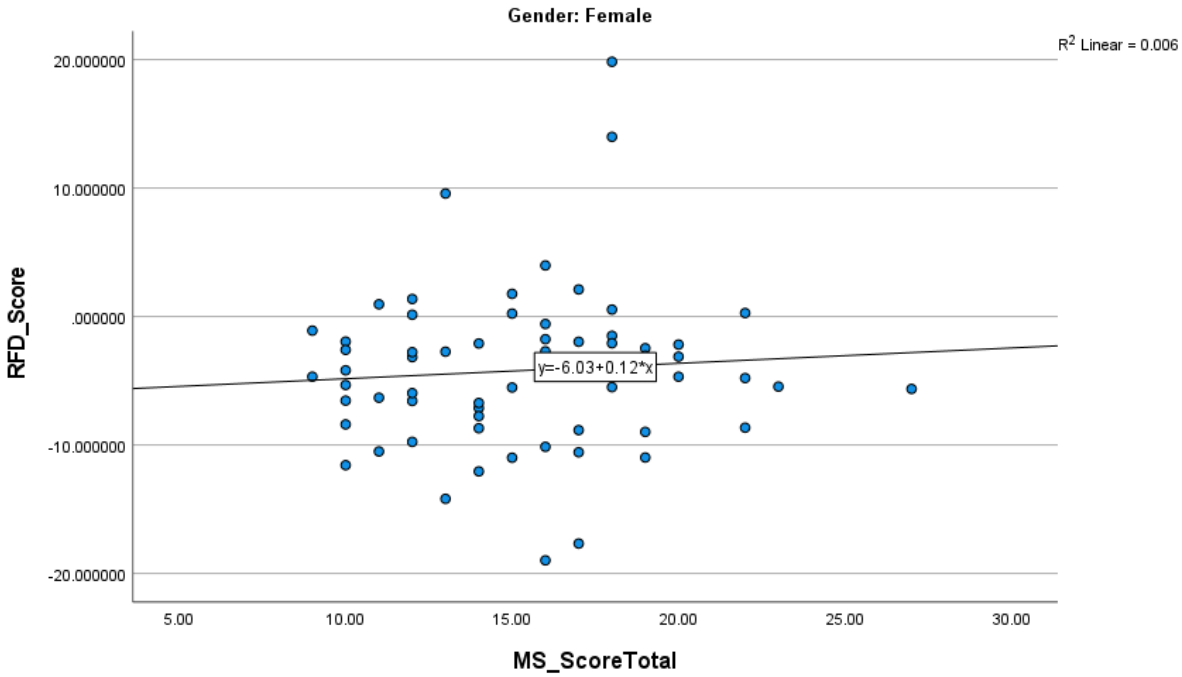


Table 4: *Pearson Product-moment correlations between RFD scores and Modern Sexism scores for female sample*

Scale	**1**	**2**	**3**	**4**
1. RFD Score	-			
2. Modern Sexism Scale	.075	-		

Statistical significance, *p < .05; **p < .01; ***p < .001

2.4.5 Mixed Between-Within Groups ANOVA

A mixed between-within groups analysis of variance was conducted to assess the impact of gender on participant's fluency scores on the FAST across the consistent and inconsistent blocks. There was a significant moderate interaction between block type fluency scores and gender, Wilk's Lambda = .91, F (1, 96) = 9.89, p = .002, partial eta squared =

.093. Due to this, main effects were interpreted with caution. There was a significant, moderate main effect for block type, Wilk's Lambda = .88, F (1, 96) = 12.68, p = .001, partial eta squared = .117, with the combined sample displaying greater fluency on the inconsistent block compared to the consistent block. This indicates an overall bias against males. This outcome appears to have arisen because scores on the inconsistent block were usually higher than scores on the consistent block for female participants. The female participants, as a group, and as observed in the RFD analysis above, appear to a show bias in the opposite direction to that of the male participants. An examination of the plot slopes (Figure 5) supports this conclusion. Specifically, the slopes are not parallel but neither do they cross, which indicates an ordinal interaction (See Figure 5). In effect, while there was an overall main effect across blocks, this was moderated significantly by gender. This being the case, it is likely that the overall main effect (in the unexpected direction, indicating a bias against males) was an effect influenced largely by one, rather than both groups of participants. The second main effect reflected the difference across genders in overall performance on both blocks combined. There was a significant effect, F (1, 96) = 6.85, p = .010, partial eta squared = .067, suggesting there was a moderate difference between genders on FAST block scores. This, however, merely indicates an overall difference in performance speed and accuracy across both blocks combined, and does not in any way indicate a bias in either direction.

Figure 5: *Interaction effect between gender and block type*

2.4.6 Independent-Samples T-Test

Descriptive statistics suggested that males and females differed in their levels of bias against females. This was tested using an independent-samples t-test to quantify differences in MS scores across the male and female participants. As a second independent samples t-test will be run to compare both genders in terms of RFD scores. A Bonferroni adjusted alpha level of .025 (.05/2) was implemented for both tests in order to protect against Type 1 errors. Levene's test was significant for MS scores p = .001, indicating that the assumption of equal variance was violated, therefore the significance values for equal variances not assumed was used in the analysis. There was a significant difference in MS scores for males (M = 20.13, SD = 6.98) and females (M = 15.39, SD = 3.88; t (40.55) = 3.58, p = .001, two-tailed). The magnitude of the differences in the means (mean difference = 4.73, 95% CI: 2.05 to 7.40)

was moderate (eta squared = 0.118). In effect, males and females differed in their levels of bias against females as measured by the Modern Sexism Scale

Due to the fact that the analysis of variance utilised individual block scores within the FAST as an independent variable, it was decided that a comparison of the overall RFD score difference across genders would also be of interest. RFD scores represent a slightly different metric in that RFD scores are a single point of data that represents a degree of bias on the individual level. For this reason, an independent-samples t-test was conducted to quantify differences in RFD score across genders. A Levene's test indicated that the assumption of equal variance for RFD scores was not violated. There was a significant difference in RFD scores for males (M = -.26, SD = 5.12) and females (M = -4.19, SD = 6.16; t (96) = 3.12, p = .002, two tailed). The magnitude of the differences in the means (mean difference = 3.93, 95% CI: 1.43 to 6.43) was moderate (eta squared = 0.09)

Table 5. *Results of the independent samples t-tests for the Modern Sexism scale and RFD scores*

	t	p	Mean Difference	Confidence Interval	Eta Squared
Modern Sexism Scale	3.58	.947	4.73	-2.05 – 7.40	0.118
RFD Score	3.12	.002	3.93	-1.43 – 6.43	0.09

2.4.7 ANCOVA

As earlier tests indicated, gender was found to be a significant variable affecting implicit bias against women. However, given the rather stark gender difference in negative attitudes that was observed here, we might wonder if this reflects recent societal changes. If this is the case, then such changes overtime may be measurable at a single point, by

examining differences in negative biases against women across people of different ages. In other words, it may be that younger participants are less prone to display negative bias than the older participants. If this is the case, it would suggest that younger members of society are growing up with less negative attitudes towards women than their predecessors. To test this idea, a one-way between-groups analysis of covariance was conducted to compare response fluency differences across the consistent and inconsistent blocks, while controlling for participant age. This ANCOVA would allow for the quantification of the contribution participant age makes to the biases observed in this study.

For the purpose of this analysis, the independent variable was the block type (i.e., consistent or inconsistent), and the dependent variable was the fluency scores on each block. In other words, this analysis consisted of all fluency scores across both blocks for all participants combined, while holding the age of participants constant.

Preliminary checks were conducted to ensure that there was no violation of the assumptions of normality, linearity, homogeneity of variances, homogeneity of regression slopes, and reliable measurement of the covariate. The results of the Kolmogorov-Smirnov test indicated that the assumption of normality was violated ($p < .001$), an examination of the histogram indicated a left skew (higher fluency scores). Levene's test was not significant, indicating that the assumption of homogeneity of variances was met. In terms of linearity, and homogeneity of regression slopes, the scatterplot suggests these assumptions were not violated. As these findings would not seriously compromise the procedure, the analysis proceeded as planned.

After adjusting for age, there was a significant difference between the consistent and inconsistent block, in the unpredicted direction $F (1, 193) = 8.99$, $p = .003$, with a small effect size as indicated by the partial eta squared = .044. While the effect size is modest, it must be

noted however that this may be a result of the disparities in the direction of bias across male and female participants. Specifically female participants in this study tended to display a large difference in rate of functional response class acquisition across the blocks, in a direction suggesting anti-male or pro-female bias between blocks. Males tended to show precisely the opposite effect. In other words, the modest difference in fluencies across the blocks is not easily attributable to a random distribution of performances across participants, and the absence of any particular bias across the participant cohort, but rather to a non-random distribution of performances across the males and females, leading ultimately to an overall small effect in response fluencies across the blocks. This effect could be demonstrated by removing the valence of the FAST scores from the raw data in an examination of differences alone, rather than, as was done here, to examine valanced (and therefore socially meaningful) differences in fluencies across the blocks. Interestingly, while an overall pro-female bias persisted even when controlling for the age of participants, a significant, relationship between age and fluency scores differences across blocks was observed, ($p < .000$), partial eta squared = .075 (medium). Expressed as a percentage, 7.5% of the variance in fluency scores was explained by age, indicating that age by itself was a determinant of fluency score in this instance. However, this does not appear to indicate that gender bias has significantly varied here with age. Rather an inspection of the raw data suggested that, as age increases, fluency scores on *both blocks* decrease, relatively equally. In effect, older participants may not differ from younger participants in their levels of gender bias but are simply slower in completing the FAST. Nevertheless, when both block differences and age are considered together, both factors accounted for approximately 11% of the variance in fluency scores, R squared = .113 $F (2, 193) = 12.29$, $p < .001$.

2.5 Discussion

The current experiment sought to assess the FAST's utility in the real-world context of assessing attitudes towards gender. A known-groups paradigm was employed wherein a sample of men and women were compared in terms of their implicit FAST scores and an explicit measure of sexism. The explicit and implicit measures were also correlated against each other to establish the convergent validity of the FAST. The descriptive statistics suggested that males and females scored relatively differently in terms of the RFD index, but in an unexpected way. That is, females scored negatively on the FAST (M = -4.19, indicating a female positive bias), while male scores clustered towards the middle of the range nearing a score of zero (-.26 indicating no particular bias). The groups' divergence in RFD score was confirmed with a t-test. On the explicit MS scale, a similar divergence between males (M = 20.13) and females (M = 15.39) was observed, with their scores indicating that men were more sexist in their self-reported views. This difference turned out to be statistically significant across the groups. In addition, a correlation between RFD scores and MS scores was observed, indicating that explicitly reported and implicitly measured verbal relations co-varied (though this was not found when men and women were examined separately). This might be interpreted as lending some convergent validity to the FAST method.

The analysis of variance indicated that there was a significant difference in fluency between the consistent and inconsistent block indicative of an overall pro-female bias. The ANOVA, however, also showed that there was a modest interaction effect between the block difference and gender, and therefore suggests that the overall group effect is accounted for by the strong pro-female bias demonstrated by the female participants. While age was not a significant covariate of the block difference in fluency scores, age explained as much as 7.5%

of variance in fluency scores, indicating that functional response class acquisition was slower as participants aged.

While it was suspected that males and females might both self-report low sexism, and that females might show either an implicit preference for men, or scores indicating lower ingroup favouritism than men, this was clearly not the case. In fact, it was generally found that females both self-reported low sexism and displayed an implicit *ingroup* bias on the FAST. Males on the other hand, reported a higher degree of sexism on the explicit measure and showed no bias in either direction on the FAST. Taken together, these results indicate that despite their lower social status relative to men, women did not hold a bias in favour of men, either explicitly or implicitly. Thus, the findings do not sit well with the predictions of SJT. In light of the results of Experiment 1, it would seem worthwhile to revisit the SJT literature to consider these findings.

The SJT literature makes an important distinction between prevalent forms of sexism that are not hostile but rather benevolent in (i.e., holding positive/paternalistic/protective attitudes towards women). While not overtly sexist in the classic sense, nonetheless benevolent sexism should be considered prejudicial (Jost et al., 2004). It has even been argued that some women may hold positive views of men who are benevolently, but not hostilely sexist (Kilianski & Rudman, 1998). In other work, Jost (1997) has shown that some women exhibit depressed entitlement in self-ratings of how much their work is worth monetarily (i.e., they valued their work less than a sample of male participants, despite independent examiners rating them equally). Such subtle, insidious forms of outgroup favouritism may not have been uncovered in the present experiment due to the stimulus sets employed.

Specifically, because the FAST employed in this experiment involved the use of binary evaluative stimulus classes (i.e., good and bad), it could only capture overt hierarchical preferences for one gender over the other. The possibility that the decision to employ binary stimulus classes in some way influenced the results should be considered. While it was presumed that the stimuli would be interpreted comparatively, it is unknown what effect employing stimuli which were literally comparative (e.g., better than, worse than) would have on participants responding. However, in its current form, incorporating relational components would be a cumbersome addition to the FAST methodology. The FAST presents only a single stimulus item at a time, therefore using relational comparative terms would be confusing for participants, as those terms always refer to a relation between two stimuli. The only way to incorporate relational components into the procedure would be to present two stimuli simultaneously, a significant modification which would require extensive laboratory testing before real world applications such as the current study. As it stands the evaluative stimuli are likely being responded to comparatively, as across a number of trials it should be apparent that they are "opposite". In its present form the FAST measures class compatibilities i.e., whether female stimuli are more compatible with positive or negative verbal stimulus classes. While this approach is straightforward, fully functionally understood, and allows for expedient testing, it does lack the capabilities of more lengthy procedures such as the IRAP where a specific relation between the stimuli could be assessed.

It could be argued therefore, that the FAST might have failed to capture the subtle verbal relations that compose benevolent forms of sexism, that apply even when males do not believe that on the whole men are in some sense generically better than women. In addition, women may hold a strong general ingroup bias in simple valanced terms, but may at the same time verbally classify stimuli related to certain traits (nurturing instincts), or behaviours (submissiveness), that are reflective of oppressive stereotypes. However, some or many

women display positive attitudes towards benevolently sexist men (Kilianski & Rudman., 1998). This may have been captured if the evaluative stimuli used here were replaced with stereotypical personality traits typical of men and women coexisting in a power relationship in relation to each other.

A study which was missed in the initial literature view of SJT may shine further light on the paradoxical findings of Experiment 1. Rudman and Goodwin (2004) hypothesised that gender attitudes may be an exception to the usual SJT finding of outgroup bias among discriminated against groups. Across several experiments, it was their finding that in contrast to SJT, women generally held a strong implicit in-group bias, and men generally held a weak ingroup bias. Using a variety of different IAT's (gender identity, parental preference, gender threat and sexual attitude) across multiple experiments they attempted to delineate the factors which cause this effect. It was their finding that an implicit preference for their mother over their father was correlated with implicit preference for women. Women also generally held a stronger female gender identity, another factor associated with implicit preference for women. Implicitly both men and women find men to be the more threatening gender, a finding which was linked with greater implicit pro-female evaluations. Finally, the sexual attitude IAT revealed that men implicitly like sex more than women and that this may influence gender attitudes. Specifically, Rudman and Goodwin found that women who implicitly liked sex also favoured men more implicitly. Men who were higher in sexual experience showed a correlation between their implicit positive sexual attitudes and their implicit gender attitudes, while the inverse was found for men with low sexual experience. In short, a liking for sex was a predictor for decreased in-group bias among sexually experience men. The findings of Rudman and Goodwin (2004) would seem to suggest that gender attitudes are a notable exception to the usual predictions of SJT concerning discriminated against groups. It is also worth noting that Rudman and Goodwin attempted to employ verbal

stimuli which were gender neutral in their IAT's, something which was not done in the current study, this too may have influenced the present findings.

In addition to the above, there is evidence to suggest that the lower the social status of a minority group, the greater the implicit outgroup favouritism they display will be, and the greater the gulf will be between implicit and explicit measures (Rudman et al., 2002). With this in mind, it could be argued that a study on gender bias might not have been the ideal starting point for a behaviour-analytic assessment of SJT. Although women are a social group who face discrimination, they nonetheless remain one of relatively high social status in comparison to several other discriminated-against groups. Therefore, knowing that implicit outgroup favouritism is likely to be found in groups of low social status (Rudman et al., 2002), we should return to the design of the study. Following the logic of Rudman et al., if we wish to maximise our chances of detecting system-justification effects we should employ a minority group of relatively low social status. As a country with a highly homogenous Irish-White population, it would stand to reason that ethnic minorities would fit this criterion. We need not take this at face value of course. In a large-scale survey, as much as 17% of respondents from a Sub-Saharan background in Ireland reported facing discrimination when looking for work (European Union Agency for Fundamental Rights, 2017). Additionally, the same survey showed that, 38% of respondents from this ethnic background had experienced hate-motivated harassment, these are among the highest rates in Europe. For these reasons, Experiment 2 will consist largely of a replication of Experiment 1 but in the context of race rather than gender.

Chapter 3

Assessing the FAST as a Means of Detecting Racial Bias in Sample of White and Non-White Individuals

3.1 Introduction

At the end of Experiment 1, it was reflected that perhaps the predictions of SJT concerning women displaying outgroup favouritism towards men was somewhat more ambiguous and lacking in empirical support than had been assumed (See Rudman and Goodwin, 2004). In the racial domain, however, SJT theorists are quite overt in their predictions (See Jost et al., 2004) and research outcomes that align with predictions in this domain would benefit from more confident assessments of convergent validity.

As outlined in Chapter 1, there is a plethora of studies showing that ethnic minorities will display an outgroup bias on implicit measures, despite self-reporting a strong ingroup bias. Nosek et al. (2002) demonstrated precisely this effect in a comparison of African Americans and European Americans. In conducting secondary analysis of large samples of data taken from the public website projectimplicit.com, it was shown that African Americans display implicit outgroup favouritism towards the more dominant European American group. Taking data from the same source (projectimplicit.com), at a later date, Jost et al. (2004) showed that almost 40% of African Americans demonstrate outgroup favouritism on an IAT. This was despite approximately 65% of African Americans from the same sample demonstrating ingroup favouritism on explicit self-reports.

While it can be said with some confidence that minority ethnic groups will display outgroup favouritism when measured according to the IAT, it is only fair to be diligent and examine other experimental situations where outgroup favouritism manifests itself. Correll et al., (2002) conducted a series of experiments using a hypothetical shooting scenario. Specifically, they had their participants complete a game wherein they had to decide whether to shoot or not shoot under time pressure. Participants would be presented with images of potential shooters of either White or African American ethnicity holding a variety of objects,

most of which were neutral but bore a passing resemblance to a gun (e.g., a silver camera, a black phone), but some of which *were* handguns. Participants had to decide whether to shoot or not shoot by pressing a specific key for each option on the keyboard. A point system was implemented, wherein participants were rewarded for correctly shooting an armed individual, but penalised for shooting an unarmed individual, and heavily penalised for not shooting an armed person. Of specific interest to this thesis, in experiment 4 they compared White and Black individuals on this game. They found that for both the White and Black sample, participants were quicker to shoot an armed African American than an armed White person. Similarly, participants were slower to select the don't shoot option when presented with an unarmed African American than an unarmed White person. The results of this study indicate that even African Americans themselves implicitly favoured White people, in the sense that they deemed them less threatening/hostile than African Americans. Now that we can safely say that outgroup favouritism among minority ethnic groups can be found in a variety of domains, it is time to detail how this will be measured in the present experiment.

As outlined at the end of Experiment 1, different types of valanced stimulus classes can be employed as attribute stimuli in these types of implicit tests. Different choices of valanced stimuli could lead to different test outcomes depending on the nature of the bias being assessed. It is clear from Experiment 1 that there is no overall general anti-female bias in the general population of males and females combined, at least in the Irish research context. While the choice of stimuli employed in Experiment 1 arguably rendered the FAST less sensitive to particular and more subtle forms of sexism than had been anticipated, such issues are not expected in Experiment 2. Racial prejudice would appear to be more simply evaluative and generic, and less subtle than sexism. Of course, as outlined in Chapter 1, the issue of racial bias has been studied quite extensively using the IAT. The current research

will employ stimuli drawn from that early IAT research work and will include binary valanced words as evaluative stimuli.

In Experiment 2 White and Non-White individuals will be recruited to complete explicit measures of racial bias, as well as a FAST configured to measure racial bias. A demographic survey (See Appendix B) administered prior to testing will assess the subjects' age, gender, ethnicity and residency in the Republic of Ireland. Their FAST performances and scores on self-report measures will be analysed in the same way as in Experiment 1. The explicit measures to be employed are the Discrimination and Diversity Scale (Wittenbrink et al., 1997; See Appendix G) and the Modern Racism scale (McConaghy 1986; See Appendix H). The Discrimination and Diversity scale is split into two subscales, the discrimination scale is designed to measure overt prejudice against ethnic minorities, such as the degree to which the participant denies that racism is a problem. The Diversity scale places more emphasis on culture and measures the degree to which the participant thinks that the integrating of other ethnic groups into their country is problematic. The Modern Racism scale is orientated more towards subtle manifestations of racism rather than classic overt White supremacy e.g., the degree to which the test-taker believes that Black people are overrepresented in media. The Modern Racism scale has been modified slightly in that references to America have been replaced with Ireland instead.

Interestingly, the FAST does not need to be modified extensively for this experiment. Importantly, the same stimuli as used in Experiment 1 can once again be employed, but classes can be reconfigured such that the functional responses classes being trained involves the establishment of compatible and incompatible positional keyboard responses for visual images of White and Non-White faces (and positive and negative evaluative stimuli), rather than for male and female faces. It is once again predicted that an overall bias against Non-

White individuals will be demonstrated by the cohort as a whole, and that this will be observable for both groups considered separately. Owing to the SJT rationale that implicit test results often contradict explicit results among minority groups in terms of outgroup bias a correlation is not expected between the explicit self-reports of racial bias, and scores on the FAST test. It is worth noting however that whether implicit and explicit results will correlate varies widely in the literature, some correlation might be found even where implicit and explicit results may diverge. For example, even if it was found that the FAST revealed an implicit outgroup bias despite an explicit ingroup one, it may still be the case that the strength of the implicit outgroup bias could be weaker for those participants with a stronger explicit ingroup bias.

3.2 Method

3.2.1 Participants

Fifty-two participants (24 identified as male, 28 as female) aged between 18 and 55 participated in this study (Mean age = 22.5, SD = 6.085). Twenty-seven participants identified as White Caucasian, and 25 identified as Non-White of various ethnicities. For the purpose of these analyses, and given a range of ethnicities identified by participants in this relatively small sample, participants were divided into two groups based on whether or not they identified as White or Non-White. There were too few exemplars of several ethnicities in order for these groups to be studied individually. Collectively these groups will all function as a discriminated against group. Participants were predominantly recruited from the student cohort of Maynooth University. However, a snowballing recruitment method was employed, wherein participants were encouraged to share the study information and access link with others. An open-ended response format for reporting gender was employed so as not to impose to a binarized forced-choice. Participation was voluntary. However, a course credit

was available to some participants who were currently enrolled in one module of an undergraduate Degree in Psychology at Maynooth University. Inclusion criteria included fluent English, normal or corrected-to-normal vision, full use of both hands and residence in the Republic of Ireland.

3.2.3 Apparatus

The Modified Modern Racism scale, Discrimination and Diversity scale and the Function Acquisition Speed Test were administered using Inquisit software hosted on the millisecond.com European server.

3.2.3.1 *Modified Modern Racism Scale*

The MMRS is an adaptation of McConaghy's (1986) original scale to the Irish context (See Appendix H). The MMRS is composed of 6 statements, to which participants respond on a 5-point scale from "Strongly Disagree" to "Strongly Agree". As the name suggests, the MMRS was designed to measure subtler, modern manifestations of racism (e.g., "Over the past few years the government and news media have shown more respect to blacks than they deserve"). The MMRS is scored on a scale of -2 to +2. The minimum score was -10 and the maximum +10. Higher scores indicated more racist attitudes.

3.2.3.2 *Discrimination and Diversity Scale (DDS)*

The Discrimination and Diversity Scale (See Appendix G) was designed to measure racial discrimination (Wittenbrink et al., 1997). It is comprised of 14 statements, split into two subscales. These scales are traditionally scored from 1/Strongly Agree to 5/Strongly Disagree. However, the scoring system was reversed to 1/Strongly Disagree to 5/Strongly Agree, in order to coincide with the direction of the MMRS (i.e., higher scores indicate more racist attitudes). The Discrimination scale contains 10-items, and is intended to measure

discrimination against ethnic minorities, in particular those identifying as Black (e.g., “Black people often blame the system instead of looking at how they could improve their situation themselves”). The maximum possible score was 50, and the minimum 10. The Diversity scale contains 4-items, intended to measure overt attitudes towards ethnic diversity (e.g., “There is a real danger that too much emphasis on cultural diversity will tear Ireland apart”). The maximum possible score was 20, and the minimum 4. As this scale was designed for an American context, the phrase *United States* was replaced with *Ireland* in three instances.

3.2.3.3 *Function Acquisition Speed Test*

The FAST administered in Experiment 2 was identical in methodology and scoring to that used in Experiment 1. Importantly, for purposes of conceptual and empirical coherence, and to eliminate differences across the experiments aside from the attitudinal dimension being assessed, the same stimuli were employed as were employed in Experiment 1. However, the contingencies employed in each training block were changed so that compatibilities and incompatibilities between ethnic (rather than gender) were assessed across the two FAST blocks. This was easily achieved using the same stimulus set as employed in Experiment 1 because half of the male and half of the female facial images employed depicted black African Americans and half depicted Caucasians. Therefore, in the current FAST preparation, and assuming a general racial bias against Black individuals, the consistent block was designed to establish common response functions for images of White people and positive stimuli, as well as common responses for images of Black people and negative stimuli. In contrast, the inconsistent block was designed to established common response functions for images of White people and negative stimuli, as well as common responses for images of Black people and positive stimuli. The instructions provided before the first block and in the interval between blocks was the same as for Experiment 1.

3.3 Procedure

Experimental sessions were conducted in the same way as Experiment 1. However, the information sheet (See Appendix I), consent form (See Appendix J) and study advertisement (See Appendix K) were of course modified to reflect the interest of the experiment in ethnic bias rather than gender bias. (The experimental tasks were completed in the following order: 1) demographic questionnaire, 2) Discrimination and Diversity scale, 3) Modern Racism Scale, 4) FAST. Like Experiment 1 the FAST was presented following the self-report measures, and for the same reasons. The debriefing information outlined the thesis behind the experiment (See Appendix L).

3.4 Results

3.4.1 Excluded Cases

A portion of the participants were excluded from the analysis due to not completing one of the self-report measures or the FAST. Additionally, if a participant scored 0 on any block of the FAST (i.e., indicating no responding at all) their data was excluded. On the bases of these criteria a total of 8 participants were excluded from the analysis.

3.4.2 Descriptive Statistics

The means and standard deviations for White and Non-White participants on the FAST blocks, RFD score, DDS and MMRS are provided in Table 6 below. There was little difference between White and Non-White participants on the DDS. It is worth noting that the mean scores of subjects on the DDS fell in the middling range of possible values, indicating neither high nor low racism according to this scale. Non-White participants generally scored lower on the MMRS indicating less racial bias, though both groups averaged scores on the lower end of the scale. In line with the hypothesis, Non-White participants generally had a

positive RFD score (indicating a White positive bias), though on average they scored slightly lower than White participants.

Table 6. *Means and standard deviations for FAST blocks, RFD scores, Modern Racism and Discrimination & Diversity scales*

	Ethnicity	N	Mean	Standard Deviation
Diversity Scale	White	27	13.00	3.56
	Non-White	25	13.08	4.89
Discrimination Scale	White	27	32.88	10.89
	Non-White	25	31.88	14.02
Modern Racism Scale	White	27	-7.63	2.34
	Non-White	25	-8.96	2.84
RFD Score	White	27	2.68	6.95
	Non-White	25	1.56	6.72
Consistent Block Fluency	White	27	19.38	4.84
	Non-White	25	16.94	7.76
Inconsistent Block Fluency	White	27	16.72	5.13
	Non-White	25	15.38	7.3

3.4.3 Correlations

The relationship between RFD score and the explicit measures (Modern Racism scale, and the Discrimination and Diversity subscales) was investigated using a Pearson product-moment correlation coefficient. Preliminary analyses were performed to ensure no violation

of the assumptions of normality, linearity, and homoscedasticity. There was no correlation between RFD scores and Modern Racism scores, between RFD scores and Discrimination scores, or between RFD scores and Diversity scores (See Table 7; Figures 6, 7, and 8). The two subscales of the DDS correlated strongly, though this was to be expected given their being two related dimensions in a standardised questionnaire. In short, there was no relationship between FAST scores and any of the self-report measures.

Table 7: *Pearson Product-moment correlations between RFD scores and explicit measures*

Scale	**1**	**2**	**3**	**4**
1. RFD Score	-			
2. Modern Racism Scale	.226	-		
3. Discrimination Scale	.171	-.019	-	
4. Diversity Scale	.122	.124	.923***	-

Statistical significance, *p < .05; **p < .01; ***p < .001

Figure 6. *Scatterplot of RFD scores by Modern Racism scores*

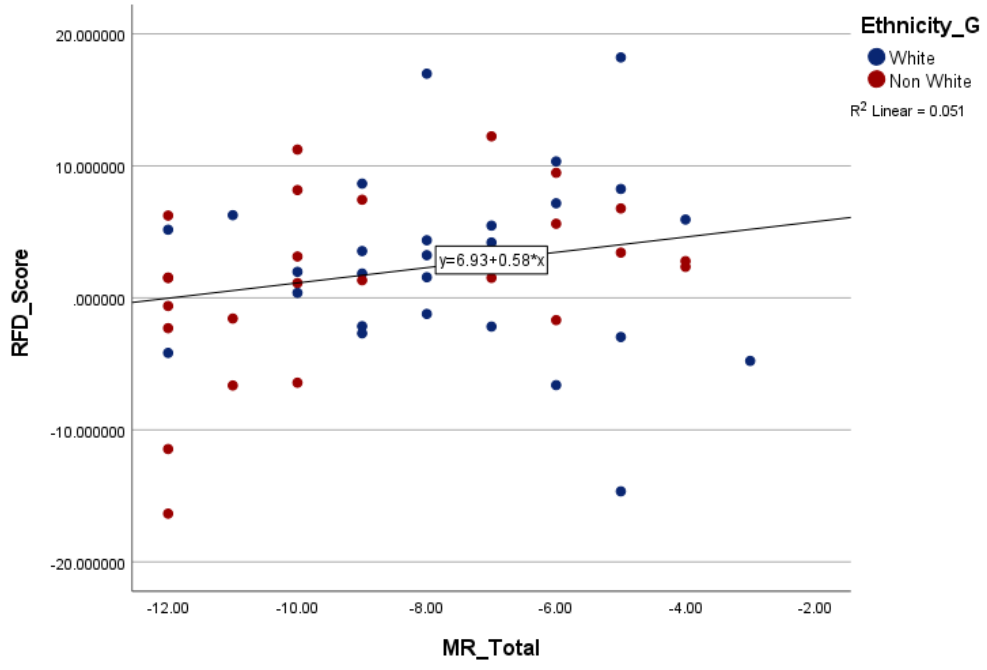


Figure 7. *Scatter plot of RFD scores by Diversity scale scores*

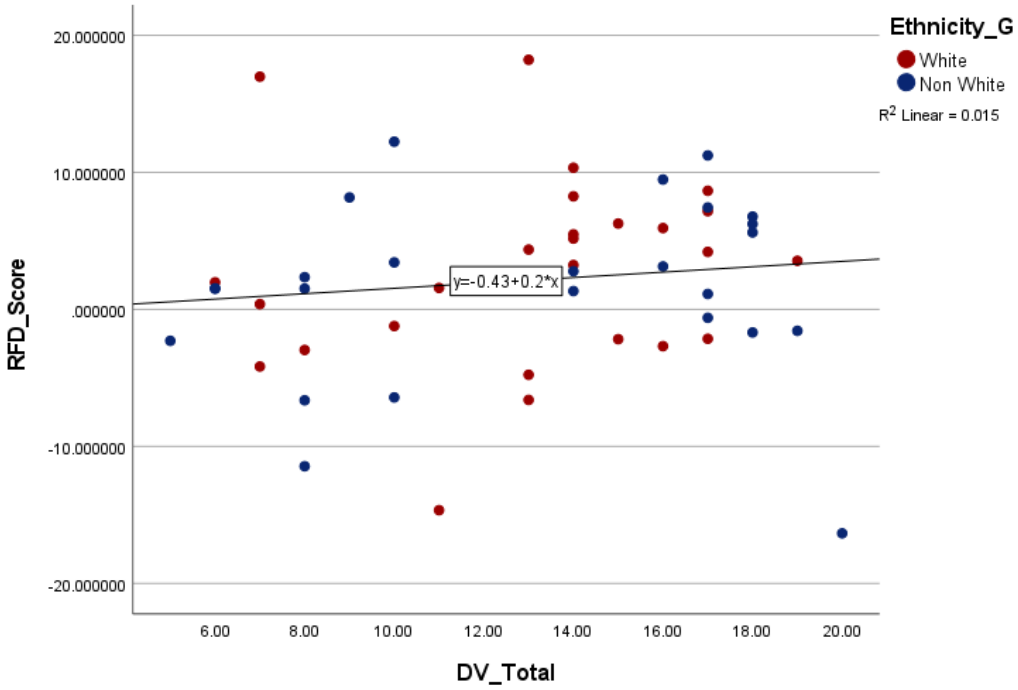
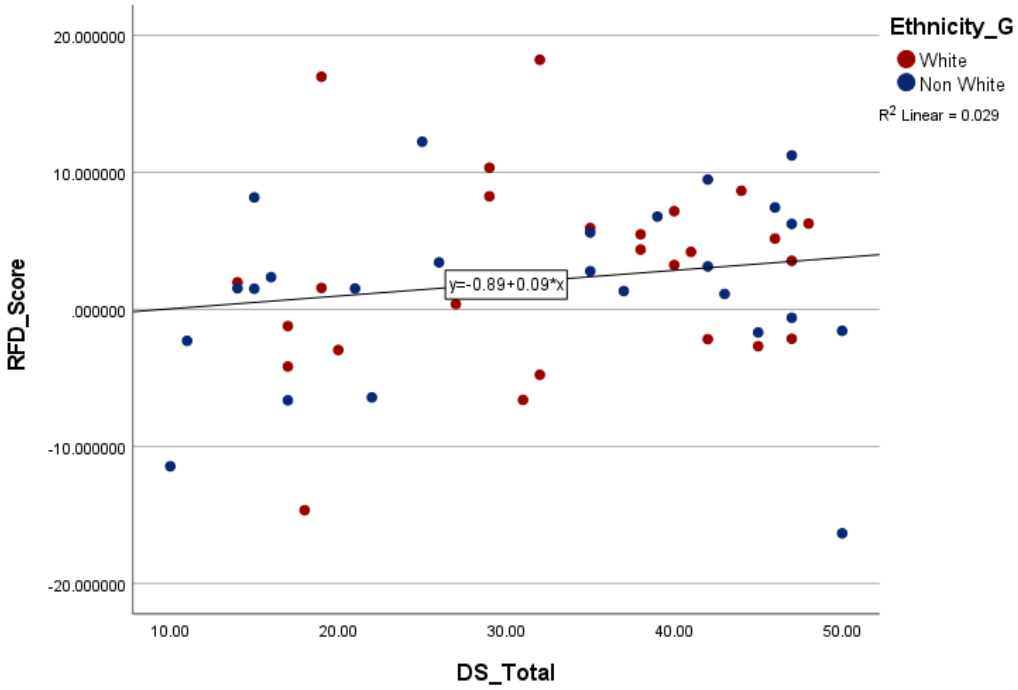


Figure 8. *Scatter plot of RFD scores by Discrimination scale scores*

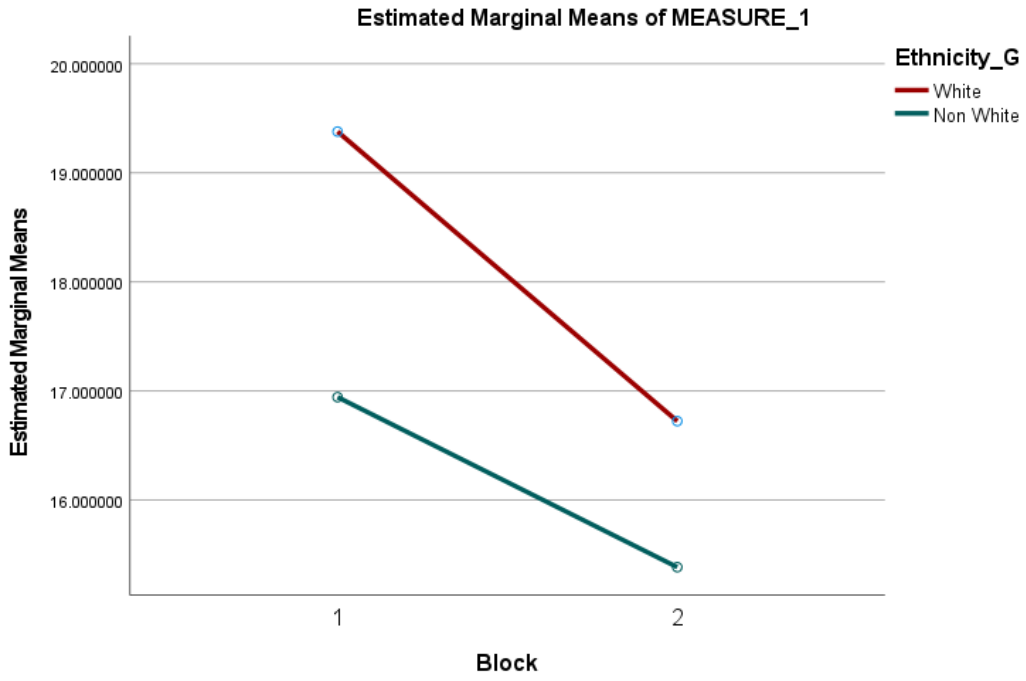


3.4.4 Mixed Between Within Groups ANOVA

A mixed between-within groups analysis of variance was conducted to assess the impact of ethnicity on participant's fluency scores on the FAST across the consistent and inconsistent blocks. There was no interaction effect between block type fluency scores and ethnicity, Wilks' Lambda = .99, F (1, 50) = .334, p = .57. There was a significant moderate main effect for block type, Wilks' Lambda = .91, F (1, 50) = 4.93, p = .031, partial eta squared = .09, with both groups showing a reduction in fluency score on the inconsistent block compared to the consistent block (See Table 8). This reduction in score on the inconsistent block suggests an overall bias against non-white groups for the participant cohort as a whole. An examination of the plot slopes (see Figure 9) indicates that while Non-White participants generally scored lower on both blocks, both White and Non-White participants exhibited a parallel decrease in scores from the consistent to the inconsistent block. The main effect comparing the difference in overall (combined) block scores across ethnicities was not significant, F (1, 50) = 1.62, p = .208, partial eta squared = .031, suggesting no difference between White and Non-White participants on FAST block scores combined. This suggests that there was no difference across the two groups in the rate of functional response class acquisition.

In summary, bias in the rate of functional response class acquisition was measured for both groups and was found to be non-different across both groups. In other words, the ethnic bias measured by the FAST in the current study was found for both groups.

Figure 9. *Interaction effect between ethnicity and block type*



Note. 1 = Consistent block scores, 2 = Inconsistent block scores

Table 8. *Means and standard deviations for White and Non-White participants across consistent and inconsistent blocks*

	White			**Non-White**		
Block Type	N	M	SD	N	M	SD
Consistent	27	19.37	4.83	25	16.94	7.76
Inconsistent	27	16.72	5.12	25	15.38	7.3

3.4.5 Independent-Samples T-Test

While the descriptive statistics suggested that White and Non-White participants differed little on self-report measures, there was some small difference in RFD score, despite scores trending in the same direction. In order to investigate these relationships more closely, an independent-samples t-test was conducted to compare the RFD scores and the explicit measure scores for White and Non-White participants. A Bonferroni adjusted alpha level of .0125 (.05/4) was implemented for all tests in order to protect against Type 1 errors. There was no significant difference in scores between the two groups on any of these dimensions (See Table 8). The results of Levene's test indicated that variance was unequal for the Diversity scale $p < .05$, and the Discrimination scale $p < .05$, therefore the values for equal variances not assumed was reported for these variables below (See Table 9). The magnitude of the differences in the means was almost non-existent for the DDS, moderate for the MMRS and small for the RFD scores (See Table 8). In short, there was little difference between White and Non-White participants along any of these indices, thereby supporting the main system justification hypothesis in the context of ethnic biases.

Table 9. *Results of the independent samples t-tests for the Discrimination and Diversity scales, Modern Racism scale and RFD scores*

	t	p	Mean Difference	Confidence Interval	Eta Squared
Diversity Scale	-.067	.947	-.080	-2.487 – 2.328	0.000
Discrimination Scale	.288	.774	1.009	-6.038 – 8.056	0.001
Modern Racism Scale	1.851	.070	1.330	-.113 – 2.774	0.064
RFD Score	.589	.559	1.117	-2.695 – 4.929	0.006

3.4.6 Bayesian Analysis

Although, an independent samples t-test indicated that the White and Non-White groups did not differ in terms of RFD scores, this does not indicate statistical equivalence. That is, the absence of evidence for difference is not the same as evidence of equivalence. While system justification theory does not claim explicitly that a low and high-status group will be equivalent in their favouritism for the high-status group, it seems to be implied in the literature that the absence of a difference is being taken as the presence of an equivalence. Given that no study to date has actually tested for equivalence in biases towards the more powerful group, but instead has inferred such equivalence from the absence of a difference, it was decided that a Bayesian analysis should be conducted in order to test for equivalence in bias scores, and to more carefully examine the alignment between these findings and an ambiguous prediction of SJT.

A Bayesian inferential statistical analysis was run in order to assess for equivalence between the two groups in terms of RFD score. Bayesian analyses are conceptually distinct from the frequentist analysis being reported thus far in the thesis. While a frequentist approach is used in order to assess the probability of a particular pattern of data given a particular hypothesis and assumption of a null difference, Bayesian analyses examine the probability of the hypothesis being correct given the data (Gelman et al., 2014).

In the current analysis, the Bayesian approach was used in order assess the probability that the RFD scores for the White and Non-White groups were statistically similar or non-different. This form of analysis will produce a Bayes factor, which refers to the relative probability for one model (i.e., the RFD scores of the two groups are similar) compared to another (i.e., the RFD scores of the two groups are not similar). A Bayesian one-way ANOVA was run in order to assess whether a model assuming RFD score varied as a

function of ethnicity was more likely than the null hypothesis. For RFD score, a BF_{10} = 0.32 indicated that these data are approximately 3 times less likely to be observed given the existence of an effect, than given the lack of existence of an effect (the next analysis can also be interpreted in a similar fashion). A Bayesian two-way ANOVA was run in order to assess whether a model assuming RFD varied as a function of block (i.e., consistent and inconsistent block fluency) and ethnicity were more likely than the null hypothesis. For the RFD score in this analysis, the BF_{10} was found to be 0.29. Taken together, these analyses are inconclusive as to whether the groups are statistically similar or different. In short, therefore, while the frequentist statistical analysis did not find these groups to be different, a Bayesian analysis also did not find them not be unambiguously equivalent.

3.5 Discussion

Experiment 2 sought to investigate the FAST's capabilities in the real-world context of assessing attitudes towards race. A known-groups paradigm was employed wherein a sample of individuals from White and Non-White ethnicities were compared in terms of their implicit FAST scores and two explicit measures of racism. The descriptive statistics suggested that the White and Non-White groups scored similarly in terms of RFD score. That is, the White sample scored positively on the FAST (M = 2.68, indicating a White positive bias), the Non-White sample also scored positively (M = 1.56). Similar findings emerged for the explicit measures. On the Modern Racism scale, White participants scored negatively (M = -7.63, indicating low racism) as did Non-White participants' (M = -8.96). Both groups displayed similar scores on both subscales of the Discrimination (White M = 32.88, Non-White M = 31.88) and Diversity Scales (White M = 13.00, Non-White M = 13.08), indicative of low racism.

The analysis of variance indicated that there was a significant difference in fluency between the consistent and inconsistent block, consistent with a pro-White bias at the entire cohort level. Interestingly, however, there was no interaction between the block differences and ethnicity, suggesting a similar pattern of performance across the two groups. The ANOVA also suggested that there was no difference between groups in rate of functional response class acquisition across the two test blocks considered together. In contrast to Experiment 1, this suggests that the pro-White bias found on the FAST was consistent across both groups.

The results of the planned t-test comparisons indicated that the groups did not differ along any of the self-report metrics or on RFD scores. A Bayesian analysis was inconclusive, indicating neither a difference nor a similarity between groups in terms of RFD scores. Furthermore, no correlation between RFD scores and either of the explicit scales was found for the participant cohort as a whole, indicating some divergent validity for the FAST in a context in which we might expect self-reporting to be unreliable. That is to say, according to the wider SJT literature considered in this thesis, a divergence between self-report and implicit bias in this context is to be expected. Therefore, the fact that such a finding *was* found lends some credence to the FAST method as the expected divergence was observed. It must be acknowledged however, that the lack of correlation found between the self-report measures and FAST RFD scores could have been a result of the reduced power in the second study. Specifically, Experiment 2 had a total of fifty-two participants, compared to Experiment one's ninety-eight, therefore it remains possible that with a larger sample size, a correlation between the self-reports and the FAST could have been found. However, it is still worth noting that the mean scores for both measures remain indicative of a general divergence between participant's explicit and implicit scores.

It was expected that both White and Non-White participants would self-report low racism, and that both groups would show an implicit preference for White people. Indeed, precisely these effects were found. Their mean scores on the explicit measures were nearly identical, and in the direction of low racism. On the FAST however, it was found that White participants held an ingroup favourable bias, and that Non-White participants held an outgroup favouritism towards White people. These results taken together cohere strongly with the predictions of SJT. That is, the presumed low status Non-White group explicitly denied a bias against their own group but at the same time demonstrated a positive bias towards the dominant White group at the implicit level. Though it is difficult to directly compare FAST and IAT scores due to differences in scoring methods, the results observed here were broadly in line with those reported by Nosek et al. (2002) and Jost et al. (2004) in their large-scale analyses of implicit outgroup favouritism among ethnic minorities. Specifically, both of these of large-scale analyses found a general pattern of explicit ingroup favouritism and implicit outgroup favouritism among low-status ethnic minority groups.

Chapter 4

General Discussion

4.1 Research Summary and Main findings

The current research sought to employ the novel FAST in a real-world context to examine one of the main hypotheses of system justification theory. Specifically, that marginalised groups will display an implicit favourable outgroup bias, not necessarily found on the explicit level. This theory was examined in the domains of gender bias and racial bias. The FAST was used as an implicit measure to compare and contrast with scores calculated for several self-report measures. A known-groups paradigm was employed, wherein, in Experiment 1, men and women were compared in terms of strength of favourable biases towards the dominant group in the domain of gender bias. In Experiment 2, the FAST performances of participants identifying as either White or Non-White were compared in the same way in the context of racial bias. It was expected that marginalised groups in each experiment (i.e., women and Non-White participants), would display an implicit outgroup favouritism in favour of the dominant group, a favouritism that might not be revealed in the explicit domain. Thus, it was expected that explicit measures would not correlate with FAST scores. The null hypothesis for system justification was not rejected in Experiment 1 (i.e., women did not display an implicit outgroup favouritism towards men). Paradoxically, according to the FAST, women participants generally had an in-group bias, while men were generally absent in favouritism in either direction. Additionally, there was almost 4 points in the mean difference between men and women in terms of RFD scores. The explicit self-reports in Experiment 1 also indicated low gender bias among the women participants. The average score for male participants was almost 5 points higher than women on the MS scale, suggesting they had greater bias against women in the explicit domain. The null hypothesis for system justification was rejected in Experiment 2. That is, both White and Non-White groups did not display significant racial bias on the explicit measures, nor was there much difference between the two groups on any of the explicit self-reports. However, their scores

on the FAST showed a White-positive favouritism for both groups. Although RFD scores were slightly lower on average for the Non-White group, the mean difference between the two groups was only 1.1, a relatively small difference between groups in comparison to Experiment 1. The results of these experiments will now be examined in the broader context of implicit testing research and system justification theory. Future avenues for research will then be considered.

4.2 Summary of Results

Although no correlation between FAST RFD scores and the MS scale was expected, in light of the descriptive data patterns, the finding of a small but positive correlation between the two metrics was perhaps not ultimately surprising. Although this correlation disappeared when men and women were examined separately, ultimately this may be an issue of lower power due each correlational analysis now working with a smaller sample. Additionally, whether a correlation between the two metrics is present or not, the point is still clearly observable from the descriptive statistics. On average men were more sexist explicitly, but neutral implicitly, and women were less sexist explicitly and strongly favoured their own group implicitly. Upon further investigation it emerged that men and women differed significantly in terms of MS scores, and the same finding was found for RFD scores. This provided some confirmatory evidence for what seemed apparent in the visual analysis of the raw data trends. That is, women displayed ingroup favouritism on both explicit and implicit measures. According to the hypothesis, men and women should have displayed a similar pattern of RFD scores, or, at the very least, women should have displayed diminished ingroup bias relative to the male ingroup bias. In particular, implicit outgroup favouritism (towards men) was expected for female participants, and an implicit ingroup favouritism was expected for men. In light of the RFD scores indicating an ingroup bias among women, an absent

finding for men, and no difference between groups on either metric, further investigation of these results was warranted.

A mixed between-within groups ANOVA was conducted to more closely examine the impact of gender on participants' fluency scores across the consistent and inconsistent block on the FAST. This analysis allowed for a more in-depth examination of the effects of gender, and block type (consistent/inconsistent), on all fluency scores taken separately. The results of this analysis revealed that block type had a moderate influence on fluency scores for both groups combined (main effect). As a reminder, the inconsistent block in this FAST assigned common responses for female and good stimuli, as well as male and bad stimuli. Fluency was higher on the inconsistent block compared to the consistent, meaning that subjects generally responded quicker, and with more accuracy on the former block. While on the face of things, this suggests an overall female positive bias, the plot slopes, coupled with the foregoing analysis, suggest that men hardly differed at all in fluency scores across both blocks. The ANOVA also indicated that gender in and of itself had a significant impact on fluency scores on both blocks combined. To be clear, it was found that on both blocks, men responded more slowly, and with more errors than women. This is a finding that warrants further discussion (See Potential Methodological Issues section below).

In an effort to make sense of this unexpected finding, it might be considered that given relatively recent social leaps in gender equality in Ireland, it may not be surprising after all if a random sample of females do not show a pro-male and anti-female bias (as captured simultaneously and inseparably on the FAST). While there is no way to test this hypothesis directly within the context of the current study, there is an analysis which may provide a clue as to the validity of such an interpretation. That is, if older participants had performances more in line with the hypothesis, then the observed findings might be explained in terms of

growing tendencies towards gender equality in Ireland more acutely experienced by a younger generation. Fortunately, there was sufficient variance in age in the present sample to examine this question more closely. An ANCOVA was employed to assess differences in fluency across the two blocks for all participants combined, while controlling for participant age. The results of this analysis indicated overall levels of bias did not vary across age. However, this analysis was not entirely unfruitful. It revealed that despite the direction of group favouritism not varying significantly by age, there was an overall reduction in response fluency across both blocks among older participants. That is, as much as 7.5% of variance in fluency score was accounted for by age. While not relevant to the experiment at hand, this finding nonetheless warrants some commentary and explanation.

While several explanations could be given for the finding that older participants were slower at the task overall, the one with the most face validity is simply that older participants may be generally less competent in the use of the relevant technology. They may also have less rapid motor responses and may already be beginning to experience some slight cognitive decline, at least relative to the much younger participants. A reduction in the speed of processing among older age groups is entirely in line with the results from the reaction time testing literature. Der and Deary (2006) conducted a large-scale study to examine how reaction times change with ageing. They administered two reaction time (RT) tasks, a simple task and a choice task. The simple task merely involved pressing a single button whenever a specific stimulus was presented as quickly as possible. The choice task featured multiple buttons, and participants had to select a specific button in the presence of a specific stimulus. It was their finding that simple RT tasks only began to increase (meaning slower responding) at around the age of 40, while choice RT tasks exhibited a steady increase with age. As the FAST requires choosing a response it would fall into the choice task category, and thus a steady increase in RT tasks with age should not be surprising. Even if it were to be argued

that the FAST is relatively simplistic task, 7.1% of the sample in the gender study were over the age of 40, thus, the fact that some variance with age was detected is an entirely normal finding. This finding is common enough that IRAP research has begun to account for it at the methodological level. Specifically, IRAP research has acknowledged higher attrition rates among older participants (Cabrera et al., 2020). Cabrera et al. attempted to combat this in a study of older age dementia caregivers (i.e., they themselves did not have dementia but were simply an older sample). In their study, they adapted the response latency and accuracy criteria to reduce the difficulties of an IRAP task for the older age group. The particular population the FAST is being administered to in an experiment should be carefully considered in light of this.

In the present sample, it could be argued, that younger participants had an additional advantage, in that, the FAST could be considered vaguely similar to other tasks younger participants may be more familiar with in their daily routines, such as social media and online gaming. This should not be considered a confound *per se*, nor should it impact on FAST RFD scores if the relative decrease in fluencies of performance compared to younger people is proportionate on both blocks of the test. However, if the FAST were to be employed in research into older age groups in the future, a change of the block contingencies could be considered in order to make the task easier for older participants. Specifically, the response window of the individual FAST trials could be elongated to allow participants more time to respond. Another change that could be considered, is to decrease the number of trials on a FAST block, this would also make the FAST easier to complete and potentially prevent participant fatigue with the task. Of course, both of the preceding changes are suggestions for research specifically into an older sample of participants. If these block changes were applied to a general sample, the FAST would likely become too easy for younger participants, thus rendering the younger participants results incomparable to the older participants. Alterations

to the procedure, such as the aforementioned suggestions, are perfectly in line with the development history of the FAST. Methodological manipulations are a far more transparent way of accounting for the difficulties faced by older age groups, than say, post-hoc statistical data stabilisation methods.

Given, the somewhat unexpected results of Experiment 1, it was decided that it would be worthwhile to test for the phenomenon of outgroup favouritism across two groups with a greater differential in social status between them. Experiment 2 was methodologically identical to Experiment 1. The FAST did not require extensive modification to be applied in the context of racism. The stimuli employed in the FAST were not varied from Experiment 1, aside from a change in the reinforcement contingencies (i.e., functional response class configuration based on race rather than gender). This is because the visual stimuli employed already featured images of both White and Non-White individuals.

Although the initial descriptive statistics were more in line with the central hypothesis, it remained to be seen whether or not the FAST was sensitive to something the explicit measures were not. The correlational analysis provided some initial evidence that this was the case. No correlation between either of the explicit measures and the FAST RFD scores for the participant cohort as a whole was found in Experiment 2. This provided some tentative evidence for the hypothesis, in that the FAST revealed an outgroup favouritism not apparent on the explicit measures. To further support this, the results of the independent-samples t-tests found no significant difference between groups along any of the dimensions of interest. To make these findings clear, these analyses suggested that despite both groups self-reporting low levels of racial bias, the implicit measure revealed the inverse. Not only this, but Non-White subjects demonstrated similar levels of White-positive bias as did the White participants. While the t-test indicated the groups were statistically different in terms

of RFD scores, the Bayesian analyses were more inconclusive, indicating that the groups were not unambiguously different in their scores. Taking these results together suggested some convergent validity between the results of the FAST methodology and wider system justification literature. Specifically, the implicit FAST converged with the findings of the IAT studies conducted within SJT, in that the low status Non-White group displayed the phenomenon of favourable outgroup bias towards the dominant White group. But of course, fair due diligence dictated that for the sake of consistency the results be subjected to the same post-hoc analyses as conducted in Experiment 1.

A mixed between-within groups ANOVA was conducted to examine the impact of ethnicity on participants fluency scores across the consistent and inconsistent blocks. This analysis found a moderate main effect for block type. That is, fluency scores for the participant cohort as a whole were lower on the inconsistent compared to the consistent block. The inconsistent block required assigning a common response to Non-White and positive stimuli, and a different common response to White and negative stimuli. Therefore, on average, the participant cohort as a whole, had greater difficulty (longer responses with more errors) in assigning a common response for Non-White with good, than White with good. Unlike Experiment 1, this overall effect was not carried by any single group. Furthermore, the main effect, which compared combined block fluency scores across ethnic groups was not significant. In other words, these groups did not differ in overall ability to acquire functional response classes. In short, it can be said with some confidence that the ethnic bias revealed at the implicit level was found for both groups.

Again, for the sake of consistency, an analysis of covariance was considered to examine bias level as a function of age. However, there was insufficient variance in age in the

present sample. Therefore, it was deemed that an ANCOVA examining block fluency differences for the entire cohort with age as a covariate would be unsuitable.

The results of these experiments taken together provide an interesting account of both implicit bias in Irish society and the utility of the FAST as a metric by which to measure it. The following sections will examine how these results can be interpreted in light of further evidence from the literature. Specifically, it will be considered how well these results cohere with system justification theory and suggestions for future research employing the FAST methodology will be made. Before this however, because the FAST was developed for an in-person laboratory context, it would be worthwhile to examine its performance here in the context of online research. This is best achieved by considering the quality of the data *vis-à-vis* the needs and grounds for data exclusion.

4.3 Data Quality and Omissions

There was some degree of attrition in both studies, possibly attributable to its online remote nature. Several participants in both studies were excluded on the basis of not completing one or more measures. Overall, the number of cases that had to be excluded was not excessive. Out of the 104 participants who took part in Experiment 1, only 6 (5.77%) had to be excluded. In Experiment 2, of the 60 participants who took part, 8 (13.33%) had to be excluded. The vast majority of exclusions were on the basis of not completing, or even attempting, either the FAST or one of the explicit measures. Only 1 participant from each experiment was excluded due to scoring 0 correct responses on one or more FAST blocks. In person laboratory studies rarely, if ever, have to exclude participants on the basis of not completing one of the measures. The reason for this being that it would be unusual for a participant to leave mid-experiment. In this online setting however, if a participant grows bored, or tired, they can exit the experiment by simply closing the task. Alternatively, a

computer error with one of the tasks may have caused them to abandon the experiment, this cannot be known without participant follow-up however, which was impossible due to the study's anonymous nature. Therefore, these factors may have contributed to the number of excluded cases in the present set of experiments. Adding to this point, there is some possibility that participants levels of engagement decreased as a result of the online setting. While participants were encouraged to conduct the study in a quiet setting free of distraction, there was simply no way to ensure these instructions were followed. It is possible, and even likely, that at least a portion of participants conducted the study in a less than ideal setting. While the idea that the online setting of this experiment may have affected participant performance has good face value, it remains conjecture, for this reason we should look to the available literature on the subject.

Semmelmann and Weigelt (2017) conducted an experiment to compare various reaction time-based measures across three settings, using different samples of participants in each instance. The first setting was the classic baseline, in lab, conducted on hardware provided by the researchers and using software present on the computer. The second (web-in-lab) was identical to the first, expect for the fact that participants conducted the experiment through web-technology i.e., conducted in the lab, on the same hardware, except through the internet. The last was conducted entirely online, using participants own hardware and accessed in a place and time of their own choosing (though these participants were briefed on the experiment in person beforehand). Semmelman and Weigelt administered five different reaction time-based measures to their participants, though perhaps of most relevance to the present study, one of these measures was the Stroop task (as discussed in the introduction, the Stroop bears some similarity to implicit tests). The researchers instructed all web-based participants to close any other internet browser windows other than the task itself so as to prevent their impact on computer performance. They found, that both web-in-lab and online

(at home) participants showed a significant increase in average reaction time in comparison to laboratory participants. It was their suggestion that the type of internet browser used by participants might have influenced reaction time measurement and accuracy. Furthermore, online participants may be using unoptimized, overloaded hardware, with many programs installed, which might impact computer performance and thus reaction time measurement. However, subsequent analysis did not conclusively show a relationship between system performance (as measured by a brief performance test before the experimental tasks) and reaction time accuracy. They instead concluded that the additional reaction time offset found among the online participants was likely due to environmental causes, or a speed-accuracy trade-off. Though their study concludes that conducting experiments online likely does have some impact on reaction times, on average this impact is relatively minimal (87ms), and importantly no change in error rates was observed across settings. Additionally, the expected effects from each task were found across all conditions (barring the priming task which was not replicated in any condition), and the only task which relied entirely on error rates (the attentional blink) was in fact more accurate when conducted using web-technology. It was the conclusion of the authors that despite the popular belief that online participants are inattentive to experimental tasks their data appears to be quite comparable to regular participants and should be a safe pool of data to draw from. It is worth noting for future online based FAST experiments that it would seem prudent to use a specific internet browser, and to close all other browser windows, and any other computer programs. It would seem safe to say that the online nature of the present study likely did not have any major impact on the data and as such we should examine other potential factors which could have impacted the data quality.

Of course, it is important to note that low fluency scores are not by themselves indicative of inattention to the task. It remains possible that the relevant word stimuli employed here were simply not salient in the vernacular of some participants. Alternatively,

random cohorts of participants will sometimes have lower rates of learning for reasons beyond the experimenter's control. Fortunately, it is likely that generally lower rates of learning on the FAST blocks did not impact the findings to any considerable degree. This is indicated by the significant differences observed across blocks in both experiments, suggesting the FAST was sensitive to cultural contingencies, even where there is a degree of behavioural competition during the task.

Given the concern regarding data exclusion as a replacement method for increased behavioural control, and given our interest in behavioural variability in behaviour analysis, data exclusion criteria were conservative rather than progressive. As a result, some participants with low (e.g., less than 10 on a block), or even negative fluency scores (i.e., more errors than correct responses in one or more blocks; 5 total across both experiments) were retained. While "low quality" data (i.e., resulting from poor stimulus control) to some extent most likely did compromise the clarity of the effects observed here, a stricter data exclusion system would be a poor remedy for this issue. An overly prescriptive algorithmic system for data inclusion and exclusion would represent a retrograde step in the development of implicit testing methods within behaviour analysis. That is, such a method would form part of a psychometric style approach to construct testing. Insofar as the test system itself would remain intact, while increasingly elaborate and progressive forms of both participant and data exclusion would be developed to enhance the statistical significance of test findings. This would come at the cost of increases in effect sizes obtained through improvements in stimulus control. This more psychometric style approach has arguably been taken in the development of the IRAP. The FAST, in contrast, was developed using a more functional-analytic strategy in bottom-up research. It would be regrettable therefore, to eliminate any but the most obviously inadmissible data (e.g., complete nonadherence to the task). This is preferable to developing graded levels of performance criteria worked out retrospectively

based on which types of exclusion criteria achieved statistically significant effects (i.e., grand scale p-hacking).

To add to the previous point, once an elaborate algorithm has been formed for either data exclusion, participant exclusion, or calculation of the key metric, the test becomes reified and psychometric by default. There is a danger therefore, that the method would become proprietary. The scoring method itself would then become a barrier to innovation and exploration. This is because data generated with even moderately alternative procedures would quickly become considered to be data not reflective of, or comparable to, data generated using the original proprietary procedure. This has also arguably occurred with the IRAP, which at times has taken the precaution of naming software versions, in order to flag the non-proprietary nature of successive and independent iterations of the software developed by independent laboratories. In contrast, the FAST was developed under the presumption that both the details of the precise methodology employed in any one study or intervention, and the precise scoring system used, are still to be considered a part of the broad rubric of the methodology. Thus, evolution across laboratories through independent research is what will make the method useful in the longer term. For these reasons, it was not considered here that methods inspired by the pursuit of statistical significance should be used to further refine data exclusion methods. After all, any cut off points for data inclusion would be arbitrary and could be guided only by such criteria as statistical significance, which would be a retrograde step for our field.

While a precise appraisal of how the FAST performs in the online environment was not a main focus for the present research, it would seem prudent to say a few words about the benefits and drawbacks of this setting given the preceding comments on data quality. By and large, these experiments have shown the FAST's utility in the online environment, at least in

principle. While it would seem that the degree of attrition is greater than in a laboratory setting, it must not be forgotten that the online environment has the potential to secure larger numbers of participants in general. There are several reasons for this, but predominantly, the online environment allows easy dissemination of invitations to participate. It achieves this easy dissemination through both online posts and the snowballing method employed here (wherein participants were encouraged to forward the study link on to others). It is also very easy for subjects to access as they need not visit a physical setting, and it allows subjects to participate in a setting of their choosing, where they may be more comfortable. Therefore, future researchers should consider a cost-benefit approach, wherein they weigh the rate of attrition and potential data quality issues against the larger data pool the online environment is likely to provide. It could be the case that the FAST is only suitable in an online setting for specific research questions. Though of course only further experimentation will reveal if this is the case. It is safe to say, however, that in the context of the experiments at present, the FAST performed well, and with this in mind future research utilising the FAST to answer questions of social relevance can now likely be safely conducted in an online setting.

While the FAST's performance in this study was good, there was some missing data from the explicit measures in Experiment 1, though not in Experiment 2. It was noted in the results for Experiment 1 that four participants had missing data. While several data replacement methods were considered, namely various imputation methods, these were ultimately not implemented. The reasoning for this was due to the very small number of missing data points. It would seem highly unlikely that such a small amount of missing data would influence the results one way or another. It was for this reason that a neutral score of three was inserted for these missing values. Given that the MS scale operates on a Likert scale of 1-5, it was thought that replacing the missing values with a value of 3 would neither bias their results upwards or downwards. However, it is acknowledged that this was a

relatively ad hoc strategy, and an imputation method may ultimately have been the more prudent choice.

4.4 System Justification Theory

It is worth noting once again, that while system justification theory served as an impetus for this study, the main purpose of the study was not to appraise the theory in its entirety. Rather, the current study was motivated by one major claim of SJT; the same claim which first distinguished it from other similar social psychological theories of group favouritism. That claim was, previous theories had failed to account for the phenomenon of outgroup favouritism by oppressed social groups. While it is outside the scope of this thesis to outline these previous theories in full, a brief reminder on how they differ from system justification is warranted.

Jost et al. (2004) maintained that social identity theory and social dominance theory adhere too closely to assumptions of self-interest. Jost et al. claimed that social identity theory emphasises the purpose of stereotypes in allowing for the rationalisation of how a person's ingroup treats the outgroup. Similarly, Jost et al. argued that social dominance theory relies too heavily on assumptions of self-interest and thus cannot account for findings of outgroup favouritism. These theories, by and large, hold that dominant groups in society impose the present social order and disadvantaged groups resist it. This is a claim that Jost et al. find lacking in evidence and suggest that it is too narrow an explanation to account for outgroup favouritism. In contrast, SJT explains the existing social hierarchy not solely as a result of dominant groups' ingroup favouritism and outgroup derogation, but also supported by non-dominant groups' outgroup favouritism.

To the behaviour-analyst, it might seem that the above-stated claims are wide sweeping macro level interpretations, and too poorly defined technically to allow for

experimental analysis of the various concepts involved. However, while we cannot assess such elaborate discursive theories in whole cloth, nothing prevents the behaviour analyst from examining particular aspects of such accounts. Such an approach can also form a part of a longer process of translational research and bridge building to mainstream social psychological theorizing. As a case in point, SJT theorists' claim that marginalised groups will show outgroup favouritism towards dominant groups was easily tested, so long as one is comfortable with a behaviour analytic interpretation of favouritism used in these studies. In the current research this favouritism was defined in terms of fluency. Specifically, a relatively enhanced fluency in the formation of functional response classes involving stimuli with positive functions and stimuli representing the outgroup, as well as the formation of classes involving aversive stimuli and stimuli representing the in group. In addition, Jost et al. (2004) specifically predicted that such outgroup biases would be more easily predictable at the implicit level. Of course, a number of previous studies employing the IAT have examined this idea. While it might be argued that these studies are a product of a different body of literature to our own, and thus bear no relevance to the behaviour analytically oriented FAST, nonetheless, there is good reason to examine them. Namely, Cartwright et al. (2016) found that the FAST co-varied with the IAT, demonstrating that both methodologies performed perfectly equally, at least in the context of detecting gender stereotypes. As such, it could be argued that using the FAST specifically to examine this topic is unnecessary as it has not been shown to diverge from the IAT. However, where it is available, it seems to be a sensible choice to use the more functionally understood test, as opposed to one based on more abstract hypothetical deductive discursive reasoning. With this in mind, it is still reasonable to examine the results of IAT's in the context of system justification to guide us in our interpretation of the present experiment.

The reader may recall the IAT data concerning SJT outlined in Chapter 1. These will now be briefly outlined again for clarity. In a sample of Blancos and Morenos from the Hispanic population, Uhlman et al. (2002) showed that Morenos displayed an implicit outgroup bias in favour of Morenos. Importantly, the Blanco-positive bias the Morenos displayed was not as strong as the ingroup favouritism the Blancos displayed. Similarly, Nosek et al. (2002) in a large analysis of data taken from the projectimplicit.com website showed that African Americans displayed an implicit outgroup favouritism. Importantly, Nosek et al. also showed a divergence between implicit and explicit measures of racial bias (i.e., African Americans self-reported ingroup favouritism). Jost et al. (2004) at a later date similarly extracted IAT data from projectimplicit.com and showed the same effects as Nosek et al. (2002). These foregoing studies are in line with the results of Experiment 2, wherein the Non-White participants as a whole displayed an outgroup favouritism (not found on an explicit measure), but not to the same degree as the White participants ingroup bias.

On a separate but related note Rudman et al. (2002) demonstrated that the phenomenon of implicit outgroup favouritism was most easily found in samples of low status social groups. Like the previous studies on race, Rudman et al. showed that that this outgroup favouritism was not displayed on explicit measures. Finally, Jost (1997) demonstrated that women exhibit a depressed self-entitlement relative to men on a task of rating how much their work was worth monetarily.

While the results of the racial IAT studies cohered almost perfectly with Experiment 2, the fact that Experiment 1 produced results less in line with SJT is a finding rightly deserving of further consideration. As has been outlined throughout, SJT does not specifically claim that women will display an outgroup favouritism in the way it was measured here. More so, it emphasises, women's depressed self-entitlement. Although at least one study

conducted by Rudman and Goodwin (2004) suggests that women may be a general exception to the typical SJT rule concerning outgroup bias among discriminated against groups. Nonetheless, the lack of ingroup bias among men and the relatively strong ingroup bias observed among the women is still an unusual finding given the broader themes of SJT. It might reasonably be asked, therefore, whether the results of Experiment 1 did not cohere with SJT as a consequence of employing the FAST methodology.

To be more specific, there is always the possibility that the FAST may produce results that sometimes diverge from those of the IAT. If this were to be the case, it may be a result of differences in the behavioural processes to which each test is sensitive due to their different methodologies. We might find therefore, that a theoretical position is supported by results obtained using one implicit test method, while it is challenged by data produced using another. Where this is the case, it will raise important questions about how test results are intrinsically a product of the test methodology. It will also shine a much-needed spotlight on the ill-founded notion that underlying the results of popular implicit tests are solid constructs that are unwavering in their nature, but which defy accurate measurement until scientists can develop the perfect procedure for their capture. In the case of the FAST, there has been a slow steady development of the method based on laboratory research, mostly using laboratory-created stimuli. Thus, claims as to the nature of the phenomenon being assessed using the FAST are well grounded. In contrast the literature surrounding the IAT and the IRAP are replete with post-hoc rationalisation based on theory, with almost no studies in either case involving laboratory-controlled stimuli and artificially created stimulus classes. The manipulation of variables at the level of administration is a very weak way in which to identify core processes. This can in theory be achieved through an exhaustive process of triangulation via application in real-world measures of verbal behaviour/attitudes but will ultimately always be inferential. The field sorely needs a test method for which every aspect

and dimension has been studied in the laboratory and its inclusion in the overall method justified. Insofar as the FAST is now perhaps more than a mere embryonic methodology, and insofar as it increasingly appears to be sensitive to the measurement of social histories, it may deserve considerable research investment to further develop this method.

The above notwithstanding, it should be remembered that the only published direct comparison of the IAT and the FAST in an experimental context was produced by Cartwright et al., (2016), who found that the FAST co-varied almost perfectly with the IAT. Given this, we can assume that based on the very limited data to date, similar results would have been observed here with the IAT, and that it too would have found its own data at odds with the predictions of SJT, at least using the current stimulus sets. As a brief aside, whether the results of future FAST studies will always co-vary with the IAT is obviously unknown. If it is the case that they do always produce the same results, the utility of the FAST will of course come into question, on the basis of the logic that it offers nothing new. The purpose of FAST research however is not to build a test which is "better" than the IAT in that it can do things the IAT cannot. The impetus behind this project is to build a test that is actually understood in a fully functional way, which has been proven to test what it claims to test, and whose results can be interpreted in a behaviour-analytic fashion, without recourse to hypothetical constructs or artificial alteration of data. Ipso facto, whether the FAST will always produce similar results to the IAT is irrelevant, however, where it does produce differing results, this will clearly delineate how differences in the underlying methodology lead to divergence in the observed results. Without engaging in too much speculation, it seems likely that the FAST will converge with the IAT in the majority of instances, though they likely will diverge in specific instances (any hypothesis about which instances would be conjecture at this stage) as research into the FAST continues.

Given the foregoing, it should now be apparent that it would be unwise to interpret the lack of support for SJT in Experiment 1 observed here as evidence of the inefficacy of the FAST methodology. Of course, it should be remembered that the findings of Experiment 1 do not directly contradict SJT, as the authors of that account never unambiguously claimed that women would display a simple valanced outgroup favouritism in the sense measured here (See also Rudman and Goodwin, 2004). Therefore, assuming a relatively sound methodology and a relatively well-run procedure, the findings of Experiment 1 should be taken at face value. Possible compromises to the experimental procedure will of course be considered later in this commentary, but for now let us consider that the data may be an accurate representation of the various fluencies in the verbal repertoires of the participants. Consequently, no significant bias against women was observed within the sample, particularly for the female participants. The most obvious reason for this outcome relates to the social change that has occurred even in quarter century since SJT was developed, at least in Western Europe.

Specifically, despite its past association with conservative and religious elements, modern Ireland has become a highly liberalised nation. As testament to this, in 2015 Ireland became the first nation to legalise same-sex marriage by popular vote. In the vein of gender equality, the 2018 abortion referendum again passed by popular vote, granted women the right to make choices regarding their own bodies and pregnancies. A right that had been aggressively resisted by the state and dominant religion for centuries. It could be argued that these specific legal changes are symptomatic of deeper social change and attitudes to gender and sexual identity within the country. This suggestion is further supported by the findings of the Global Gender Gap Report which indicated that Ireland ranks favourably in terms of gender equality (World Economic Forum, 2022). Specifically, equality between men and women was measured in terms of political representation, wealth and education, as well as a

variety of other factors. The report ranked Ireland 9th globally in terms of gender equality and stated that Ireland had closed roughly 80% of the gap between men and women (World Economic Forum, 2022).

Pointing to such evidence as the aforementioned surely helps to make sense of womens' positive categorisations of themselves as inferred from the first procedure. While the absence of a pro-female favouritism for the male participants was not observed, this may not be surprising, in that the absence of a pro-female bias is indicative of no bias at all. In addition, perhaps it should be noted that an anti-female bias was expected, the absence of such a bias again points to progressive social change. To add on to this, it might not be surprising that the greater attitudinal change was observed for female participants given their vested interest in women's rights. Where an attitude object has important perceived consequences to the individual, there is more likely to be a link between their attitude and behaviour (Sivacek & Crano, 1982). That is to say, when the potential outcome of an attitude is likely to have a high impact on the individual, they are more likely to actually behave in accordance with that attitude. If we interpret the findings of Experiment 1 in this light a more coherent explanation arises. Women have a vested interest in the advancement of their own rights, therefore their strong pro-female responding on the FAST (i.e., attitude-behaviour consistency) may be a result of this vested interest.

While the preceding argument may be intellectually appealing and surely holds some water, there is no doubt that all over the world, and including Ireland, female emancipation has not yet been fully realized. Reports from the Central Statistics Office in Ireland indicate that 29.1% of workplace discrimination incidents reported by women cited their gender as the cause of discrimination, compared to only 7.8% for males (Equality and Discrimination Report., 2019). From this, it can be deduced that gender-based discrimination remains a

problem in Ireland, at the very least more so for women than men. Taking this into account, it is possible that despite these difficult to change cultural conditions, there is a genuine shift occurring in the way we speak about gender. Therefore, a similar change in the configuration of verbal classes in the vernacular could be expected, even where oppressive practices still persist. It is important for the reader to understand, at this point, there is no requirement that explicit and implicit verbal behaviours cohere, and the expected incoherence is part of the reason why researchers have become so interested in the outcomes of implicit tests. In other words, perhaps unusually, the current study may have identified implicit positive biases towards women, or at least absence of negative bias, in a context in which overt behaviour is in fact still biased in men's favour. Understanding the complexities of the relationships between overt behaviours, explicit verbal reports and implicit test performances is a complex matter, and it would be unwise here to become overly entrenched in any one particular narrative explanation for these conjunctions and disjunctions. However, the foregoing gives a flavour of that complexity and the types of questions that lie ahead for researchers interested in implicit verbal behaviour.

With that said, it remains interesting to note that while weak a correlation between implicit and explicit measures was observed in Experiment 1, no such correlation was found in Experiment 2. Though given that the degree of covariance between implicit and explicit measures was similar across both experiments, this finding might have resulted from a lack of power in Experiment 2. However, this experiment found no evidence of self-reported racial bias, while simultaneously detecting clearly significant levels using the implicit test measure. Despite the anonymity provided by the online environment, subjects nonetheless self-reported racial bias was quite low on the MMR, and modest on the DS/DV. This was true for both White and Non-White subjects. Importantly, self-reported racial bias did not differ between White and Non-White participants. While according to the independent samples t-

test the difference between groups was not significant in terms of RFD scores, average levels of bias were lower for Non-White than White participants. Furthermore, the results of the Bayesian analysis were inconclusive, neither indicating that White and Non-White participants were equivalent nor different in terms of RFD score. However, this finding does not constitute a direct contradiction of SJT, which claims that discriminated lower-status groups will have an outgroup favouritism, but not necessarily that this bias will be equivalent to that of the dominant group's ingroup favouritism (Jost et al., 2004).

The foregoing notwithstanding, it would be difficult to fully explain the lack of correlation between implicit and explicit measures in Experiment 2, given the presence of such a correlation in Experiment 1. Even acknowledging that the correlational results may have been a result of insufficient power in Experiment 2, the finding of an implicit outgroup bias among some participants still stands in stark contrast to the results of Experiment 1. Fortunately, the defining difference between the experiments may shine some light on the issue. That is, these experiments focused on forms of prejudice involving relative bias against two groups from within and across those groups, but they differed in terms of the relative power relationships between the groups and questions. More specifically, it is quite likely that Non-White individuals face greater discrimination than women in Ireland due to their minority status. In effect, the results of Rudman et al. (2002) may explain this lack of correlation, as they found the greatest disparity between implicit and explicit bias for groups of particularly low status. Accordingly, while women face discrimination, they remain a relatively high-status group in comparison to ethnic minorities. This argument is supported by reports of discrimination from Non-White respondents in Ireland. Those of a Sub-Saharan background in Ireland have frequently reported receiving hate-motivated harassment, and facing discrimination in the hiring process (European Union Agency for Fundamental Rights, 2017). In short it is clear that Non-White individuals experience a high degree of

discrimination in Ireland. If we take the suggestions of Rudman et al. (2002) concerning the social status of groups being a predictor of implicit outgroup favouritism and the reports of discrimination faced by an ethnic minority group together, a neat explanation for the results arises. The implicit outgroup favouritism of a discriminated group was revealed as per system justification, and the large discrepancy between implicit and explicit measures was a result of Non-White individuals' low status in society. Overall, the results of Experiment 2 provide some support for the conclusions of system justification theorists and bulwarks the IAT studies already conducted in this domain. This should also provide some encouragement to future researchers employing the FAST methodology. Though there are stark differences in methodology between the FAST and IAT, it seems likely that in essence they measure the same types of behaviours. Therefore, the large body of IAT research available provides a useful starting point for avenues of further investigation with the FAST.

While the previous sections might provide a reasonable explanation of why implicit and explicit measures might not cohere in the context of system justification, it does not explain how this can actually occur. That is, what is the difference in the type of verbal relations incurred when responding between an implicit and explicit measure? While the variety of social reasons for why this might occur have been discussed, *how* it can occur has not yet been explored. It seems likely that answer rests in how these relations are formed and in what way explicit vs. implicit measures tap into the relations under examination.

4.5 Explicit and Implicit or Direct and Derived?

Where implicit measures correlate with explicit ones, it raises the question of the utility of implicit tests. The fact that they do occasionally correlate suggests that implicit measures are not equally useful in all domains. From the cognitive perspective it could be said that where the two measures correlate, they may not be measuring distinct constructs.

From the behavioural perspective however, we could look to a possible difference between the nature of the relational types being assessed in an implicit and explicit test for an explanation of this divergence. Specifically, it could be argued that complex attitudes which have been highly elaborated at the explicit level may be difficult to measure accurately using simple binary implicit tests such as the IAT or the FAST. The reader may at this stage note the IRAP was specifically designed to assess nuanced complex relationships between words. As has been discussed throughout however, the IRAP takes an extensive amount of time to administer, and suffers from high attrition rates, possibly in part as a product of its time intensiveness. Additionally, this thesis has raised concerns over its psychometric style scoring system, which is similar in style to the IAT, and is generally too opaque for the field of behaviour analysis to inherit without considerable empirical justification. The FAST currently measures only simple relationships of equivalence between verbal relations. While developing a relational version of the FAST to overcome this should not be ruled out, such a methodology is outside of the scope of the current thesis. For the time being, however, it could well be worthwhile to examine the difference between explicit and implicit measures in terms of the nature of relations that each metric captures.

While it has been noted throughout that a social desirability bias may be responsible for the divergence between implicit and explicit measures, it remains the case that subjects may not always be attempting to hide their true response. Subjects may not always be able to discriminate a clear attitudinal position on a topic, and instead may at best be able to poorly discriminate a general dispositional “feeling” towards the object of the question. Therefore, employing questions with greater specificity, as advocated by Ajzen & Fishbein (1977), may overcome this problem to a certain extent. However, in this instance, an implicit test may be more suitable for identifying verbal histories, because even emotional dispositions are affective responses mediated by verbal contingencies in the history of the subject. Such

affective responses should be detectable using the current FAST methodology. In instances where one's own opinion is not fully developed verbally, there may be a tendency to answer questions based on perceived appropriateness. This is as opposed to reflective discriminations of one's past verbal behaviour both public and private. This explanation may go some way in explaining why subjects can be seemingly quite surprised by their implicit test results.

Nosek (2007) argued that in cases where explicit and implicit test results diverge, test takers may refuse the implicit test results and insist upon the truthfulness of their explicit test results. In his words "implicit evaluation reflects accumulated experience that may not be available to introspection and may not be wanted or endorsed but is still attitudinal because of its potential to influence individual perception, judgement or action" (Nosek 2007, p. 68). Of course, questions concerning whether or not some attitudes really are completely introspectively unavailable to the individual are completely outside of the behavioural domain. However, Nosek's commentary suggests an explanation may be found in the nature of the relations under examination. Specifically, "accumulated experience" not readily available to introspection may refer to relations that are derived rather than directly taught. That is, a subject's verbal environment may have indirectly taught a relation such as Black-bad, while the subject remains aware that this is an unacceptable or "wrong" way to respond as a result of social norms (i.e., they may have observed others being scolded for expressing such a view in the past). Therefore, despite the implicit test correctly detecting this history of verbal relations, the subject explicitly rejects these findings. In order to understand this explanation in full we shall need to return to the RFT account of attitudes.

O'Reilly et al. (2015) assume a rather simplistic and uncontroversial RFT model of attitudes, in which, attitudes are understood to be defined functionally in terms of complex relational networks facilitating transformations of stimulus functions. In the case of attitudes

these are most usually affective stimulus functions. An attitude therefore is simply a complex relational/affective response. Furthermore, in the same way that verbal relations can be both directly trained and derived, attitudes could emerge from direct experience in the form of directly reinforced verbal relations. Alternatively, they could emerge indirectly as derived relations between relata in a complex relational network, but in the absence of direct reinforcement for derivation. To this extent, an attitude can be quite literally implied by a relational network but never explicitly derived. It does not follow, however, that an individual who has never consciously discriminated that say, African Americans are bad people, does not participate in a culture and display a verbal repertoire that does not contain within it support for the derivation of such a relation. This derivation may only emerge at some appropriate juncture in the future when the stimulus conditions are correct (e.g., being asked for the first time if they think that African Americans are bad people). From an RFT perspective, even the relational network that supports the derivation of a racially biased verbal relation need not itself have arisen through direct training of relations of coordination between stimuli. In contrast, it is possible that many equivalent relations within a relational network arose through the training of opposite relations, rendering the derivation of verbal relations supporting racial bias to be even more indirect.

The previous point is somewhat conceptually dense, perhaps a situation in which this might occur in the real-world would better illustrate the point being made. As an example, an individual need never be told that all members of a particular ethnic group are not very bright for this relation to emerge. The relation could arise through regular social reinforcement of observations that members of that ethnic group are never equated with intelligence. Perhaps, for instance, this particular ethnic group is rarely represented in TV programs in the role of college professor or brain surgeon, and is usually represented as the school janitor and hospital patient. Given such a social history, terms referring to the ethnic group in question

will quickly lead to responding to equivalent terms which are by definition non inclusive of terms describing high levels of cognitive functioning (e.g., the word “stupid”). The inter-relatedness of series of relations like this, each derived from the reinforcement of relations other than this, could lead to the derivation of still further relations that are seemingly very remote from the verbal statements heard within the verbal community. Despite this, their verbal culture still directly and logically supports the derivation of such a relation. O’Reilly et al. (2015) suggest that it is in such a way that behaviourists can best view what we mean by implicit relations, rather than in terms of the indirect measure of explicit relations.

In the previous section it was pointed out that a verbal culture may logically support a derivation of a relation - this point is worth further clarification. In short, baseline relations might be trained which logically support a derived relation which could be called ‘racist’, ‘sexist’ etc. While the subsequent derivation might go unreinforced for an extended period of time, at some point in the future in a specific context, the subject might ‘spontaneously’ respond in a prejudicial way on the basis of the baseline relations they learned. If enough training has occurred in the subject’s culture to imply that prejudicial relations should be derived, then they will eventually one day be derived for the first time. It is worth acknowledging that for some participants, the FAST may in fact be providing those contingencies for the first time. Additionally, for some people a relation of opposition between men/women, Black/White people may have been taught. For example, if a person is taught that “White people are fantastic” then by definition there has to be either punishment or at least non-reinforcement of relating White people to negative terms. Furthermore, teaching an individual that one ethnic group is good at one set of skills (e.g., chess) and a different ethnic is good at a different set of skills (e.g., gardening) may lead to the emergence of a relation of difference between these groups. Such training supports a relational network in which terms which are equivalent to White people are mutually exclusive with terms

relating to Black people. Despite a prejudicial relation never explicitly being derived, this verbal background should in principle be easily detected under the contingencies of the FAST. While for subjects' whose verbal background did not teach such relations the contingencies of the FAST may be entirely unfamiliar, as they bear little functional similarity to their prior learning. Such participants are likely to respond slower on FAST trials, but importantly they should respond equally slowly to all trial types. The FAST may in fact teach a relation of opposition to participants but given its brevity, and the fact that for these individuals it does not cohere with their prior learning, it would seem likely that this relation would not have any lasting effect on their behaviour.

If an equivalence relation can emerge in as subtle a way as the mere training of opposite relations it should not be surprising that an explicit test may be unable to detect the resulting attitude. It seems likely explicit tests are limited to the detection of directly trained relations, whereas implicit tests are more sensitive to the more subtle derived and implied underived relations that are likely to emerge from a naturalistic relational network. On the face of things, it would seem a reasonable suggestion that directly taught relations would be immediately available to the subject (i.e., "the first thing that springs to mind"). As opposed to this, relations of greater nodal distance (number of intervening stimuli in an equivalence relation, or other relations and relations between relations) might not be immediately accessible, but nevertheless serve as contingencies to control verbal responses.

In effect, the O'Reilly et al. (2015) suggestion that many attitudes can be considered in terms of highly complex relational networks, consisting sometimes of underived relations among the verbal stimuli dovetails nicely with the view of Nosek (2007). That is, participants bring to these tests an "accumulated experience" which is an entity in itself sometimes possibly competing with explicitly reinforced and more simplistic verbal rules brought to the

test context by participants. In both Nosek's terms, and from the point of view of RFT, it may be difficult or impossible for an individual to discriminate the types of potential verbal relations supported by their verbal history. Whereas it may be easy in contrast, to verbally report well reinforced simplistic verbal equivalences. The two need not cohere and language is likely replete with such logical incoherencies.

It would be an enormously complex matter to begin to speculate on the social contingencies that lead to the emergence of an individual behavioural repertoire that consisted of various forms of coherent and incoherent relational responding in different contexts. In effect, we would be speculating on the reasons why a particular individual might deny being racially biased while at the same time behaving overtly in ways that appear to characterize this very bias. Of course, a core assumption of Relational Frame Theory is that language coherence is itself reinforcing for verbally able humans, and so over time relational coherences may develop within the verbal repertoire of the individual. We need not take this assumption on face value however. Bordieri et al. (2015) demonstrated this in a matching-to-sample task. Specifically, Bordieri et al. found that participants responded to ambiguous stimuli in an MTS procedure in ways which were coherent with their previous learning histories. In the absence of reinforcement, participants tended to categorise stimuli according to how they had done so in a previous MTS procedure, despite their responding never being reinforced in this procedure either. Participants who were not exposed to the first procedure responded to the stimuli in the second procedure according to the sample stimuli provided. In effect, this study suggests that individuals find coherence itself reinforcing. Taking from this, it could be reasoned that the "accumulated experience" Nosek (2007) refers to could in fact be past patterns of coherent responding. This past coherent responding may subsequently influence an individual's response to novel contexts.

As stated already, the historical relations being measured by the FAST need not have been explicitly taught. Therefore, implicit tests may in various circumstances be detecting directly established relations *or* derived relations between stimuli of varying nodal distance. It is curious that in the literature concerning the widely explored and applied IRAP, no empirical study has examined the difference in effects between its assessment of derived versus directly trained relations. This would seem to be an important distinction and might explain why test effects are sometimes weak. Knowing that the relations involved are derived in nature, rather than directly established, would partly explain any lower than expected levels of relatedness between stimuli in verbal relations. This in turn might modulate conclusions regarding the presence or absence of relations between the verbal classes of interest. However, the important point here in the current context is that derived relations between stimuli within a larger stimulus set may be incoherent with relations among exemplars from that same stimulus set that have been directly established (i.e., the difference between racial bias proliferated through innuendo and that proliferated by instruction).

Fortunately, we need not take the argument that implicit tests are sensitive to derived relations as well as directly trained relations entirely on face value. The reader may recall the experiment conducted by Cummins and Roche (2020) outlined in the introductory section. In short, Cummins and Roche (2020) demonstrated that the FAST was sensitive to relations of varying nodal distance. FAST effects were larger for stimuli of shorter nodal distance. This outcome supports the thesis above, implicit tests in real world research may be targeting relations of varying nodal distance and complexity, and therefore result in different implicit test effect sizes. More specifically, probes for relations between word pairs that involve directly related stimuli were likely to lead to larger effects than probes for relationships between word pairs that participate only in derived relations. Probes for relations between stimuli separated by multiple nodes may lead to weaker still test effect indices, or may lead to

contradictory effects where incoherencies remain across stimuli in a larger set. While a weak effect may be interpreted, therefore, as evidence of a weak bias, it is perhaps more accurately described as a measure of the strength of a yet to be explicitly derived relation. In fact, even where this is the case, the networks may be highly elaborated and reinforced, and highly supportive of the derivation of a yet to be derived relation between two words. At present, we have no way to distinguish between these two cases, and this represents a very important research question not yet addressed by other researchers within or outside our field. Suffice it to say for the time being, that though large nodal distance between stimuli may present itself as a weak bias on the FAST, such a relation would likely be completely undetectable by a standard explicit test.

With the above taken into account, it is now not unreasonable to suggest that the FAST and other implicit tests should have more utility than explicit tests in measuring attitudes which have not been directly taught. Although, implicit tests are still sensitive to directly taught relations, it may just be the case that their utility in these instances would not exceed that of explicit measures. In addition, where a social desirability bias is weak or absent, and the relations are directly taught, a high degree of correspondence between implicit and explicit measures should be expected. To support this claim, it may be worthwhile to examine what the IAT literature has to say on the subject.

Nosek (2007) provided numerous examples of where implicit and explicit tests converge and diverge. He showed the strongest correlations for political opinions, such as IAT's structured around pro-choice/pro-life, democrats/republicans, and feminism/traditional values. The weakest correlations were found for measures of implicit bias involving Asians/Whites, thin/fat, and tall/short. But what is the essential difference between these two sets, and how can we understand this difference in effects obtained with different types of

target stimuli at a process level? One explanation might point to the fact that political questions typically reach a high level of specificity, as advocated by Ajzen and Fishbein (1977). In other words, the level of specificity in the question likely controlled responses that are equally specific and controlled by clear and salient verbal contingencies. Specific questions exert specific contextual control over answers. In contrast, broad sweeping non-specific questions attempt to tap into broad and poorly defined verbal repertoires (e.g., do you care about the protection of the environment?). This latter type of question may be of genuine interest to the researcher but will lead to much more behavioural variability across participants and administrations. This is because, such questions exert poor contextual control, and allow for too broad a range of answers that could be coherent with the question. In technical terms, there is not a sufficient deployment of relational (Crel) and functional (Cfunc) cues in the question for the listener to know precisely how to answer and to which behaviours it is referring in their verbal past. The listener is also not in a position to conduct an empirically sound functional analysis of their verbal history upon the presentation of such a question and is likely to resort to immediately available cliché answers of the type that are usually socially reinforced by listeners (e.g., Of course I care about the environment).

Interestingly, some empirical work has been conducted to examine response differences at the neural level between directly trained and derived relations. Specifically, Schlund et al. (2008) demonstrated that symmetrical relations (i.e., direct) elicit activation in the parahippocampus, while transitive and equivalence relations (i.e., derived) elicit bilateral activation in the anterior hippocampus. Therefore, it is likely that different but related neural processes are occurring in response to probes for relations of different types. It stands to reason that these different response types could coexist. The main point here, however, is that they may well be in competition and that explicit and implicit tests may be suitable for measurement of relations of different types. Indeed, early research in derived relational

responding provided some support for the idea that verbal relations do not hang together perfectly as single units but can tolerate a degree of incoherence across different nodal distances (e.g., symmetrical and transitive relations among a set of stimuli).

Specifically, several studies have shown that symmetrical and transitive relations need not cohere within the same equivalence class. Although reversing the symmetrical relations underlying an equivalence class may be effective, this may still fail to result in a reversal of the transitive relation. For example, in one study by Pilgrim and Galizio (1990), two three-member equivalence classes were established. The researchers then exposed participants to one or more changes in the reinforcement contingencies controlling the baseline discriminations. Specifically, in one condition they reversed the A-C relations (i.e., choice of C2 was reinforced and C1 punished when A1 was the sample, the reverse when A2 was the sample). In another condition they randomised the A-C relations by reinforcing and punishing the choice of C1 and C2 equally often in the presence of A1 and A2. In yet another condition, they reversed the original A-B and A-C relations. Despite establishing a stability in responding under the direct control of those novel contingencies, Pilgrim and Galizio (1990) showed that performances on the transitivity probes remained consistent with the initial equivalence class preceding the interventions. They similarly showed that baseline and symmetry probes were extremely sensitive to baseline modifications, while leaving the transitivity/equivalence probes consistent with the initial equivalence class. However, these findings may be a result of only partial reversal of the initially trained relations. It has been shown that where all initially trained relations are reversed, equivalence reversal is reliably produced (Smeets et al. 2003).

In understanding why baseline and derived relations do not always cohere, it is important to remember that in Relational Frame Theory a relational frame is itself a form of

generalised operant behaviour (Gómez et al. 2002). That is to say, the emergence of stimulus equivalence (or any other relational frame) is not something a subject can innately do but arises as a result of past reinforcement. Once the frame has been taught in one context it can generalise to a multitude of others. Given this, frames should in theory be susceptible to contextual control i.e., there may be contexts in which a previously learned relational frame does not emerge because in that context a different response has been reinforced. There is some evidence to support this assertion arising from research into 'breaking equivalence' (i.e., the conditions under which equivalence will not emerge) which we shall now explore.

Gómez et al. (2002) first trained their subjects on a set of baseline relations (A1-B1, B1-C1, A2-B2, B2-C2), followed by training for symmetry (B1-A1, B2-A2, C1-B1, C2-B2), transitivity (A1-C1, A2-C2) and finally to break equivalence (C1-A2, C2-A1). All training was done using a standard matching-to-sample procedure with feedback. Importantly, in trials where subjects were required to break equivalence a contextual clue was incorporated into the trial (e.g., "&&&&", "@@@@" at the top of the screen). The same set of trials (excluding baseline trials) was then readministered without feedback, if a subject successfully broke equivalence at this stage they proceeded to the next set of trials. If they did not training (with feedback) was repeated until the behaviour occurred on trials without feedback. Successful subjects were then taught the baseline relations again with an entirely new set of stimuli. Without any further training they proceeded straight to the no feedback test phase. This phase was to test whether the subjects would generalise the break equivalence frame in the presence of the contextual clues without further training. Several such generalisation tests were administered. Of the five participants, two exhibited the break equivalence pattern on their first generalisation test, the remainder exhibited the pattern on subsequent generalisation tests. In summary, even a very basic frame such as equivalence can be taught to occur in some contexts but not in others. In principle there is always a context which controls current

responding, and in naturalistic language we can expect to find countless inconsistencies, incoherencies and contradictions due to the nigh infinite number of contingencies which control behaviour in different contexts.

Of importance however, the preceding studies provide support for the idea that under certain conditions directly trained or symmetrical relations between words can coexist within relational networks of transitive relations with which they are incoherent. This phenomenon has implications for our interpretation of the results of implicit tests. Tests designed to index the strength of a relationship between stimuli in a symmetrical relation may indicate compatibilities or incompatibilities that are incoherent with the results of a test designed to index the strength of transitive relations involving some of the same stimuli and related stimuli. In other words, it is of significant importance to consider the nature of the relations being assessed. The potential fallibility of implicit test outcomes should not be underestimated if one does not have a good conceptual grasp of, and ideally empirical control over, the functional nature of the verbal relations being indexed in that test. Similarly, if explicit tests probe only for symmetrical relations, they may then overlook transitive relations that are in contradiction to the probed symmetrical relations. At least a portion of the reason that implicit tests can detect transitive relations over symmetrical ones is likely a result of the time pressure built into implicit tests.

Indeed, the very reason for response time pressure in an implicit test such as the FAST is to allow remote contingencies to take control of the behaviour. The reader may recall the behavioural definition of implicit provided in chapter 1, the example given was that someone may be taught a relation between “Irish people” and “Drunkards” and separately a relation between “Drunkards” and “Ignorance”. Therefore, the relation between Irish people and ignorance is implied by the relational network. It is thought that under time pressure, the

remote relation (i.e., the contingency controlling the behaviour) between Irish and ignorance is more likely to be observed. This is because time pressure effectively eliminates the control of secondary verbal responses that mediate the ultimate pressing of a button on the keyboard. It might be the case as the REC model (Barnes-Holmes D., Barnes-Holmes Y, Power, Hayden, Milne & Stewart, 2010) suggests that the extended amount of time permitted in an explicit test allows for non-automatic responses consisting of verbal responses to one's verbal response to the initial likely response. That is to say, when a subject is asked if they believe White people are more intelligent than Black people, their initial "gut response" might be to respond yes. Before this response takes place however, a secondary verbal response occurs in contradiction to their initial response "I had better not say that". This secondary response moderates the initial one and may cause them to respond in the opposite way. This secondary response may be a result of the individual encountering aversive consequences to their initial response in the past, e.g., being called a racist, or suffering social ostracization for expressing such views. Despite sharing a common stimulus (i.e., being asked to compare White and Black people), the responses are not merely topographically distinct, they are functionally distinct. As discussed, the initial response may elicit a punishing consequence, while the secondary response (and thus resulting opposite behaviour) avoids this consequence (i.e., it is negatively reinforcing), or may even be reinforced e.g., being praised for expressing liberal values. The responses do not merely differ in form but also in consequence, and as the consequence may vary with context (e.g., different friend groups with differing social values or experimentally with time pressure) they can be said to be functionally distinct.

The sort of elaborated responding just described is impossible under the time pressure of a test such as the FAST. This however, is at present a very difficult hypothesis to test. A simpler way of viewing this is simply that under time pressure, the dominant contingencies will exert the most pressure over behaviour, and that of course behaviour may become more

variable over longer periods of time if response windows are extended. The current account is not openly rejecting the idea that it is possible for an individual, given enough response time, to form a rule involving responding in the opposite way to which they initially begin to respond. This response pattern might be referred to as a social desirability bias. However, let us refocus on the suggestion of the REC model concerning what happens when the test taker is indeed under time pressure due to response windows or strict instructions.

In such a case as the aforementioned, it is suggested and indeed supported by the REC model, that only one set of contingencies would control behaviour, and these contingencies would be the ones most well established in the history of the test taker. These contingencies may control behaviour in a relatively indirect way, or in quite a direct way that has involved direct reinforcement of the particular verbal relations. What the REC model fails to distinguish between, are the cases in which an individual with limited response time has had a history of both reinforced verbal responses, and a history of responding to relational networks that merely imply further yet to be derived relations. By this it is meant that their history of learning logically supports a derived relation, but this specific relation has never been reinforced e.g., if taught A > B and B > C, then the relation A > C is logically supported but has never been explicitly derived. As has already been outlined, symmetric and transitive relations need not be coherent within the same equivalence class. As, a real-world example, consider an individual who has learned from their family that they should not like Black people. However, everyone around them, including their family, teaches them to never state as such and to always say that they do in fact like Black people. Now the question arises as to which of these contingencies will dominate this individual's responses. In an explicit test, it is likely they will respond in the way which they have been directly taught to respond i.e. "I like Black people". In an implicit test however, it is questionable as to which contingencies will dominate responding. The suggestion being made here for the first time, is that responses will

emerge exactly as they would under the conditions of any competing contingencies. That is, it is simply a matter of which set of contingencies have exerted the most control in the past and which responses have the greatest behavioural momentum/probability. Of course, it would be very difficult to discern on a case-by-case basis ahead of the administration of an implicit test whether or not the relations of interest have been well rehearsed or are derived in nature and perhaps even incoherent with socially acceptable and well-rehearsed verbal practices.

While the point stated above concerning when to use implicit/explicit tests remains a good rule of thumb, it is important to remember that the behaviour a stimulus elicits is always dependent on context. It may be the case that people who hold a prejudice against another group such as Black people are not universally prejudiced in every context. For example, in the context of employment a recruiter may withhold a job offer to a qualified Black person on the basis of their ethnicity, but in another situation that same recruiter may refute the stereotype that Black people are more likely to be criminals. In short, even individuals we would consider to be racist may not be universally racist in all situations. In real-world situations the momentum argument applies, if we were to look at all of the responses an individual has produced in similar contexts (i.e., under similar contingencies of reinforcement) the one with the longest history of reinforcements is the one most likely to be produced. Perhaps our recruiter is congratulated by her superiors for a good batch of recruits whenever she hires exclusively White people but ignored or even admonished when she hires Black people. In the stereotype situation, she may be congratulated by her friends for standing up against racial stereotypes. In both situations however, the response which is most likely to be produced is that which has the longest history of reinforcement. But how might any of this apply to implicit tests?

The FAST and other implicit tests are stripped of almost all social context, thus there are no cues to guide a participant's behaviour. As the FAST merely requires responding to images of different people and a particular class of words in a common way it can likely be assumed that these positive or negative response patterns are dominant across most of the history of the participant. In the case of our recruiter, she may produce a response pattern indicative of negative evaluations of Black people, but of course this does not necessitate that she is universally racist in all situations. The methodology of the FAST functions precisely by decontextualising relational responses to ascertain what the dominant pattern of responding to specific stimuli are when free of potential social cues. If by contrast, the research question specified a social cue then the FAST can easily be adjusted to that cue. In the case of intelligence, it would merely be a matter of loading the FAST with stimuli concerning intelligence e.g., smart, stupid, genius, idiot, rather than general positive/negative stimuli. Cartwright et al., (2016) adopted this approach in the context of gendered stereotypes, so while in this set of experiments the FAST functioned as a general litmus test, it can be made context dependent with only minor adjustment. The question remains however, as to whether there are situations in which an explicit test is better suited than an implicit one. As a bold and progressive suggestion, it might be offered here that the defining criterion on when to decide upon the use of an implicit versus an explicit measure comes down to how the relations of interest were trained. Accordingly, if relations of interest can be reasonably estimated to be reinforced in the verbal history of the test taker, then an explicit test should be employed. Alternatively, if the relations of interest are likely to be derived and supported only indirectly by past verbal contingencies, then an implicit measure should be employed. The more indirectly the means by which the verbal relations have arisen, and the less likely it is that the verbal relations have been directly reinforced, the more useful an implicit test may be to ascertain the verbal contingencies at work in the repertoire of the

individual test taker. In contrast, the more likely it is that the relations of interest have been reinforced, the more likely it is that verbal contingencies implying the derivation of competing verbal relations will fail to exert control in the test context.

Now that the various implications of findings of the present set of studies have been explored, it would seem time to examine any potential issues which may have impacted the present study. In so doing, some avenues for future research may also be suggested.

4.6 Potential Methodological Issues

4.6.1 Convergent Vs. Divergent Validity

There were differing approaches to validity between the first and second experiment which deserve some further explanation as this was not explored in depth when explaining the hypothesis. The standard reasoning behind any assessment of a novel test's validity is to compare it with a pre-existing test which is thought to measure the same construct, should the tests correspond with one another then this lends validity to the novel test. While it was initially conjectured on the basis of the SJT rationale that the FAST and Modern sexism scale would diverge the results contradicted this claim, and subsequent research noted that gender may be an exception to the general SJT rule (Rudman & Goodwin, 2004). To be clear, that rule being, that groups who face discrimination will self-report ingroup bias while implicit tests may reveal a contradictory outgroup bias. Therefore, the fact that the FAST and Modern sexism scale co-varied in Exp 1. lends convergent validity to the FAST as it was clear the FAST was tapping into the same attitudes about women as the explicit test was. Even if the results of the correlation were to be discarded entirely, the general trend of the data exhibits the same results, namely that men scored higher on both the explicit and implicit test, indicative of more prejudicial attitudes towards women. Women on the other hand averaged

lower results on both tests indicative of less prejudice towards women, in sum a degree of convergent validity for the FAST was established in the first experiment.

Of course, this leads us to the question of why this approach to validity was abandoned in the second experiment. The reason again lies in the SJT rationale at work in this experiment. Whereas validity is usually established when two metrics converge in their results, it is by now well known that in the domain of racial prejudice explicit and implicit measures often diverge. Specifically, several large-scale studies (See Nosek et al., 2002; Jost et al., 2004) have demonstrated that minority groups often self-report in-group biases but implicitly favour the dominant outgroup. Similarly, these studies have shown that the dominant group will self-report less in-group bias but implicitly favour their own ingroup. Therefore, it can be said with confidence that at least in the domain of racial bias a divergence between self-report and implicit measures is likely to be found. For Exp. 2 even if the lack of correlation were to be discarded on the basis of low power the general data trends are still indicative of a divergence. On average both White and Non-White groups self-reported low racial bias, while their FAST RFD scores indicated the opposite, a bias in favour of White people. The lack of covariance between the implicit and explicit measures in Experiment 2 therefore lend divergent validity to the FAST on the basis of findings within SJT.

4.6.2 The RFD Scoring Method

The reader may wish to note that in the current study the RFD scoring system was used, this as opposed to the difference in learning slopes across blocks method reported in published FAST studies. It was mentioned in Chapter 1 of this thesis that the slope scoring method carries risks of inflation or deflation based on the level of random responding in which a participant engages. In contrast, the RFD method protects against inflated or deflated

scores caused by bursts of rapid responding through calculating a learning rate for the block in terms of the correct response per minute rate, corrected by the incorrect response per minute rate necessarily incurred in random responding.

As the RFD scoring method is a relatively novel approach to the scoring of an implicit measure we might reasonably question whether this renders the results incomparable to other studies or in some way diminishes the contribution of this research to the wider literature. However, it must be remembered that from the outset the developers of the FAST method have emphasised the importance of researchers employing their own scoring algorithms and procedural modifications, so long as these are openly described and acknowledged. In other words, from the within the behaviour-analytic tradition, it is recognised that the data produced by all procedures are a result of the contingencies at work in the test. Therefore, no test is a direct window into an extant construct which it indexes in some objective way (give or take some random test error). To that extent, there can be no perfect scoring algorithm. Ideal scoring algorithms will only be identified based on pragmatic utility. Seeing as the pragmatic utility of the FAST has not yet been identified and has merely been explored in this research, it would be premature to claim that one particular scoring algorithm is proprietary and must be used in all future research. In short, the FAST is a method, a paradigm, and a research framework. It is not a psychometric test and does not measure a construct. To that extent, the scoring measure should, and will be varied according to the needs of the researcher without in any way jeopardizing the core process at work in the test. In line with the emphasis behaviour-analysts place on variation and selection in the evolution of behaviour, variation and selection should also take place in the development of the current methods.

4.6.3 Ethnicity Classification

The reader may have noted the binary classification of ethnicity in Experiment 2. For the purposes of analysis, all ethnicities were classified as either White or Non-White. It could be argued that the rather gross classification of participants as either White or Non-White obfuscated patterns of racial bias that might be particular to one or more ethnicities. It is certainly fair to say that there is no such ethnic group as Non-White, and one should not generalise about all individuals who do not identify as White. However, for both statistical convenience and due to the very small number of individuals who identified as anything other than White, it was considered pragmatic to combine these participants for the current analysis. Furthermore, it should also be remembered that Experiment 2 was not particularly interested in any one ethnicity, but rather in power relationships between dominant ethnic groups and all individuals who are not members of that dominant ethnic group. To that extent, the generalizations engaged in here are justified insofar as, while they do not speak directly to one ethnic group, these generalisations do speak collectively to ethnic minorities who are disempowered and disenfranchised relative to the White citizens of Ireland.

While the above method was sufficient for this particular experiment, there is an alternative approach which could prove worthwhile in future experiments into the phenomenon of outgroup favouritism. Rather than coding the data according to the gender or ethnicity of the participants a purely functional approach could have been employed. Namely, the conditions could have been coded in a simple ingroup/outgroup format. For example, rather than stating that a Black participant exhibited a White-positive bias, this could be rephrased so that no reference is made to the subject's ethnicity, merely that they exhibited greater fluency in either the ingroup or the outgroup block. The current method, while pragmatic, is admittedly inspired by the social-cognitive literature from which SJT is derived,

and a more functionally aligned approach might elect to use the ingroup/outgroup format just outlined.

4.6.4 Alternative Explanations for the Results of Experiment 1

The results of the first experiment were unusual both in terms of SJT and of conventional theories of ingroup bias. While female subjects displayed a strong ingroup bias according to their RFD scores, males showed no bias in either direction. While the results of Rudman and Goodwin (2004) go some way in explaining this, the results remain sufficiently unusual to warrant an examination of how the FAST was structured methodologically in this experiment. A deeper dive into the data reveals some interesting patterns that might help explain the gender difference in responding. Specifically, it was found that gender in and of itself had a significant impact on fluency scores. While it could be the case, that women are simply faster at responding in general than men, there is no real theoretical explanation for why this would be the case. Therefore, alternative explanations will be considered.

By digging into the raw data further, it was observed that the male incorrect responses per minute were approximately equal for both blocks (Con - 4.4, Incon - 4.34), while women demonstrated a noticeable decrease in error rates for the inconsistent block (Con – 3.91, Incon – 2.5). From this, it is easily deduced that, error rates were likely the primary influence on the observed differences in responding. The reader might recall at this point the experiments performed by Camp et al. (1967) and Rabbit and Rodgers (1977) outlined in the introductory section. The essential argument of these experiments taken together, is that errors are likely to lead to slower rates of responding, and errors tend to beget further additional errors. This observation may prove relevant to the current data patterns. Specifically, while women still produced fewer errors than men in both blocks overall, error rates were lowest and most notably diverged in the inconsistent block. Additionally, due to

the fact that males produced relatively equal error rates in both blocks, their overall rate of responding was likely inhibited. Their higher rate of incorrect responses accounts for the fact that males were slower than females in general. While this is apparent from the data, the overall difference in error rates between males and females remains unexplained.

The most transparent explanation for the divergence in error rates between males and females lies in their pre-experimental history. On the basis of their error rates, it would seem likely that women in this sample had more experience in relating both men and women in positive ways than their male counterparts. Nonetheless, their experience in relating women to positive verbal stimuli was still greater than their experience in relating males to positive stimuli. Indeed, this is not so much a matter of conjecture, but by definition insofar as the FAST was designed precisely to measure the fluency of relating words in particular ways. On the basis of this a possible explanation emerges. The women in this sample may have a greater history of responding to both men and women in a complementary fashion than their male counterparts, while having a greater still history of responding to women in a complementary fashion than men. Similarly, the men in this sample may have a weaker history of both complimentary and derogatory appraisals of both men and women than their female counterparts but are relatively equal in their patterns of responding towards both male and females.

The speculation above is of course purely intellectually satisfying in the current context and is not a substitution for rigorous empirical work to elucidate these relevant processes. However, working at the intersection between sound behaviour analytic laboratory work and application of our principles in the real world through translational research is always a somewhat uncomfortable position to be in. The translational researcher must tread carefully when venturing into domains of analysis usually dominated by terminology and

concepts emerging from an orthogonal philosophical paradigm. An argument of the sort being presented here would be unorthodox in a mainstream behaviour analytic journal. With that said, it may represent a small part of the discursive process of identifying potentially manipulable variables in the course of developing empirically sound and functional accounts of behaviour. Of course, there remains some possibility that something other than participants pre-experimental history may have influenced the results found in Experiment 1. It is possible that some facet of the experimental procedure may have had an impact on the observed findings.

Indeed, the following IAT study offers an alternative explanation suggested by empirical data. Ramos et al. (2015) presents an alternative explanation and raises the possibility of an error in the procedure. Ramos et al. performed an experiment with an IAT structured around stereotypes of men as competent and women as warm. The experimental manipulation in this study was as follows. Their sample was split into three groups of men and women. Each group was asked to perform a memory task where they observed, memorized and then recalled the association between six sentences and pictures. The pictures featured men and women interacting with each other. The sentences varied between groups. One group was presented with hostile sexist statements, another with benevolent sexist statements, and the last group arbitrary flower related sentences (i.e., the no sexism group). It was found that the women in the sexist statement groups demonstrated weaker gender stereotype bias, relative to the no sexist statements group, while men were unaffected by the exposure to sexist statements. This gender stereotype bias was measured using an implicit Go/No-Go Association Task. The authors reasoned that because women suffer the consequences of sexism in real life when exposed to it in a study context they may react in defiance to it. Of high relevance to this study, the sexist statements the experimental groups

were exposed to were derived from the Ambivalent Sexism Inventory, an explicit measure of gender stereotypic beliefs.

While the aforementioned study is an interesting explanation for the absence of a negative ingroup bias among women in Experiment 1 of the current study, it raises a further interesting issue regarding the order of the tests delivered in the current experiments. That is, given this consideration, it may have been suboptimal to administer the MS scale before the FAST in the experimental sequence. It could be the case, as Ramos et al. (2015) suggest, that by merely exposing the female participants to sexist statements from the questionnaire their implicit gender bias on the FAST could have been affected by the activation of an oppositional verbal repertoire. In behaviour analytic terms we would refer to this as counter control, a phenomenon that is quite well understood by both basic researchers and applied behaviour analysts. Indeed, coercion is ineffective precisely for this reason and has been written about extensively (Sidman, 1989). Future research might consider examining this account in the context of the FAST. In the short term, however, it may be more prudent for future studies to not present any explicit questionnaires until after the implicit tests have been delivered.

4.6.5 Employing Different Kinds of Stimuli in FAST Research

The choice of stimuli for the FAST in Experiment 1 may have been suboptimal for detecting gender bias amongst both the male and female participants. More specifically, the evaluative terms used here were taken from among those used by early implicit association test study stimulus sets and were broadly positive and negatively valanced verbal stimuli. In contrast, however, IAT studies into sexism have in the past used more specific stimuli representing common stereotypes of men and women. For example, Ramos et al. (2015) used verbal stimuli relating to stereotypes of men as competent and women as warm instead of

broadly positive and negative evaluative terms. This approach represents a more focused effort in detecting gender stereotypes and enhances the specificity of the test. Put simply, an individual may not have a broad dislike for one gender over the other, but it may nevertheless be the case that they view one gender as more competent or warmer than the other. However, it was estimated in the current study that such a pursuit of specificity might be at the cost of a general negative evaluation of females. That is, as a study assessing the utility of the FAST, and not knowing what form negative categorizations of females might take for participants, it seemed more conservative to search for general negative evaluations than for very particular negative evaluations. Nevertheless, future studies employing the FAST methodology might consider examining the formation of functional response classes between male and female target stimuli and class of stimuli representing orthogonal stereotypes regarding the genders. No doubt, different results will emerge from different tests using different stimulus sets, but in so doing it will help us map out the ways in which the vernacular maintains sexist behaviour and perhaps some specific ways in which it does not.

One early FAST development study based on the Watt et al. (1991) stimulus equivalence procedure has already suggested that evaluative term specificity can be used in innovative ways to identify a history of verbal behaviour. But more particularly membership of various social categories. Specifically, Roche et al. (2005) suggested the use of stimulus categories that may be more familiar to one group than another. The use of such specific stimulus categories can aid in the identification of group membership, insofar as knowledge of the meaning of the stimuli involved will lead to clearer differences in performance across two test blocks. In that study, employing an embryonic version of the FAST, the authors were interested in identifying the history of sexual interest in minors amongst a group of sex-offenders against children, a group of sex-offenders against adults, and a sample of non-sex-offenders. Rather than simply examining the relatedness of images of children faces with

sexual terms, the authors chose to examine the relatedness of images of children's faces with terms used for children almost exclusively by child sex-offenders. The idea was not so much to simply examine whether sexual and child related stimuli already participated in functional or derived verbal relations (e.g., Dawson et al., 2009; Gavin et al., 2012; Roche et al. 2012), but to examine specifically if that relation applied to terms used in a subculture. The recognition of such a subcultural term would itself be indicative of past paedophilic activity. This study was not peer-reviewed and was reported in a book chapter on sex offending, but the results were promising in terms of identifying the sex offenders against children by their implicit test performance. Future research using implicit tests to identify group membership might benefit from including such a strategy.

The preceding Roche et al. (2005) procedure, in effect, involved the tailoring of stimuli used in an implicit test to a particular population. However, it would also be possible to tailor stimuli to particular individuals and deploy a personalized test, such as has been achieved with idiographic IAT's (Greenwald & Farnham, 2000; Bluemke & Friese, 2012). It may be worthwhile for future FAST research to reconsider its generally nomothetic approach in favour of a more idiographic one. Nomothetic designs employ the same stimuli for all participants, while an idiographic one would use different stimuli for each individual participant. This would be in line with the philosophy of the FAST and behavioural research more generally. Stimulus control is always an individual matter and the use of global stimulus control methods in the absence of control of the history of each participant needs to be strongly justified. Such an approach has been taken thus far out of pure experimental convenience, but it is a broad-brush approach that lacks precise stimulus control. That is, using such generic words as love, peace, happy, filthy, rotten was assumed here to produce broadly the same affective and relational responses for most participants. This assumption was based entirely on fact that these words are widely used and that most participants are

members of broadly the same culture. However, it is easy to see how responses could vary both between and across participants to each of these words given the various vicissitudes of each individual participant's learning history.

In an idiographic approach, subjects themselves would select the words to be used as stimuli in the FAST prior to its administration. Such a selection procedure could include free association tests, or simple categorizations of already narrowed batteries of potential stimuli. Tests of the usual social meaning of each word could also be conducted as is done in the categorization tests before both the IAT and the IRAP. This procedure is designed to ensure that participants categorize these words in ways typical of other participants in the study. This alone is a procedure that the FAST would do well to adopt so as to at least to protect the integrity of the nomothetic approach. However, it would add considerable information about the functions of stimuli for each individual participant if they were to choose their own exemplars for each category set before the commencement of the FAST itself. This would have the advantage of enhancing stimulus control and ensuring that stimuli evoke relevant affective response functions. It would also ensure that all stimuli in the exemplar groups constituted an already existing verbal stimulus class. Furthermore, the pre-test stimulus categorisation could serve as a screening procedure, wherein participants who have no familiarity with discriminating on the bases of the selected category labels would not advance to the critical test phase. As we are interested in measuring participants' history with the verbal classes of interest, there is little point in testing participants with no such history and their incorporation in the data would likely only obscure the effects of interest. Effectively, this would clean the data through tighter stimulus control; a method far more in line with behaviour analytic philosophy than a complex statistical algorithm such as that employed in the IAT. If future researchers wished to employ a FAST with a more idiographic design, they

would do well to consider how the IAT has been employed in similar ways in the past with regard to stimulus categorisation tests.

Greenwald and Farnham (2000) used a list-based idiographic self-concept IAT, wherein subjects were allowed to choose their stimuli from lists of "me" and "not me" items. They were also allowed to delete items from the evaluative list. That is, given a list of pleasant and unpleasant items, subjects were allowed to delete items from each list that in their view did not represent the respective category. Using this form of IAT, the researchers found a higher degree of correlation between IAT D-score indices and an explicit measure of self-concept than they did for a typical nomothetic self-concept IAT. Interestingly, however, the authors concluded that the additional time and effort required to produce the idiographic IAT for each individual participant was too costly for the return in increased score accuracy. This is a crucial perspective to highlight, in that it puts in sharp relief the difference between the social cognitive researcher and the behaviour analyst. That is, the social-cognitive research is content with lower levels of stimulus control where group level statistical significance can still be obtained. Behavioural researchers, in contrast, should be mindful to do the opposite. Specifically, they should focus on an increase in stimulus control for each individual participant, even where this leads to inter-subject variance that threatens group level statistical significance.

Another study employing the idiographic approach was conducted by Bluemke and Friese (2012) in the domain of self-concept. These researchers did not use a complex stimulus selection procedure, but simply used personal details for stimuli, including the subjects' first names, family names, birthdays etc. to represent the "me" concept. Each subject also completed a generic self-concept IAT. It was their conclusion that an idiographic IAT in this domain was more valid, in that it produced index scores that correlated more

impressively with those produced by explicit self-report measures. Hofmann et al. (2005), in a meta-analysis on IAT research found a higher degree of correlation with explicit measures for idiographic IAT's than generic nomothetic ones (r = .32 vs. .15). However, this analysis relied on only 6 idiographic studies. It is worth noting, however, that one study employing an idiographic IAT to examine levels of anxiety (Stieger et al., 2010) found no superior performance of the idiographic method over the nomothetic.

In behaviour analytic research more broadly, it has been recognized that verbal behaviour research likely needs to consider idiographic approaches in translational research areas. For instance, Eilertson and Arntzen (2020) reported on a study designed to examine the transfer of pain functions produced by visual stimuli through laboratory-controlled equivalence relations. Rather than choose visual stimuli that the researchers guessed would be evocative of a covert pain response for most participants, they tailored the painful stimuli used for each participant. They achieved this by having subjects rate which images, from an array of images of a needle injection, they thought to be most painful and which they thought would be least painful on a Likert scale. They then implemented a training structure employing these stimuli which will now be outlined.

Eilertson and Arntzen (2020) trained participants in six conditional discriminations with abstract shapes as stimuli and tested for the formation of three three-member equivalence classes in a one-to-many training structure (i.e., A-B, A-C relations were trained). The image deemed most painful by the subject was then labelled D1, the least painful D2, and a third image of the needle replaced with a Q-tip as D3. They then extended the original equivalence class by training a D stimulus to a respective A stimulus. Following the D-A training, testing for the formation of three four-member equivalence classes was conducted using a matching-to-sample test. After the test for emergent relations subjects were

brought to a different room containing three identical bottles of water labelled with the B stimuli. The experimenter then instructed them to choose a bottle and bring it to them. The experimenter waited outside the room while they did so. After this phase had been completed, participants were then presented with sheets depicting the B stimuli and asked to rate them in terms of painfulness on a Likert scale. They found that the B and D stimuli were not rated significantly differently (i.e., a transfer of pain response functions), and that participants avoided the bottle labelled with B1 (i.e., the B stimulus participating in a derived relationship with D1, the needle image they deemed most painful), but did not differentiate between the bottles labelled with the B2 and B3 stimuli. This study demonstrates that in principle stimuli can be tailored to individual subjects to ensure the experimentally intended connotative meaning and therefore guarantee high yield rates in the transfer of functions. Such a design can aid in reducing between-subject variability in procedure outcomes through increased stimulus control.

In a study conceptually similar to the previous one, Arntzen and Eilertson (2020) used an idiographic style approach in an experiment on using stimulus equivalence to teach nutritional skills. That is, a number of exemplars from different nutritional categories of food were presented to subjects at baseline. Those foods which subjects could not correctly categorize in terms of their carbohydrate content were used in the conditional discrimination training. The idea here was to teach the participants to categorize these food items correctly in order to make healthy eating decisions in the future. Had the researchers employed verbal stimuli related to well-known food items whose carbohydrate content was known little could have been achieved by the study. In contrast, the researchers ensured that all of the stimuli involved were stimuli the participants were not yet able to categorize in terms of carbohydrate context (into one of three categories, low, medium, and high content). Given the increased stimulus control exerted during the procedures such methods considerably improve

the quality of research designs and complement our ability to draw conclusions about relevant behavioural processes.

Taking together the idiographic studies conducted with the IAT, the Roche et al. (2005) argument, and now the behavioural experiments conducted with tailored stimuli, we can see that there is an entirely new direction available for future FAST research. While a full outline of an idiographic FAST is outside the scope of the present thesis, in principle, such an approach is entirely possible. Such an approach could have great potential in increasing the specificity of the FAST. To bring this back within the scope of the present study we might briefly consider the potential advantages such a method might have had in Experiment 1.

In the context of assessing sexism, the stimuli could be tailored for a cohort to include words typically used in a misogynistic way to refer to women. This would not only assess these relations in their own right, but might indicate the use of those terms in the verbal repertoire of the test takers. This would be indicated by whether the functional response classes containing the specified stimuli can be formed with relative ease. Of course, as suggested by the various accounts outlined in the foregoing sections, the challenge remains that any homogenous class of verbal stimuli may not represent well the stereotypes (i.e., relational responding biases) inherent in the verbal behaviour of the participants. For instance, as outlined, Jost et al. (2004) have suggested that modern forms of sexism are often more benevolent in nature than older forms. Thus, the stereotypes involved in the sexist language will vary from person to person and generation to generation. This could be combated by a selection procedure involving a free association test or a simple categorization task. These methods could involve a list of potential stereotypes narrowed down by the researcher and would serve to enhance stimulus control in that it would ensure the stimuli invoke relevant affective response functions for each participant.

Of course, in some contexts the functions of stimuli are practically universal; and researchers may still wish to exercise discretion in complex and high-end translational research to continue with the nomothetic approach. Nevertheless, the field would be well served if different researchers use different approaches for the purpose of indexing specific stereotypes and prejudices. This would allow comparison across studies of the nomothetic and idiographic approach in particular contexts.

4.7 Concluding Remarks

Outside of some of the methodological issues noted, the FAST performed remarkably well for a test that has only recently begun to be applied in the assessment of non-laboratory-controlled relations. Its ability to detect naturalistic stimulus relations was noted by Cartwright et al. (2016) but further supported here. The current thesis sought to assess the utility of the Function Acquisition Speed Test as a new behaviourally oriented measure of "implicit" bias. While the test was developed primarily under laboratory conditions, the two experiments reported here aimed to assess its utility in translational research aimed at building bridges with those in the social psychological research sphere.

The FAST proved itself to be somewhat sensitive to naturalistic stimulus relations in the context of sexism and racism. The results of Experiment 2 in particular point to its potential utility in the place of explicit self-report measures in certain research contexts. While the current studies did not compare the FAST to the IAT directly, it would appear that a behaviour-analytic alternative is at least a viable option. Only further research will reveal the advantages and disadvantages of the current method, but no doubt at present the FAST approach is more fitting for research in the behavioural domain than is the IAT.

The current study also did not and cannot comment on the relative merits of the IRAP approach in complex social research. While the RFT-inspired approaches of both tests are

similar, the formats of the tests diverge enormously. It has been expressed from the outset of the FAST research program however, that the aim was to develop a test from the ground-up. Consequently, the FAST research program has sought to avoid becoming distracted by progressive leaps based on theory, or the procedural aspects of other tests that are not yet fully understood in functional analytic terms. However, it is certainly worth noting that the FAST procedure could now safely be extended to encompass relational components. For example, the test could continue to establish functional response classes across two blocks, but stimulus classes could include relational evaluation stimuli as exemplars. To elaborate on this, participants could be trained in the formation of functional response classes between the phrase "women are smart" and the word "true" as well as the classes "men are smart" and the word "false". On an alternate training block the true and false keyboard response positions could be reversed, so that in effect we could index the relatedness of more complex concepts than mere words participating in equivalence classes. Such procedures could easily be elaborated to involve relations of other kinds. Crucially, however, the research would progress slowly and always in a bottom-up fashion. Conceivably, a procedure as elaborate, nuanced and as far reaching as the IRAP could be developed. However, given over a decade of experimentation, both published and unpublished, in the development of the FAST method, it likely measures the same sorts of effects as measured by the IAT. The key differences lie in a format that is notably distinct in multiple ways and involves a different conceptualisation as well as scoring method. The same is likely to be true of a relational FAST (FASTr). Indeed, as such research developments progress, we can be more confident that any procedure we offer in translational research will be well-grounded and achieve the standards of empirical validation at every step that we expect of our methods in the experimental analysis of behaviour. Insofar as the current thesis outlined two simple

experiments to help begin that process of translational research and development, it has already made its contribution to the field.

References

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888–918. https://doi.org/10.1037/0033-2909.84.5.888

Arntzen, E., & Eilertsen, J. M. (2020). Using stimulus-equivalence technology to teach skills about nutritional content. *Perspectives on Behavior Science*, *43*(3), 469–485. https://doi.org/10.1007/s40614-020-00250-2

Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Delacorte Press.

Barchard, K. A., & Williams, J. (2008). Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. *Behavior Research Methods*, *40*(4), 1111–1128. https://doi.org/10.3758/brm.40.4.1111

Barnes-Holmes, D., Barnes-Holmes, Y., Luciano, C., & McEnteggart, C. (2017). From the IRAP and REC model to a multi-dimensional multi-level framework for analyzing the dynamics of arbitrarily applicable relational responding. *Journal of Contextual Behavioral Science*, 6(4) https://doi.org/10.1016/j.jcbs.2017.08.001

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

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). *A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model*. *The Psychological Record, 60*(3), 527-542. https://doi.org/10.1007/BF03395726.

Barnes-Holmes, D., Hayden, E., Barnes-Holmes, Y., & Stewart, I. (2008). The Implicit Relational Assessment Procedure (IRAP) as a response-time and event-related-potentials methodology for testing natural verbal relations: A preliminary study. *The Psychological Record, 58*(4), 497-516.

Barnes-Holmes, D., Murtagh, L., Barnes-Holmes, Y., & Stewart, I. (2010). Using the Implicit Association Test and the implicit relational assessment procedure to measure attitudes toward meat and vegetables in vegetarians and meat-eaters. *The Psychological Record, 60*(2), 287–305. https://doi.org/10.1007/bf03395708

Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst, 19*(2), 163–197. https://doi.org/10.1007/bf03393163

Bluemke, M., & Friese, M. (2012). On the validity of idiographic and Generic Self–Concept Implicit Association tests: A core–concept model. *European Journal of Personality, 26*(5), 515–528. https://doi.org/10.1002/per.850

Bordieri, M. J., Kellum, K. K., Wilson, K. G., & Whiteman, K. C. (2015). Basic properties of coherence: Testing a core assumption of relational frame theory. *The Psychological Record, 66*(1), 83–98. https://doi.org/10.1007/s40732-015-0154-z

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? evaluating the inference of prejudice in the implicit association test.

Journal of Personality and Social Psychology, *81*(5), 760–773. https://doi.org/10.1037/0022-3514.81.5.760

Brunel, F. F., Tietje, B. C., & Greenwald, A. G. (2004). Is the implicit association test a valid and valuable measure of implicit consumer social cognition? *Journal of Consumer Psychology*, *14*(4), 385–404. https://doi.org/10.1207/s15327663jcp1404_8

Cabrera, I., Márquez-González, M., Kishita, N., Vara-García, C., & Losada, A. (2020). Development and validation of an implicit relational assessment procedure (IRAP) to measure implicit dysfunctional beliefs about caregiving in dementia family caregivers. *The Psychological Record*, *71*(1), 41–54. https://doi.org/10.1007/s40732-020-00445-8

Camp, D. S., Raymond, G. A., & Church, R. M. (1967). Temporal relationship between response and punishment. *Journal of Experimental Psychology*, *74*(1), 114–123. https://doi.org/10.1037/h0024518

Cartwright, A. (2013). *Developing the function acquisition speed test for homonegativity*. [Masters thesis, National University of Ireland Maynooth]. https://mural.maynoothuniversity.ie/5388/

Cartwright, A., Roche, B., Gogarty, M., O'Reilly, A., & Stewart, I. (2016). Using a modified function acquisition speed test (FAST) for assessing implicit gender stereotypes. *The Psychological Record*, *66*(2), 223–233. https://doi.org/10.1007/s40732-016-0164-5

Central Statistics Office. (2019, July 4). *Equality and discrimination 2019 - CSO - central statistics office*. CSO. Retrieved October 5, 2021, from https://www.cso.ie/en/releasesandpublications/er/ed/equalityanddiscrimination2019/.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals. *Journal of personality and social psychology*, *83*(6), 1314.

Cullen, C., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). The implicit relational assessment procedure (IRAP) and the malleability of ageist attitudes. *The Psychological Record*, *59*(4), 591–620. https://doi.org/10.1007/bf03395683

Cummins, J. (2017) *Quantifying differential stimulus relatedness using the Function Acquisition Speed Test.* [Masters thesis, National University of Ireland Maynooth]. https://mural.maynoothuniversity.ie/9901/

Cummins, J., & Roche, B. (2020). Measuring differential nodal distance using the function acquisition speed test. *Behavioural Processes*, *178*, https://doi.org/10.1016/j.beproc.2020.104179

Cummins, J., Tyndall, I., Curtis, A., & Roche, B. (2019). The function acquisition speed test (FAST) as a measure of verbal stimulus relations in the context of condom use. *The Psychological Record*, *69*(1), 107–115. https://doi.org/10.1007/s40732-018-0321-0

Cummins, J., Roche, B., Tyndall, I., & Cartwright, A. (2018). The relationship between differential stimulus relatedness and implicit measure effect sizes. *Journal of the Experimental Analysis of Behavior*, *110*(1), 24–38. https://doi.org/10.1002/jeab.437

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Dawson, D. L., Barnes-Holmes, D., Gresswell, D. M., Hart, A. J., & Gore, N. J. (2009). Assessing the implicit beliefs of sexual offenders using the implicit relational

assessment procedure. *Sexual Abuse*, *21*(1), 57–75. https://doi.org/10.1177/1079063208326928

De Houwer, J. (2006). What are implicit measures and why are we using them. *Handbook of Implicit Cognition and Addiction*. R. W. Wiers and A. W. Stacy, Sage Publications.

Der, G., & Deary, I. J. (2006). Age and sex differences in reaction time in adulthood: results from the United Kingdom Health and Lifestyle Survey. *Psychology and aging*, *21*(1), 62.

Drake, C. E., Seymour, K. H., & Habib, R. (2016). Testing the IRAP: exploring the reliability and fakability of an idiographic approach to interpersonal attitudes. *The Psychological Record*, *66*(1), 153-163.

Dymond, S., & Roche, B. (Eds.). (2013). *Advances in relational frame theory: Research and application.* New Harbinger Publications, Inc.

Eilertsen, J. M., & Arntzen, E. (2020). Tailoring of painful stimuli used for exploring transfer of function. *The Psychological Record*, *70*(2), 317–326. https://doi.org/10.1007/s40732-020-00381-7

Fields, L., Arntzen, E., Nartey, R., & Eilifsen, C. (2012). Effects of a meaningful, a discriminative, and a meaningless stimulus on equivalence class formation. *Journal Of The Experimental Analysis Of Behavior*, *97*(2), 163-181. https://doi.org/10.1901/jeab.2012.97-163

Gavin, A. (2008) *The Functional-Analytic Development of a Test for Behavioural History using the Concept of Derived Stimulus Relations.* [PhD thesis, National University of Ireland Maynooth] https://mural.maynoothuniversity.ie/1877/

Gavin, A., Roche, B., & Ruiz, M. R. (2008). Competing contingencies over derived relational responding: A behavioral model of the Implicit Association Test. *The Psychological Record, 58*(3), 427–441. https://doi.org/10.1007/bf03395627

Gavin, A., Roche, B., Ruiz, M. R., Hogan, M., & O'Reilly, A. (2012). A behavior analytically modified implicit association test for measuring sexual categorization of children. *The Psychological Record, 62*(1), 55-68. https://doi.org/10.3402/snp.v2i0.17335

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2014). Bayesian data analysis (3rd ed.). Florida: CRC Press.

Glenn, S. S. (1988). Contingencies and metacontingencies: Toward a synthesis of behavior analysis and cultural materialism. *The Behavior Analyst, 11*(2), 161-179.

Golijani-Moghaddam, N., Hart, A., & Dawson, D. L. (2013). The implicit relational assessment procedure: Emerging reliability and validity data. *Journal of Contextual Behavioral Science, 2*(3-4), 105-119.

Gomez, S., Barnes-Holmes, D., & Luciano, M. C. (2002). Generalized break equivalence II: Contextual control over a generalized pattern of stimulus relations. *The Psychological Record, 52*(2), 203-220.

Greenwald, A. G, & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review, 102*(1), 4. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., & Farnham, S. D. (2000). Using the implicit association test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology, 79*(6), 1022–1038. https://doi.org/10.1037/0022-3514.79.6.1022

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, *71*(1), 419-445.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., Banaji, M. R., & Klauer, K. C. (2005). Validity of the salience asymmetry interpretation of the implicit association test: Comment on Rothermund and Wentura (2004). *Journal of Experimental Psychology: General*, *134*(3), 420–425. https://doi.org/10.1037/0096-3445.134.3.420

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Grey, I. M., & Barnes, D. (1996). Stimulus equivalence and attitudes. *The Psychological Record*, *46*(2), 243.

Hall, G., Mitchell, C., Graham, S., & Lavis, Y. (2003). Acquired equivalence and distinctiveness in human discrimination learning: Evidence for associative mediation. *Journal of Experimental Psychology: General*, *132*(2), 266–276. https://doi.org/10.1037/0096-3445.132.2.266

Hattie, J., & Cooksey, R. W. (1984). Procedures for assessing the validities of tests using the "known-groups" method. *Applied Psychological Measurement*, *8*(3), 295–305. https://doi.org/10.1177/014662168400800306

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. Springer Science+Business Media, LLC.

Hayes, S. C., & Brownstein, A. J. (1986). Mentalism, behavior-behavior relations, and a behavior-analytic view of the purposes of science. *The Behavior Analyst*, *9*(2), 175–190. https://doi.org/10.1007/bf03391944

Hayes, S., & Long, D. (2013). Contextual Behavioral Science, Evolution, and Scientific Epistemology. In I. Stewart & B. Roche (Eds.). (2013). *Advances in relational frame theory: Research and application* (pp. 5-26). New Harbinger Publications, Inc.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*(2), 340–347. https://doi.org/10.1037/0033-2909.109.2.340

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

Hughes, S., Barnes-Holmes, D., & Vahey, N. (2012). Holding on to our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. *Journal of Contextual Behavioral Science*, *1*(1-2), 17-38.

Hughes, S., Hussey, I., Corrigan, B., Jolie, K., Murphy, C., & Barnes-Holmes, D. (2016). Faking revisited: Exerting strategic control over performance on the Implicit Relational Assessment Procedure. *European Journal of Social Psychology*, *46*(5), 632-648.

Jost, J. T. (1997). An experimental replication of the depressed-entitlement effect among women. *Psychology of Women Quarterly*, *21*(3), 387–393. https://doi.org/10.1111/j.1471-6402.1997.tb00120.x

Jost, J. T. (2019). A quarter century of system justification theory: Questions, answers, Criticisms, and societal applications. *British Journal of Social Psychology*, *58*(2), 263–314. https://doi.org/10.1111/bjso.12297

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, *33*(1), 1–27. https://doi.org/10.1111/j.2044-8309.1994.tb01008.x

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, *25*(6), 881–919. https://doi.org/10.1111/j.1467-9221.2004.00402.x

Jost, J. T., Pelham, B. W., & Carvallo, M. R. (2002). Non-conscious forms of system justification: Implicit and behavioral preferences for higher status groups. *Journal of Experimental Social Psychology*, *38*(6), 586–602. https://doi.org/10.1016/s0022-1031(02)00505-x

Kilianski, S.E., Rudman, L.A. (1998). Wanting It Both Ways: Do Women Approve of Benevolent Sexism? *Sex Roles* 39**,** 333–352. https://doi.org/10.1023/A:1018814924402

Lalor, I. (2019) *Developing an Empirically Valid Function Acquisition Speed Test for Assessing Attitudes to and Predicting Real-world Behaviour*. [Masters thesis, National University of Ireland Maynooth] https://mural.maynoothuniversity.ie/13647/

Mace, F. C., & Critchfield, T. S. (2010). Translational research in behavior analysis: Historical traditions and imperative for the future. *Journal of the Experimental Analysis of Behavior*, *93*(3), 293–312. https://doi.org/10.1901/jeab.2010.93-293

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An Integrative Review. *Psychological Bulletin*, *109*(2), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

Mandell, C., & Sheen, V. (1994). Equivalence class formation as a function of the pronounceability of the sample stimulus. *Behavioural Processes*, *32*(1), 29-46. https://doi.org/10.1016/0376-6357(94)90025-6

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). Academic Press.

Merwin, R. M., & Wilson, K. G. (2005). Preliminary findings on the effects of self-referring and evaluative stimuli on stimulus equivalence class formation. *The Psychological Record*, *55*(4), 561–575. https://doi.org/10.1007/bf03395527

Moss-Lourenco, P., & Fields, L. (2011). Nodal structure and stimulus relatedness in equivalence classes: Post-class formation preference tests. *Journal of the Experimental Analysis of Behavior*, *95*(3), 343–368. https://doi.org/10.1901/jeab.2011.95-343

Moxon, P. D., Keenan, M., & Hine, L. (1993). Gender-role stereotyping and stimulus equivalence. *The Psychological Record*, *43*(3), 381.

Nevin, J., & Grace, R. (2000). Behavioral momentum and the Law of Effect. *Behavioral And Brain Sciences*, *23*(1), 73-90. https://doi.org/10.1017/s0140525x00002405

Nicholson, E., & Barnes-Holmes, D. (2012). The implicit relational assessment procedure (IRAP) as a measure of Spider Fear. *The Psychological Record, 62*(2), 263–277. https://doi.org/10.1007/bf03395801

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science, 16*(2), 65–69. https://doi.org/10.1111/j.1467-8721.2007.00477.x

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101

Nunnally, J.C. (1978) Psychometric theory. 2nd Edition, McGraw-Hill, New York.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological science, 12*(5), 413-417.

O'Reilly, A (2012) *Developing the Function Acquisition Speed Test: Using a Functional Research Approach to Build a Novel Implicit Test.* [PhD thesis, National University of Ireland Maynooth] https://mural.maynoothuniversity.ie/5396/

O'Reilly, A., Roche, B., & Cartwright, A. (2015). Function over form: A behavioral approach to implicit attitudes. In *Exploring implicit cognition: learning, memory, and social cognitive processes* (pp. 162-182).

O'Reilly, A., Roche, B., Gavin, A. Ruiz M., Ryan A., & Campion G., (2013). A Function Acquisition Speed Test for Equivalence relations (Faster). *The Psychological Record* 63, 707–724. https://doi.org/10.11133/j.tpr.2013.63.4.001

O'Reilly, A., Roche, B., Ruiz, M., Tyndall, I., & Gavin, A. (2012). The Function Acquisition Speed Test (Fast): A behavior analytic implicit test for assessing stimulus relations. *The Psychological Record, 62*(3), 507–528. https://doi.org/10.1007/bf03395817

O' Shea, B., Watson, D. G., & Brown, G. D. (2016). Measuring implicit attitudes: A positive framing bias flaw in the Implicit Relational Assessment Procedure (IRAP). *Psychological assessment*, *28*(2), 158.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT Criterion Studies. *Journal of Personality and Social Psychology*, *105*(2), 171–192. https://doi.org/10.1037/a0032734

Pilgrim, C. (2011). Translational Behavior Analysis and practical benefits. *The Behavior Analyst*, *34*(1), 37–40. https://doi.org/10.1007/bf03392232

Pilgrim, C., & Galizio, M. (1990). Relations between baseline contingencies and equivalence probe performances. *Journal of the Experimental Analysis of Behavior*, *54*(3), 213–224. https://doi.org/10.1901/jeab.1990.54-213

Plaud, J. J. (1995). The formation of stimulus equivalences: Fear-relevant versus fear-irrelevant stimulus classes. *The Psychological Record 45*(2): 207-222. https://doi.org/10.1007/BF03395929

Plaud, J. J. (1997). Behavioral analysis of fear-related responding using a modified matching-to-sample procedure. *Cognitive Behaviour Therapy* 26(4): 157-170. https://doi.org/10.1080/16506079708412485

Plaud, J. J., Gaither, G. A., Franklin, M., Weller, L. A., & Barth, J. (1998). The effects of sexually explicit words on the formation of stimulus equivalence classes. *The Psychological Record*, *48*(1), 63–79. https://doi.org/10.1007/bf03395259

Rabbitt, P., & Rodgers, B. (1977). What does a man do after he makes an error? an analysis of response programming. *Quarterly Journal of Experimental Psychology*, *29*(4), 727–743. https://doi.org/10.1080/14640747708400645

Ridgeway, I., Roche, B., Gavin, A., & Ruiz, M. R. (2010). Establishing and eliminating Implicit Association test effects in the Laboratory: Extending the behavior-analytic model of the IAT. *European Journal of Behavior Analysis*, *11*(2), 133–150. https://doi.org/10.1080/15021149.2010.11434339

Ramos, M. R., Barreto, M., Ellemers, N., Moya, M., Ferreira, L., & Calanchini, J. (2015). Exposure to sexism can decrease implicit gender stereotype bias. *European Journal of Social Psychology*, *46*(4), 455–466. https://doi.org/10.1002/ejsp.2165

Roche, B., & Barnes, D. (1997). A transformation of respondently conditioned stimulus function in accordance with arbitrarily applicable relations. *Journal of the Experimental Analysis of Behavior*, *67*(3), 275–301. https://doi.org/10.1901/jeab.1997.67-275

Roche, B., Barnes-Holmes, Y., Barnes-Holmes, D., Stewart, I., & O'Hora, D. (2002). Relational frame theory: A new paradigm for the analysis of social behavior. *The Behavior Analyst*, *25*(1), 75–91. https://doi.org/10.1007/bf03392046

Roche, B., O'Reilly, A., Gavin, A., Ruiz, M. R., & Arancibia, G. (2012). Using behavior-analytic implicit tests to assess sexual interests among normal and sex-offender populations. *Socioaffective Neuroscience & Psychology*, *2*(1), 17335. https://doi.org/10.3402/snp.v2i0.17335

Roche, B., Ruiz, M. and Hand, K. (2003) An Experimental Analysis of Social Discrimination using Relational Frame Theory. Paper presented at the Annual Conference of the Association for Behavior Analysis, May 23-27

Roche, B., Ruiz, M., O'Riordan, M., & Hand, K. (2005). A relational frame approach to the psychological assessment of sex offenders. In e. Quayle & m. Taylor, Viewing Child Pornography on the Internet: Understanding the Offence, Managing the Offender, and Helping the Victims (pp. 109-125).

Rosenberg, M. J., Rosenthal, R., & Rosnow, R. (1969). The conditions and consequences of evaluation apprehension.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the implicit association test: Dissociating salience from associations. *Journal of Experimental Psychology: General*, *133*(2), 139–165. https://doi.org/10.1037/0096-3445.133.2.139

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. https://doi.org/10.1521/soco.20.4.294.19908

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.

Schlund, M. W., Cataldo, M. F., & Hoehn-Saric, R. (2008). Neural correlates of derived relational responding on tests of stimulus equivalence. *Behavioral and Brain Functions*, *4*(1), 6. https://doi.org/10.1186/1744-9081-4-6

Second European Union minorities and discrimination survey. (2017) European Union Agency For Fundamental Rights. Retrieved October 05, 2021, from

https://fra.europa.eu/en/publication/2017/second-european-union-minorities-and-discrimination-survey-main-results

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*(4), 1241-1260.

Sidman, M. (1960). Tactics of scientific research.

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of speech and Hearing Research*, *14*(1), 5-13.

Sidman, M. (1989). *Coercion and its fallout*. Authors Cooperative.

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Authors Cooperative.

Sivacek, J., & Crano, W. D. (1982). Vested interest as a moderator of attitude–behavior consistency. *Journal of Personality and Social Psychology*, *43*(2), 210–221. https://doi.org/10.1037/0022-3514.43.2.210

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Skinner, B. F. (1974). *About behaviorism*. New York: Knopf

Smeets, P. M., Akpinar, D., Barnes-Holmes, D., & Barnes-Holmes, Y. (2003). Reversal of equivalence relations. *The Psychological Record*, *53*(1), 91-119. ISSN: 0033-2933

Stewart, I., & Roche, B. (2013). Relational Frame Theory: An Overview. In I. Stewart & B. Roche (Eds.). (2013). *Advances in relational frame theory: Research and application* (pp. 51-71). New Harbinger Publications, Inc.

Stieger, S., Göritz, A. S., & Burger, C. (2010). Personalizing the IAT and the SC-IAT: Impact of idiographic stimulus selection in the measurement of implicit anxiety.

Personality and Individual Differences, *48*(8), 940–944. https://doi.org/10.1016/j.paid.2010.02.027

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology*, *68*(2), 199–214. https://doi.org/10.1037/0022-3514.68.2.199

Thorndike, E., 1920. A constant error in psychological ratings. *Journal of Applied Psychology*, 4(1), pp.25-29. https://doi.org/10.1037/h0071663

Tyndall, I. T., Roche, B., & James, J. E. (2004). The relation between stimulus function and equivalence class formation. *Journal of the Experimental Analysis of Behavior*, *81*(3), 257–266. https://doi.org/10.1901/jeab.2004.81-257

Uhlmann, E., Dasgupta, N., Elgueta, A., Greenwald, A. G., & Swanson, J. (2002). Subgroup prejudice based on skin color among Hispanics in the United States and Latin America. *Social Cognition*, *20*(3), 198–226. https://doi.org/10.1521/soco.20.3.198.21104

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of Criterion Effects for the implicit relational assessment procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry*, *48*, 59–65. https://doi.org/10.1016/j.jbtep.2015.01.004

Watt, A., Keenan, M., Barnes, D., & Cairns, E. (1991). Social categorization and stimulus equivalence. *The Psychological Record*, *41*(1), 33–50. https://doi.org/10.1007/bf03395092

World Economic Forum. (2022). *Global Gender Gap Report*. https://www.weforum.org/reports/global-gender-gap-report-2022

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology, 72*(2), 262–274. https://doi.org/10.1037/0022-3514.72.2.262

Zajonc, R., 1968. Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2, Pt.2), pp.1-27. https://doi.org/10.1037/h0025848

Appendix A

The Modern Sexism Scale

Below are a number of statements measuring your attitudes and beliefs toward gender. You may decline to answer a question for any reason, if you so wish. You are reminded once again that all data collected is completely anonymous. **Please read each statement carefully using the scale below to make your choice.**

Note: Questions 4, 5, 6, 7, 8, 10, 12, 13 comprise the Modern Sexism subscale

1	2	3	4	5
Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree

1.Women are generally not as smart as men	1	2	3	4	5
2.I would be equally comfortable having a woman as a boss as a man.	1	2	3	4	5
3.When both parents are employed and their child gets sick at school, the school should call the mother rather than the father.	1	2	3	4	5
4.It is easy to understand why women's groups are still concerned about societal limitations of women's opportunities.	1	2	3	4	5
5.Discrimination against women is no longer a problem in Ireland.	1	2	3	4	5
6.Women often miss out on good jobs due to sexual discrimination.	1	2	3	4	5
7.On average, people in our society treat husbands and wives equally.	1	2	3	4	5
8.It is rare to see women treated in a sexist manner on television.	1	2	3	4	5
9.Women are just as capable of thinking logically as men.	1	2	3	4	5
10.It is easy to understand the anger of women's groups in Ireland.	1	2	3	4	5

11.It is more important to encourage boys than to encourage girls to participate in athletics.	1	2	3	4	5
12.Over the past few years, the government and news media have been showing more concern about the treatment of women than is warranted by women's actual experiences.	1	2	3	4	5
13.Society has reached a point where women and men have equal opportunities for achievement.	1	2	3	4	5

Appendix B

Demographic Questionnaire

Please provide the following information to help us with this research

What is your age in years?

Q. What is your gender?

- female
- male
- non-binary
- other

Q. What is your current country of residence

Q. What is your primary ethnicity?

- Caucasian
- Latino/Hispanic
- Middle Eastern
- African
- Caribbean
- South Asian
- East Asian
- Mixed
- Other

Please record this unique random code as proof of participation (e.g., in case you wish to request that your data be removed from this study at a later date).

Your number is: xxxx

When you have taken a note of this, press the spacebar to continue

Appendix C

Information Sheet (Gender Attitudes Study)

This research is being conducted by Matthew Wall (contact: **matthew.wall.2017@mumail.ie**), a postgraduate student at the Department of Psychology, Maynooth University, under the supervision of Dr. Bryan Roche (contact: **Bryan.T.Roche@nuim.ie** / +353 (1) 708 6026). It is the responsibility of this student to adhere to professional ethical guidelines in their dealings with participants and the collection and handling of data. If you have any concerns about participation you may refuse to participate, or withdraw at any stage.

This study involves examining the effectiveness of a new type of computer-based attitude test, called an implicit test. In this study, the topic of interest is attitudes towards gender in society. The implicit test being used is called the Function Acquisition Speed Test (FAST). It works by seeing how fast you can learn to categorise words and images in different ways. This can sometimes indicate a bias towards categorising words and images in a particular way, and this can suggest a particular attitude.

During the test, words and images will appear on the screen one by one. All you need to do is press one of two computer keys, and let the software teach you how to do it correctly. The words that will appear on the screen are either positive words (e.g. Good) or negative words (e.g. Bad). The images that appear on screen are an assortment of different faces.

As part of the experiment you will be asked to identify your gender, ethnicity, age and country of residence. You will also be asked to complete a questionnaire regarding your attitudes to gender. Finally, you will be presented with the FAST test. These tests will take approximately 10-15 minutes to complete and you may take a short break between tests should you require it.

All data from the study will be confidential, and it is not possible for us to link your identity to the test performance data we record from you. However, you will be provided on screen with a randomly generated four-digit code. You should note this for proof of participation, if for any reason you wish to ask a question of the researchers or ask for your data to be withdrawn and destroyed.

The data gathered will be compiled and, analysed at a group level only and submitted in a postgraduate thesis. This data may also be used as part of analyses for a scientific publication. All data collected will be retained on a University computer in the Department of Psychology for a duration of 10 years as per University regulations. No personally identifying information will be gathered or stored in any form.

At the conclusion of your participation, you will be provided with more information about the purpose of the study, and you will be invited to email the researchers with any further queries you may have. If you do request test scores, we wish to tell you in advance that we can do so, but we cannot interpret these for you, or make any sort of diagnosis or comment on your character.

Participants should be over 18 years of age, and should not suffer from any known condition that will make concentrating or problem solving difficult in any way.

While we will hold no personal data of any kind on participants, it must be recognised that, in some circumstances, confidentiality of research data and records may be overridden by courts in the event of litigation or in the course of investigation by lawful authority. In such circumstances the University will take all reasonable steps within law to ensure that confidentiality is maintained to the greatest possible extent.

If during your participation in this study you feel the information and guidelines that you were given have been neglected or disregarded in any way, or if you are unhappy about the process, please contact the Secretary of the National University of Ireland Maynooth Ethics Committee at **research.ethics@nuim.ie** or +353 (0)1 708 6019. Please be assured that your concerns will be dealt with in a sensitive manner.

Appendix D

Social Media Advertisement (Gender Attitudes Study)

If you are over the age of 18 and a resident of Ireland, you are invited to participate in a psychology experiment concerning attitudes towards gender in society today. The experiment will involve providing us with your age, ethnicity and gender and completing a very brief attitude questionnaire as well as a simple learning task that will allow the researchers to gauge your gender attitudes. It will take 10-15 minutes to complete.

Your participation would be completely anonymous and no personally identifying information will be collected.

This research is being conducted by Matthew Wall (contact: **matthew.wall.2017@mumail.ie**), a postgraduate student at the Department of Psychology, Maynooth University, under the supervision of Dr. Bryan Roche (contact: **Bryan.T.Roche@nuim.ie** / +353 (1) 708 6026).

More information about the nature of the experiment, as well as the consent form you need to complete before participation can be found at the following link **www.millisecond.com/xxxxxxxxxxxx**

Appendix E

Consent Form (Gender Attitudes Study)

This research is being conducted by Matthew Wall, a postgraduate student at the Department of Psychology, Maynooth University. The method proposed for this research project has been approved in principle by the Maynooth University Research Ethics Committee, which means that the Committee does not have concerns about the procedure. It is the responsibility of the researcher to adhere to ethical guidelines in dealing with the participants and the collection and handling of data. If you have any concerns about participation you may refuse to participate or withdraw at any stage. At the beginning of testing, you will be provided with a randomly generated four-digit code. Please take note of this code as if you decide to withdraw your data from analysis it will be impossible to identify it without this information precisely because we hold no personal information about you whatsoever.

In this study we will ask you to provide some demographic information (age, gender, ethnicity and country of residence), followed by a computer based implicit test that takes the form of a learning task, and finally a short questionnaire regarding your opinions on gender. You may take a break between tasks if you so desire.

If you are under the age of 18 or feel uncomfortable with the topic of this research, or for any other reason wish to not participate, you should leave now before any data is collected. If at any point during experimentation you decide you no longer want to participate you may leave and your data will not be utilised. Please self-exclude also if English is not your first language, or if you have vision difficulties that cannot be corrected with spectacles.

All of the data collected in this study will be aggregated and will be included in a Masters thesis report completed by the researcher. The research may also be published in a scientific journal.

Your participation in this study will require approximately 10-15 minutes. If you have any concerns or queries about the study you can contact the researcher Matthew Wall at **matthew.wall.2017@mumail.ie**, or supervisor Dr. Bryan Roche at **Bryan.T.Roche@nuim.ie**. Please take note of these details now. It is the responsibility of this student researcher to adhere to professional ethical guidelines in their dealings with participants and the collection and handling of data. If you have any concerns about participation you may refuse to participate, or withdraw at any stage.

If during your participation in this study you feel the information and guidelines that you were given have been neglected or disregarded in any way, or if you are unhappy about the process, please contact the Secretary of the National University of Ireland Maynooth Ethics

Committee at **research.ethics@nuim.ie** or +353 (0)1 708 6019. Please be assured that your concerns will be dealt with in a sensitive manner.

By checking the box below, you are agreeing that: (1) you have read and understood the Participant Information Sheet on the previous page. (2) you are taking part in this research study voluntarily (without coercion), (3) you are over 18 years of age and do not suffer from any medical condition, which may make participating in a computer based learning task dangerous for you, and (4) that you understand that it is only possible to withdraw data by contacting the researchers at one of the email addresses above and providing the unique four-digit code provided to you on the next page.

Consent and Proceed ☐

Appendix F

Debriefing Information (Gender Attitudes Study)

Thank you for taking the time to participate in this study. The purpose of this experiment was to test the hypothesis that marginalised groups, such as women or Black individuals can display a bias towards their own social group. Of course, this may not be true for all minority or marginalised groups but the effect may be apparent when average bias test scores are examined carefully. This research was also interested in comparing the degree of bias shown by minority or marginalised groups, compared to majority groups or groups holding social power.

The FAST test that you took measures reflexive associations, in this case between faces of different genders and both positive and negative terms. Recall that there were two blocks in the test. You may have found that one block was easier to complete then the other. This might be because of the way in which the various faces and words were indirectly associated in that block by sharing a common keyboard response, such as the Z or the M key. Researchers can examine the speed and accuracy of test takers on each of these blocks. By seeing which one was associated with the fastest responses and the most correct responses, they can work out whether the test taker is more comfortable indirectly associating faces of a given gender with positive or negative words. In this way they can infer the attitude of the test taker. This study sought to examine whether there was any difference in implicit gender bias scores between men and women.

A secondary question under examination was whether there would be any difference between your stated opinion (on the questionnaire you answered) and your score on the FAST implicit test. This question was posed because the FAST is a relatively new test and it will be useful to know if it is measuring similar biases to those people report in questionnaires.

If you later wish to request access to your data, you will need to do so by email. Please provide your randomly generated four-digit code provided to each participant for such requests. The data will be sent with no accompanying text in a separate email, to an email address of your choice, from a Maynooth University email address. Your email and the reply to it will then be deleted immediately from MU servers. There will be no identifying information in the body of that email. As explained earlier, we will not be able to interpret data for you for ethical reasons.

Should you have any questions or concerns about the study you can contact me at **matthew.wall.2017@mumail.ie** or my supervisor for this research Dr. Bryan Roche at **Bryan.T.Roche@nuim.ie** / +353 (1) 708 6026. If during your participation in this study you feel the information and guidelines that you were given have been neglected or disregarded in any way, or if you are unhappy about the process, please contact the Secretary of the National University of Ireland Maynooth Ethics Committee at **research.ethics@nuim.ie** or +353 (0)1 708 6019. Please be assured that your concerns will be dealt with in a sensitive manner.

Appendix G

Discrimination and Diversity Scales

Note that this is the original scoring system of the DDS, in the current study the direction of the scale was reversed when administered. That is to say, a score of 1 indicated "Strongly Disagree" and a score of 5 indicated "Strongly Agree". The scale was not otherwise altered from its original scoring system. The phrase *United States* was replaced with *Ireland* in three instances.

DS

1. Members of ethnic minorities have a tendency to blame Whites too much for problems that are their own doing.

2. Members of ethnic minorities often exaggerate the extent to which they suffer from racial inequality.

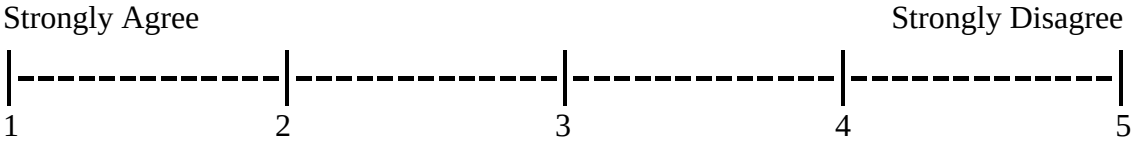
3. Black people often blame the system instead of looking at how they could improve their situation themselves.

4. These days, reverse discrimination against Whites is as much a problem as discrimination against Blacks itself.

5. More and more, Blacks use accusations of racism for their own advantage.

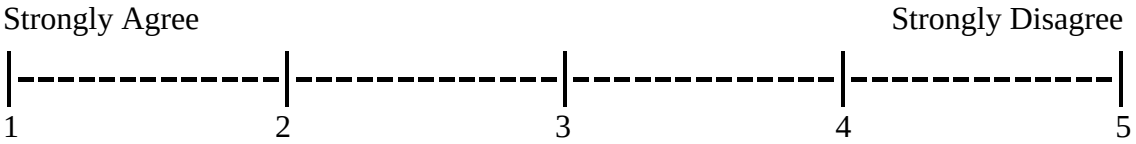
6. Blacks are ultimately responsible for the state of race relations in this country.

7. Discrimination against Blacks is not a problem in Ireland.

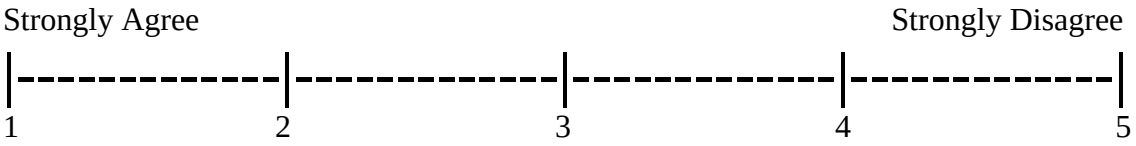


8. A primary reason that ethnic minorities tend to stay in lower paying jobs is that they lack the motivation required for moving up.

9. Many ethnic minorities do not understand how hard one has to work to achieve success.



10. In Ireland people are not judged by their skin colour.



Appendix H

Modified Modern Racism Scale

Please answer *all* of the following six statements and please do so as *honestly* as you possibly can. Your answers are completely confidential and anonymous.

1. It is easy to understand the anger of black people.

Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
-2	-1	0	1	2

2. Blacks are getting too demanding in their push for equal rights.

Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
-2	-1	0	1	2

3. Over the past few years blacks have gotten more economically than they deserve.

Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
-2	-1	0	1	2

4. Over the past few years the government and news media have shown more respect to blacks than they deserve.

Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
-2	-1	0	1	2

5. Blacks should not push themselves where they're not wanted.

Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree

-2	-1	0	1	2

6. Discrimination against blacks is no longer a problem.

Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
-2	-1	0	1	2

Appendix I

Information Sheet (Racial Bias Study)

This research is being conducted by Matthew Wall (contact: **matthew.wall.2017@mumail.ie**), a postgraduate student at the Department of Psychology, Maynooth University, under the supervision of Dr. Bryan Roche (contact: **Bryan.T.Roche@nuim.ie** / +353 (1) 708 6026). It is the responsibility of this student to adhere to professional ethical guidelines in their dealings with participants and the collection and handling of data. If you have any concerns about participation you may refuse to participate, or withdraw at any stage.

This study involves examining the effectiveness of a new type of computer-based attitude test, called an implicit test. In this study, the topic of interest is attitudes towards race in society. The implicit test being used is called the Function Acquisition Speed Test (FAST). It works by seeing how fast you can learn to categorise words and images in different ways. This can sometimes indicate a bias towards categorising words and images in a particular way, and this can suggest a particular attitude.

During the test, words and images will appear on the screen one by one. All you need to do is press one of two computer keys, and let the software teach you how to do it correctly. The words that will appear on the screen are either positive words (e.g. Good) or negative words (e.g. Bad). The images that appear on screen are an assortment of different faces.

As part of the experiment you will be asked to identify your gender, ethnicity, age and country of residence. You will also be asked to complete a questionnaire regarding your attitudes to race. Finally, you will be presented with the FAST test. These tests will take approximately 10-15 minutes to complete and you may take a short break between tests should you require it.

All data from the study will be confidential, and it is not possible for us to link your identity to the test performance data we record from you. However, you will be provided on screen with a randomly generated four-digit code. You should note this for proof of participation, if for any reason you wish to ask a question of the researchers or ask for your data to be withdrawn and destroyed.

The data gathered will be compiled and, analysed at a group level only and submitted in a postgraduate thesis. This data may also be used as part of analyses for a scientific publication. All data collected will be retained on a University computer in the Department of Psychology for a duration of 10 years as per University regulations. No personally identifying information will be gathered or stored in any form.

At the conclusion of your participation, you will be provided with more information about the purpose of the study, and you will be invited to email the researchers with any further queries you may have. If you do request test scores, we wish to tell you in advance that we can do so, but we cannot interpret these for you, or make any sort of diagnosis or comment on your character.

Participants should be over 18 years of age, and should not suffer from any known condition that will make concentrating or problem solving difficult in any way.

While we will hold no personal data of any kind on participants, it must be recognised that, in some circumstances, confidentiality of research data and records may be overridden by courts in the event of litigation or in the course of investigation by lawful authority. In such circumstances the University will take all reasonable steps within law to ensure that confidentiality is maintained to the greatest possible extent.

If during your participation in this study you feel the information and guidelines that you were given have been neglected or disregarded in any way, or if you are unhappy about the process, please contact the Secretary of the National University of Ireland Maynooth Ethics Committee at **research.ethics@nuim.ie** or +353 (0)1 708 6019. Please be assured that your concerns will be dealt with in a sensitive manner.

Appendix J

Consent Form (Racial Bias Study)

This research is being conducted by Matthew Wall, a postgraduate student at the Department of Psychology, Maynooth University. The method proposed for this research project has been approved in principle by the Maynooth University Research Ethics Committee, which means that the Committee does not have concerns about the procedure. It is the responsibility of the researcher to adhere to ethical guidelines in dealing with the participants and the collection and handling of data. If you have any concerns about participation you may refuse to participate or withdraw at any stage. At the beginning of testing, you will be provided with a randomly generated four-digit code. Please take note of this code as if you decide to withdraw your data from analysis it will be impossible to identify it without this information precisely because we hold no personal information about you whatsoever.

In this study we will ask you to provide some demographic information (age, ethnicity, gender and country of residence) followed by a computer based implicit test that takes the form of a learning task, and finally two short questionnaires regarding your opinions on Black and White people. You may take a break in between tasks if you so desire. If you are under the age of 18 or feel uncomfortable with the topic of this research, or for any other reason wish to not participate, you should leave now before any data is collected. If at any point during experimentation you decide you no longer want to participate you may leave and your data will not be utilised. Please self-exclude also if English is not your first language, or if you have vision difficulties that cannot be corrected with spectacles.

All of the data collected in this study will be aggregated and will be included in a Masters thesis report completed by the researcher. The research may also be published in a scientific journal.

Your participation in this study will require approximately 10-15 minutes. If you have any concerns or queries about the study you can contact the researcher Matthew Wall at **matthew.wall.2017@mumail.ie**, or supervisor Dr. Bryan Roche at **Bryan.T.Roche@nuim.ie**. Please take note of these details now. It is the responsibility of this student researcher to adhere to professional ethical guidelines in their dealings with participants and the collection and handling of data. If you have any concerns about participation you may refuse to participate, or withdraw at any stage.

If during your participation in this study you feel the information and guidelines that you were given have been neglected or disregarded in any way, or if you are unhappy about the process, please contact the Secretary of the National University of Ireland Maynooth Ethics Committee at **research.ethics@nuim.ie** or +353 (0)1 708 6019. Please be assured that your concerns will be dealt with in a sensitive manner.

By checking the box below, you are agreeing that: (1) you have read and understood the Participant Information Sheet on the previous page. (2) you are taking part in this research study voluntarily (without coercion), (3) you are over 18 years of age and do not suffer from any medical condition, which may make participating in a computer based learning task dangerous for you, and (4) that you understand that it is only possible to withdraw data by contacting the researchers at one of the email addresses above and providing the unique four-digit code provided to you on the next page.

Consent and Proceed ☐

Appendix K

Social Media Advertisement (Racial Bias Study)

If you are over the age of 18 and a resident of Ireland you are invited to participate in a psychology experiment concerning attitudes towards minority ethnicities. We particularly invite people who do not consider their ethnicity to be White (Caucasian).

The experiment will involve providing us with your age, ethnicity and gender and completing two brief attitude questionnaires as well as a simple learning task that will allow the researchers to gauge your attitudes towards Black and White people. It will take 10-15 minutes to complete.

Your participation would be completely anonymous and no personally identifying information will be collected.

This research is being conducted by Matthew Wall (contact: **matthew.wall.2017@mumail.ie**), a postgraduate student at the Department of Psychology, Maynooth University, under the supervision of Dr. Bryan Roche (contact: **Bryan.T.Roche@nuim.ie** / +353 (1) 708 6026).

More information about the nature of the experiment, as well as the consent form you need to complete before participation can be found at the following link **www.millisecond.com/xxxxxxxxxxxx**

Appendix L

Debriefing Information (Racial Bias Study)

Thank you for taking the time to participate in this study. The purpose of this experiment was to test the hypothesis that marginalised groups, such as women or Black individuals can display a bias towards their own social group. Of course, this may not be true for all minority or marginalised groups but the effect may be apparent when average bias test scores are examined carefully. This research was also interested in comparing the degree of bias shown by minority or marginalised groups, compared to majority groups or groups holding social power.

The FAST test that you took measures reflexive associations, in this case between faces of different ethnicities and both positive and negative terms. Recall that there were two blocks in the test. You may have found that one block was easier to complete then the other. This might be because of the way in which the various faces and words were indirectly associated in that block by sharing a common keyboard response, such as the Z or the M key. Researchers can examine the speed and accuracy of test takers on each of these blocks. By seeing which one was associated with the fastest responses and the most correct responses, they can work out whether the test taker is more comfortable indirectly associating faces of a given ethnicity with positive or negative words. In this way they can infer the attitude of the test taker. This study sought to examine whether there was any difference in implicit racial bias scores between Black and White individuals.

A secondary question under examination was whether there would be any difference between your stated opinion (on the questionnaire you answered) and your score on the FAST implicit test. This question was posed because the FAST is a relatively new test and it will be useful to know if it is measuring similar biases to those people report in questionnaires.

If you later wish to request access to your data, you will need to do so by email. Please provide your randomly generated four-digit code provided to each participant for such requests. The data will be sent with no accompanying text in a separate email, to an email address of your choice, from a Maynooth University email address. Your email and the reply to it will then be deleted immediately from MU servers. There will be no identifying information in the body of that email. As explained earlier, we will not be able to interpret data for you for ethical reasons.

Should you have any questions or concerns about the study you can contact me at **matthew.wall.2017@mumail.ie** or my supervisor for this research Dr. Bryan Roche at **Bryan.T.Roche@nuim.ie** / +353 (1) 708 6026. If during your participation in this study you feel the information and guidelines that you were given have been neglected or disregarded in any way, or if you are unhappy about the process, please contact the Secretary of the National University of Ireland Maynooth Ethics Committee at **research.ethics@nuim.ie** or +353 (0)1 708 6019. Please be assured that your concerns will be dealt with in a sensitive manner.

Appendix M

Gender, Modern Sexism Scores and FAST Data (Gender Attitudes Study)

Note, "C-" refers to the consistent block, "I-" refers to the inconsistent block

CRPM = Correct Responses Per Minute

IRPM = Incorrect Responses Per Minute

MS = Modern Sexism Scale

Subject	Gender	MS Scale	C-CRPM	C-IRPM	I-CRPM	I-IRPM	RFD_Score
1.00	Female	19.00	21.55	5.39	27.70	.57	-10.97
2.00	Female	19.00	19.13	4.78	24.34	1.01	-8.97
3.00	Female	10.00	19.91	4.37	27.67	.56	-11.57
4.00	Female	15.00	24.33	1.01	23.60	2.05	1.77
5.00	Female	22.00	21.49	2.39	21.79	2.97	.28
6.00	Female	23.00	26.18	2.28	29.35	.00	-5.45
7.00	Female	9.00	24.83	1.03	25.41	.52	-1.09
8.00	Female	15.00	23.04	2.56	26.54	.54	-5.52
9.00	Female	20.00	22.14	2.46	24.87	.51	-4.68
10.00	Male	28.00	16.04	9.02	14.01	10.14	3.15
11.00	Female	22.00	21.74	4.77	25.19	3.44	-4.79
12.00	Male	25.00	12.86	10.10	16.55	7.09	-6.70
13.00	Male	9.00	27.19	1.74	27.88	.57	-1.86
14.00	Male	17.00	21.14	2.35	17.66	4.41	5.54
15.00	Male	20.00	23.32	1.49	23.98	1.53	-.61
16.00	Male	17.00	19.98	4.39	20.23	3.85	-.78
17.00	Female	10.00	24.78	2.15	27.95	.00	-5.32
18.00	Female	10.00	22.78	3.11	28.06	.00	-8.39
19.00	Male	9.00	13.69	8.39	17.39	4.91	-7.19
20.00	Female	22.00	20.62	5.16	25.17	1.05	-8.65
21.00	Male	32.00	22.46	1.95	21.89	3.56	2.18
22.00	Female	12.00	21.24	3.46	26.67	2.32	-6.57
23.00	Male	34.00	18.38	6.46	17.12	6.66	1.46
24.00	Male	8.00	24.30	3.31	24.63	3.36	-.29
25.00	Female	16.00	25.56	.52	27.74	.00	-2.71
26.00	Female	16.00	24.84	2.16	26.10	1.67	-1.76
27.00	Female	20.00	25.05	2.18	26.76	1.71	-2.18
28.00	Female	10.00	26.41	1.69	28.49	1.82	-1.95
29.00	Female	13.00	24.28	2.11	26.59	1.70	-2.73
30.00	Female	11.00	22.12	4.86	28.34	.58	-10.50
31.00	Female	14.00	21.47	2.93	26.20	.53	-7.12
32.00	Female	12.00	21.18	5.29	24.57	2.73	-5.96
33.00	Female	12.00	26.46	1.10	25.63	1.64	1.37
34.00	Male	29.00	25.78	1.07	25.40	1.06	.37
35.00	Female	17.00	24.87	.51	26.32	.00	-1.96
36.00	Male	12.00	26.04	2.89	27.73	1.16	-3.43
37.00	Female	27.00	22.40	3.65	25.45	1.06	-5.63
38.00	Female	19.00	22.78	1.45	26.20	1.09	-3.78

39.00	Male	27.00	16.12	7.59	17.71	6.22	-2.95
40.00	Male	17.00	28.00	.57	27.18	1.73	1.98
41.00	Female	18.00	25.28	3.45	28.52	1.19	-5.49
42.00	Female	10.00	22.49	1.96	25.79	1.07	-4.18
43.00	Male	20.00	24.51	2.13	26.51	1.10	-3.02
44.00	Female	10.00	23.96	2.66	25.52	1.63	-2.59
45.00	Female	11.00	26.43	2.30	25.38	2.21	.96
46.00	Female	16.00	23.55	2.62	25.78	1.07	-3.78
47.00	Female	14.00	25.13	1.60	26.73	1.11	-2.09
48.00	Female	17.00	21.00	4.00	23.30	2.59	-3.71
49.00	Female	16.00	26.54	1.11	27.79	1.77	-.58
50.00	Female	19.00	21.49	2.39	23.62	2.05	-2.46
51.00	Male	19.00	19.67	6.21	22.94	3.13	-6.35
52.00	Female	14.00	18.55	4.64	25.96	.00	-12.05
53.00	Female	11.00	22.65	3.69	26.38	1.10	-6.31
54.00	Female	12.00	20.97	4.60	27.26	1.14	-9.75
55.00	Female	19.00	23.42	2.60	27.12	1.73	-4.57
56.00	Female	18.00	27.11	.55	27.14	1.13	.55
57.00	Female	14.00	24.86	2.16	29.42	.00	-6.72
58.00	Female	17.00	16.98	10.41	22.84	5.71	-10.56
59.00	Female	17.00	23.96	1.53	22.86	2.54	2.11
60.00	Male	15.00	26.10	.53	25.67	1.07	.97
61.00	Female	16.00	24.62	4.01	27.14	2.36	-4.17
62.00	Female	12.00	26.33	1.68	27.81	.00	-3.15
63.00	Female	9.00	22.05	3.01	25.34	1.62	-4.68
64.00	Female	14.00	19.71	5.56	25.02	2.18	-8.69
65.00	Female	17.00	21.08	5.27	26.33	1.68	-8.84
66.00	Female	14.00	23.45	3.20	28.58	.58	-7.75
67.00	Female	16.00	16.46	6.40	22.72	2.52	-10.14
68.00	Female	15.00	24.60	4.00	23.57	3.21	.23
69.00	Male	13.00	22.21	1.42	18.44	5.20	7.56
70.00	Female	17.00	8.56	11.83	20.06	5.66	-17.67
71.00	Male	24.00	20.06	6.34	20.81	5.87	-1.21
72.00	Female	18.00	17.87	3.92	19.79	4.34	-1.50
73.00	Male	21.00	18.98	4.17	22.07	3.59	-3.66
74.00	Male	23.00	27.30	.00	23.85	1.52	4.97
75.00	Male	21.00	18.26	6.42	17.75	7.61	1.70
76.00	Female	19.00	21.11	3.44	22.70	1.97	-3.05
77.00	Female	16.00	22.93	2.55	20.26	3.86	3.98
78.00	Male	16.00	14.98	4.73	21.49	2.39	-8.85
79.00	Male	16.00	16.05	10.70	21.31	3.47	-12.49
80.00	Female	18.00	20.37	4.47	10.34	14.27	19.84
81.00	Female	10.00	17.09	5.40	21.78	3.55	-6.54
82.00	Male	31.00	15.58	8.76	17.79	7.62	-3.35
83.00	Female	12.00	8.03	14.28	10.12	16.51	.14
84.00	Male	18.00	17.63	6.20	18.58	6.53	-.62
85.00	Male	19.00	22.41	3.06	23.34	2.59	-1.39
86.00	Male	32.00	23.05	1.47	18.06	5.70	9.22
87.00	Female	20.00	18.85	5.32	21.34	4.68	-3.12

88.00	Female	15.00	16.82	6.54	23.93	2.66	-10.99
89.00	Female	18.00	22.89	1.99	15.79	8.88	13.99
90.00	Male	12.00	23.39	3.19	16.57	5.82	9.46
91.00	Male	19.00	21.44	2.38	20.86	3.97	2.17
92.00	Female	13.00	18.50	5.22	13.43	9.72	9.58
93.00	Male	19.00	15.03	9.21	11.35	13.32	7.79
94.00	Male	22.00	19.75	2.69	21.56	2.40	-2.11
95.00	Female	16.00	13.07	11.13	23.53	2.61	-18.98
96.00	Female	12.00	19.89	6.28	20.99	4.61	-2.77
97.00	Female	18.00	15.65	10.43	15.04	7.75	-2.07
98.00	Female	13.00	18.11	7.04	26.97	1.72	-14.19

Appendix N

Ethnicity, Discrimination and Diversity Scale Scores, Modern Racism Scores and FAST Data (Racial Bias Study)

Note, "C-" refers to the consistent block "I-" refers to the inconsistent block

CRPM = Correct Responses per minute

IRPM = Incorrect Responses per minute

DV = Diversity Scale

DS = Discrimination Scale

MR = Modified Modern Racism Scale

Subject	Gender	Ethnicity	DV	DS	MR	C-CRPM	C-IRPM	I-CRPM	I-IRPM	RFD
1.00	Male	White	14	29	-6	25.68	2.23	20.20	7.10	10.34
2.00	Male	Non White	10	26	-5	26.15	2.91	23.66	3.85	3.43
3.00	Female	Non White	8	21	-7	27.23	.56	26.24	1.09	1.52
4.00	Male	White	6	14	-10	24.03	1.00	23.05	2.00	1.98
5.00	Female	White	7	27	-10	23.57	.98	23.72	1.51	.39
6.00	Female	White	13	31	-6	19.27	5.44	24.41	3.97	-6.60
7.00	Male	White	11	18	-5	15.54	10.36	22.96	3.13	-14.65
8.00	Female	White	8	20	-5	24.45	2.13	26.38	1.10	-2.96
9.00	Male	White	14	29	-5	21.28	4.67	15.78	7.43	8.26
10.00	Female	White	10	17	-8	18.63	3.55	19.46	3.17	-1.21
11.00	Male	White	11	19	-8	26.01	1.66	25.63	2.85	1.57
12.00	Female	White	7	17	-12	17.65	6.20	20.00	4.39	-4.16
13.00	Female	Non White	8	17	-11	18.21	5.14	22.82	3.11	-6.63
14.00	Female	Non White	8	16	-4	8.14	15.81	3.34	13.36	2.35
15.00	Male	White	16	35	-4	26.02	1.08	21.37	2.37	5.94
16.00	Male	Non White	10	25	-7	23.37	1.49	15.77	6.13	12.24
17.00	Male	White	7	19	-8	26.19	.00	16.10	6.90	16.99
18.00	Female	Non White	6	14	-12	22.03	6.21	20.86	6.59	1.54
19.00	Female	Non White	6	15	-12	27.47	.00	26.50	.54	1.51
20.00	Female	Non White	5	11	-12	22.74	3.10	24.67	2.74	-2.29
21.00	Female	Non White	9	15	-10	26.56	.54	20.67	2.82	8.17
22.00	Male	Non White	10	22	-10	18.19	7.07	20.95	3.41	-6.42

23.00	Male	Non White	14	35	-4	20.58	8.00	17.13	7.34	2.79
24.00	Male	Non White	8	10	-12	16.75	5.89	23.28	.97	-11.45
25.00	Female	White	16	45	-9	22.92	2.00	25.86	2.25	-2.68
26.00	Female	White	17	41	-7	24.87	2.16	22.10	3.60	4.20
27.00	Female	White	15	42	-7	22.40	3.64	23.59	7.76	-2.17
28.00	Male	White	13	32	-3	13.10	9.49	15.84	7.46	-4.77
29.00	Male	White	13	33	-9	16.53	5.29	14.89	5.53	1.85
30.00	Female	White	19	47	-9	23.74	2.42	20.60	3.16	3.55
31.00	Male	White	14	38	-7	15.86	4.25	14.22	5.87	5.48
32.00	Female	White	15	48	-11	21.84	3.42	14.16	5.85	6.27
33.00	Female	White	14	46	-12	23.59	2.05	20.22	3.85	5.17
34.00	Female	White	14	40	-8	23.29	2.59	22.31	3.04	3.24
35.00	Male	White	17	47	-9	21.91	3.58	23.37	2.77	-2.14
36.00	Female	White	17	44	-9	26.10	2.27	20.23	5.06	8.66
37.00	Female	White	13	38	-8	14.56	5.81	10.76	7.57	4.37
38.00	Male	White	13	32	-5	24.29	2.11	14.35	10.39	18.22
39.00	Male	White	17	40	-6	25.02	2.18	20.08	4.01	7.17
40.00	Female	Non White	19	50	-11	19.52	4.29	20.75	3.96	-1.56
41.00	Female	Non White	17	47	-12	15.70	8.09	15.99	9.00	-.61
42.00	Male	Non White	17	47	-10	22.33	3.64	16.07	7.76	11.24
43.00	Male	Non White	18	45	-6	22.74	1.86	23.80	2.01	-1.68
44.00	Female	Non White	20	50	-12	14.62	10.59	22.32	1.94	-16.34
45.00	Male	Non White	14	37	-9	20.69	5.17	21.60	4.74	1.34
46.00	Female	Non White	18	35	-6	24.25	2.70	21.24	5.31	5.62
47.00	Male	Non White	18	39	-5	23.91	3.26	19.32	5.45	6.78
48.00	Female	Non White	16	42	-6	23.73	2.06	17.81	5.62	9.48
49.00	Female	Non White	16	42	-10	18.78	4.70	20.57	3.35	3.14
50.00	Male	Non White	17	43	-10	21.16	1.84	20.46	2.27	1.13
51.00	Female	Non White	17	46	-9	21.99	3.00	17.82	6.26	7.44
52.00	Male	Non White	18	47	-12	24.35	2.12	19.11	3.11	6.24