Maynooth University
National University of Ireland Maynooth

Assessing the Function Acquisition Speed Test (FAST) as a novel implicit measure of salient emotional experiences.

Aideen Watters

Maynooth University, Department of Psychology

February 2023

Head of Department: Dr. Michael Cooke

Research Supervisor: Prof. Bryan Roche

Contents

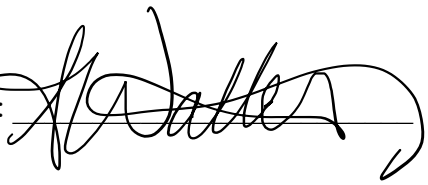
List of Tables

List of Figures

List of Equations

Declaration

I, the undersigned, hereby certify that this material, which I now submit in fulfilment of a M.Sc. degree, has not been previously submitted as an exercise for a degree at this or any other University, and is, unless otherwise stated, entirely my own work.

Signed: 

Aideen Watters

17337283

Date: 28/02/2023

Acknowledgements

Firstly, to my supervisor, Prof. Bryan Roche. Your everlasting patience throughout the writing process, the research journey and indeed my personal development as an academic are deeply appreciated. Your unlimited support has encouraged me to strive for things I would have otherwise ignored (like all three scholarship applications), and many of the achievements I've made over the past year likely wouldn't have happened without it.

Second, the PG room crew- you all have been an immense support and have been the source of calm in my chaotic approach to thesis writing. From the constant cheerleading to the endless chats over cups of coffee, you have all made this process a pleasure from the very beginning. I would be remis if I didn't mention Matthew; no longer an official MU student, but nonetheless hugely influential and encouraging of my journey through postgrad life. Of course, my appreciation also extends toward the whole Psychology Department. In particular, to Dr Maguire, who was very positive about my aspirations to undertake this degree.

Third, Kyle, thank you for the continued support throughout my many theses- I promise I'll be done soon. I'm sure it's not easy listening to me droning on about the differences between ANOVAS and T-tests, or giving out about the stress of it all but you do it anyway. As with every other accomplishment in life you're my biggest supporter. I love having you on my team.

Lastly, my family. The last year has been full of ups and downs but regardless of what's going on I always know I can depend on you for comfort and reassurance. All the women in my family who made me value education as much as I do, in particular Bridie. You've hilariously implored me to earn whatever degree is going to make me the most

money and through this probably isn't the one, I know you're still proud of me. I might still be waiting on a ring but hey, at least I can support myself if it never comes.

Abstract

The purpose of this thesis is to address one of the final outstanding questions from the basic research program into the Function Acquisition Speed Test (FAST), and to contribute to the knowledge on the FAST using the same ground-up approach taken by the developers of the method. This research investigated the utility of the FAST, a novel behaviour-analytic "implicit" test as a measure of stimulus relatedness as a function of stimulus salience. The impact of experimental setting on data quality was also explored. Following a critique of the widely used Implicit Association Test (IAT), the empirical development of the FAST method is outlined. Data for Experiment 1 (n=62) were collected remotely. An evaluative conditioning procedure attempted to establish positive and negative emotional functions for two neutral stimulus classes across three conditions, differentiated by Unconditioned Stimulus (US) salience. Explicit evaluations of the Conditioned Stimulus (CS) were recorded post-conditioning. A FAST, employing the CS and novel evaluative words, was then administered to assess the relatedness of the CSs to the positive and negative evaluative terms. The FAST proved sensitive to the conditioning contingency (i.e., performances reflected the intended evaluative associations), but did not vary as a function of the salience of the US employed during the conditioning phase. Due to unacceptably high attrition levels, Experiment 2 (n=217) replicated Experiment 1 with a larger, remunerated sample of participants. Again, the FAST proved sensitive to conditioning contingencies. An interaction between block fluency scores and CS salience was also observed. Experiment 3 (n=56) aimed to replicate these results with a smaller, supervised and non-renumerated sample. Main

effects were again found, but interaction effects were not. Analysis of attrition rates across samples demonstrated that the paid, online sample in Experiment 2 produced the highest quality data, resulting in the lowest levels of attrition. Challenges, including poor data quality, low sample sizes, and methodological issues that may have compromised stimulus control are discussed in depth. These issues notwithstanding, this study provides in-principle evidence for the FASTs ability to measure the occurrence and intensity of emotional/evaluative learning experiences.

Chapter 1

General Introduction

1.1 Introduction

In 1998, Greenwald and Banaji first introduced their Implicit Association Test (IAT), promising an indirect and discrete measure of "unconscious bias" or "mental associations". This single test has made an enormous impact on the field of social psychology, and psychology more generally. The Greenwald et al. (1998) paper has been cited more than 15,000 times, and the IAT has been used in hundreds of studies attempting to measure implicit attitudes (e.g., ethnic/racial discrimination: Oswald et al., 2013; gender: Hansen et al., 2019; automatic white preference: Dasgupta et al., 2000; self-biases: Nosek et al., 2000; voting intention: Greenwald et al., 2009).

The idea that unconscious cognitive events and mental representations (i.e., implicit biases) could be measured by a simple test intrigued psychologists, to say the least. While behaviour analysts might be initially sceptical of such claims and take issue with the use of mentalistic, non-functionally defined terms, procedures such as the IAT are extremely amenable to conceptual analysis from a behaviour-analytic perspective. In fact, the development of such "implicit measures" directly parallels developments of similar procedures within the behaviour analytic field for different, but related purposes.

In this section , the basic methodology of the IAT is outlined, and its conception as a measure of the strength of "mental associations", from which "unconscious biases" and "attitudes" are inferred, is illustrated. A review of the history of similarly focused behavioural research that led inexorably to parallel developments, resulting in behaviour-analytic IAT-style tasks: the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes et al., 2010) and the Function Acquisition Speed Test (FAST; O'Reilly et al., 2012) will follow. This thesis will focus specifically on the FAST methodology, as this approach is the subject matter of the current research study.

1.2 The Implicit Association Test

The IAT was partly developed in response to the problem of presentation bias, which may be understood as the act of outwardly altering one's beliefs, attitudes or behaviour in order to obtain social approval (see Goffman, 1959; Greenwald & Breckler, 1985). While its developers did not assume that implicit biases were any more authentic than self-presentational biases, they were nevertheless of research interest, especially in cases in which an individual may be likely to misreport their conscious attitudes, including, for example, attempts to conceal racist beliefs (Greenwald & Banaji, 1995).

The IAT is a computer-based test used to assess "mental associations" thought to underlie implicit biases or attitudes. Derived from a connectionist perspective on cognition, the creators of the test conceived an "attitude" as the probability that the activation of a mental concept (e.g., the mental concept of a particular racial group) will lead to the activation of a valence attribute mental concept (e.g., positive valence; Greenwald et al., 2003). The "implicit" aspect of the test derives from the fact that the authors believed the measure captured the activation of these concepts in an unconscious manner. The developers of the test advocated that such implicit attitudes may then have a downstream impact on behaviour in a similarly unconscious manner. From such a perspective, implicit bias is viewed in essence as a latent causal variable; a mental structure in the form of associations that impacts upon behaviour in an unconscious manner (De Houwer , 2019).

In the IAT, participants are presented with stimuli representing two distinct categories of target stimuli, (e.g., flowers and insects), and two categories of attribute stimuli (e.g., "good" and "bad"). Stimuli representing each of these four concepts are presented individually on a computer screen, with each "trial" of the task involving the presentation of one stimulus. The critical task blocks are preceded and intermixed with a series of practice blocks. During the critical blocks of the test, a participant is instructed to press a left or right

keyboard button (e.g. left: 'E' key, right: 'I' key) on each trial. The specific response requirements are outlined in rules presented at the beginning of each block, and reminders for these rules remain present at the top of the computer screen during the block (e.g., "press left for names of plants and good words, press right for names of bugs and bad words"). Each block of the test (i.e., "consistent" and "inconsistent") typically contains 60 trials. In the consistent block, required response configurations are assumed to be consistent with the associations between mental concepts of participants (e.g., flowers and good words share a response, bugs and negative words share another response). In the inconsistent block, the response configurations are assumed to be inconsistent with these associations. Relatively faster responding during one block compared to the other is assumed to reflect associations between mental concepts.

Despite enormous interest (or more likely, because of it), the IAT has also been the subject of considerable conceptual and methodological critique. The following section will consider the most prominent of these critiques, with an eye to outlining those that would be of most interest and relevance to a behaviour-analytic audience.

1.2.2 The IAT and its curious methodology

The IAT is premised upon several assumptions. Some of these assumptions are testable, but some represent *a priori* mentalistic assumptions and explanations of behaviour which are likely unpalatable to the behaviour analyst. For instance, researchers have critiqued the assumption that attitudes are best understood as associations between mental concepts (Hughes et al., 2011). Its developers suggest that the presentation of a stimulus in the task *activates* a mental representation of another related stimulus, and that this activated association affects tasks performance. This associative assumption has taken on an immutable quality within social cognitive attitude research, in spite of the fact that it cannot be easily examined directly (Hughes et al., 2011). Indeed, as others have argued (De Houwer et al.,

2013), this assumption often confounds the interpretation of IAT scores by virtue of failing to adequately separate between observed effects in the measure (e.g., an IAT score) and corresponding explanatory accounts (e.g., association activation).

Another issue of concern is the lack of evidence that variance in IAT scores reflect commensurate differences in the strength of underlying mental associations. Specifically, social cognitive researchers approach this shortcoming from a psychometric perspective in terms of failures to establish sufficiently satisfactory levels of construct validity (Schimmack, 2019). Others have specified particular processes that undermine the reliability of IAT scores as direct measures of mental associations. For example, Calanchini et al. (2014) identified non-attitudinal processes influencing IAT scores. Specifically, they found that detection ability, a mental process associated with discriminating the target stimuli according to the rules, played a large role in IAT performances and therefore scores. That is, the ability to "detect" the correct response required on each task impacted IAT scores, irrespective of attitudes toward the target concept. Nevertheless, researchers continue to work by the assumption that performance on the IAT directly mirrors the organization and strength of mental associations, often based merely on an explicit – implicit measure correlations (e.g., Banse et al., 2001). However, identifying correlations across measures at the group level tells little about underlying mental associations at the individual-level. Another strategy, therefore, is to employ 'known-groups' paradigms, in which differences in test scores across social groups are predicted based on "known" differences in the attitudes of those groups (see Teige-Mocigemba et al., 2010). However, these studies suffer from the same issues as correlational studies: differences at the group-level tell us little about individual-level processes (cf. the mereological fallacy; Bennett & Hacker, 2003).

Several other critiques have been offered by researchers, such as those relating to confounds in stimulus exemplar selection (e.g., De Houwer, 2002), susceptibility to

conscious control (Fiedler et al., 2006), cognitive ability (Klauer et al., 2010), and the confounding impact of stimulus asymmetry (Rothermund & Wentura, 2004) to name but a few.

Most published critiques, and much of the ongoing controversy regarding the IAT, relates to the concerns from those operating at the mental level of analysis. However, behaviour analysts have also weighed in on these issues. Behaviourists propose the IAT may be considered in essence as a learning task, in the sense that corrective feedback is provided during task blocks. Curiously, however, correct responses are never consequated in the task. Incorrect responses *are* consequated via the presentation of a red X and a requirement to adjust one's response. Most behaviour analysts would consider this to be an extremely inefficient way to teach. In addition, it has been well-established that the presentation of aversive events (e.g., a red "X" with likely generalized conditioned punisher functions) during fluent responding, have a detrimental effect on behavioural fluency (e.g., Church & Raymond 1967). Indeed, some supporting evidence for this view was generated by social cognitive researchers before the advent of the IAT in the context of research into the Stroop task (Stroop, 1935). Specifically, Rabbit & Rodgers (1977) found significant decreases in response fluency following negative feedback during Stroop tasks, that were sufficiently unrelated to the task itself that they suggested the omission of response time recordings for all Stroop trials following errors. In effect, it is unknown, the extent to which IAT scores are enhanced by the proliferation of negative feedback during inconsistent trial blocks by the inclusion of response times for trials following negative feedback.

There is yet another aspect of the negative feedback procedure that consciously exaggerates response times during error trials and which has been discussed in several research papers (e.g., Greenwald et al., 2003). More specifically, the IAT does not in fact record response times from the point of stimulus presentation to the emission of a response,

but from the point of stimulus presentation to the production of a *correct response*. The time taken to correct an error response following the presentation of the red X, is roughly 400 milliseconds (see Greenwald et al. 2003). Thus, response latencies on all error trials are enhanced by approximately 400 milliseconds. Greenwald et al. (2003) outlined how this 400ms additional "built-in" time penalty was sufficient to secure reliable and stronger IAT effects and should be retained as part of the standard procedure going forward. In other words, instead of enhancing the IAT effect by improved stimulus control, the effect was enhanced by a scoring algorithm that involved increasing the recorded response times for errors above those which were actually observed. Surprisingly for a behaviour analyst, previous versions of the IAT achieved this arbitrary data inflation manually by recording response times for error responses as the average response time for the whole block + 600ms.

Another aspect of the IAT and its scoring mechanism that may appear curious to the behavioural scientist, is the focus on response time over response accuracy or fluency. This is not uncommon within cognitive psychology and dates to at least the 1880s (McLeod, 1991). In this tradition, response time is used as an index of mental effort and was conceptualized as such perhaps most famously in the Stroop task. This task involves ascertaining response compatibilities by examining response times in a task requiring participants to respond to incompatible features of a stimulus across different blocks of the task (e.g., name the color of a printed word which semantically represents a different color). However, within the behavioural tradition, stimulus control is usually indexed by response accuracy first, and response speed only after accuracy has been established (Binder 1996). Thus, given that the fluencies of the various relational response repertoires being measured in the IAT are unknown except for how they are indexed by the IAT itself, it would be more prudent to use response accuracy as a measure of stimulus compatibility until those very compatibilities have been established in principle in the first instance. Indeed, using accuracy

as the primary metric in IAT-like procedures appears to offer advantages beyond response times; (cf: Cummins & De Houwer, 2022). To use response speed alone as the index of the presence or absence of relations between stimuli, whose degree of relatedness is otherwise unknown, is conceptually questionable at a minimum.

Given that the research program to be discussed originated in an interest in stimulus equivalence (Sidman & Tailby, 1986), it is worth pointing out that the effectiveness of training intended to lead to the emergence of derived stimulus equivalence relations is usually assessed using an accuracy criterion alone (e.g., Fields, et al., 1990), although response times have occasionally been used as an auxiliary metric (see Fields, et al., 2014). While behaviour analysis does not fundamentally object to the use of response times in such research contexts, it is worth considering that response times are typically not normally distributed and that most statistical methods used to analyze them assume that they are (see Heathcote, et al., 1991). The use of inappropriate inferential statistical methods for analyzing response times can lead to distorted effect sizes and difficulties in the interpretation of data (see Whelan, 2008).

Much of the response time truncation, re-coding, and elimination methodology encapsulated in the current IAT scoring algorithm, along with its response correction procedure (later inherited by the IRAP), was intended to normalize data in order to maximize the chances of finding statistically significant test effects (see Greenwald et al., 2003). However, this has been achieved at the expense of interpretability; an IAT D score cannot provide immediate and direct insights into the behaviour that produced it due to these extensive statistical transformations. From a behavioural perspective, a more desirable approach to achieving these goals (while also maintaining interpretability) would be to improve the stimulus control exerted over behaviour within the task itself.

As a particular example, let us consider the issue of "ensuring rapid responses" within the IAT. At present, the task does this via (i) instructions to participants, (ii) the post-hoc removal of response times above 10000ms, (iii) the removal of participants from the data who exhibit greater than 10% responses beyond 10000ms, (iv) the recoding of all response times between 3000ms and 10000ms to 3000ms and (v) the recoding of all response times less than 300ms to 300ms. As an alternative to this level of convolution of the raw data, the task could instead simply instate a limited hold to achieve this (i.e., a response window). This would lead to the shaping of increasingly rapid responding for all participants, or at the very least result in response times being recorded that tally with the response times actually produced by the participant, and simultaneously provide a more simplified statistical approach than that offered by the recommended IAT algorithm. At least ostensibly, this is a superior method of achieving control over the rapidity of responding in the task.

1.3 A Behavioural Approach to Implicit Testing

This section will illustrate that from a behavioural perspective, the IAT is *prima facie* a measure of the relative "strengths" of various stimulus relations (Roche et al., 2005). Until recently, the concept of relation "strength" or stimulus relatedness, was relatively novel in the behaviour-analytic field. Nonetheless, several researchers have attempted to functionally define this concept. For instance, differences in stimulus equivalence (i.e., the spontaneous, untrained relationship that emerger between A1 and C1, through a shared relation with B1) yields following probe tests for emergent relations of different nodal distance can be understood in terms of differences in stimulus relatedness (e.g., Moss-Lourenco & Fields, 2011), as can differences in the probability of the transfer of response functions (Fields et al., 1995; see also Arntzen et al., 2016; Fields, 2015; Fields et al., 2012; Mizael et al., 2016). Probabilities in derived relation yields have also been manipulated using overtraining in baseline conditional discriminations designed to lead to their emergence (e.g., Bortoloti, et

al., 2013). The “strength” of a stimulus relation, or the relatedness of stimuli within a relation, can also be conceptualized in terms of its resistance to change given competing reinforcement contingencies; also referred to by Tyndall et al. (2009) as class “stickiness”. While the emphases of these conceptualizations differ, the outcome they refer to is synonymous (i.e., the probability of functional or equivalence class emergence). It is also worth noting that a longstanding goal of stimulus equivalence researchers is to develop a measure of stimulus relatedness as a function of training procedures (see Bentall et al., 1999; Bortoloti and de Rose, 2009; Doughty et al.2014; Moss-Lourenco & Fields, 2011; Sidman et al., 1985).

Such training procedures are generally employed to establish an equivalence relation amongst various stimuli. Depending on the nature of said training, relations may be formal; for example, orange is equivalent to mandarin as both are spherical in shape and orange in colour; they share common physical characteristics. Relations can also be trained on arbitrary dimensions (i.e., value) for verbally able organisms. Once the individual relations between the individual stimuli are sufficiently trained through continuous pairing and reinforcement procedures, an individual will begin to spontaneously relate stimuli that have not been explicitly paired previously. This process is described as the formation of an equivalence class.

As described above, equivalence classes may be formed on the basis of shared functional or arbitrary qualities. Another way in which an equivalence class may form, however, is on the basis of the functional response the exemplar stimuli of that class elicit. For instance, if a dog, a spider and a snake all elicit a similar fear response (i.e., a functional response) in an individual, those three stimuli may become related to each other vis á vis the shared functional response (e.g., fear), thus creating a class. That is, when stimuli are related though only the response they elicit, it is known as a functional response class.

Given the foregoing IAT example, imagine an individual with a long history of responding to flowers and bug stimulus exemplars as verbally equivalent to positive and negative evaluative terms, respectively. In colloquial terms, the individual *likes* flowers and *dislikes* bugs. Stated technically, the verbal and nonverbal response functions of positive and negative evaluative terms have also been established for flowers and bugs, respectively. Such an individual is likely to demonstrate response differences across IAT blocks which are differently configured as "consistent" and "inconsistent". In other words, the IAT arguably measures the degree to which the establishment of functional response classes is facilitated or impeded by the existence of previously established functional or equivalence classes involving the relevant stimuli. Such a simple description of the IAT process avoids appeal to mentalistic concepts, such as implicit bias, and focuses analysis on the learning history of the participant completing the task, rather than on hypothetical mental events.

While latent variables are not appealed to within behaviour analysis, this does not mean that concepts like implicit bias or attitudes are not amenable to study from a behavioural perspective. For instance, De Houwer (2019) recently argued that implicit bias (or "attitudes") may in fact be reconceptualized as instances of behaviour *qua* behaviour, without much cost to the cognitive perspective. To this end, implicit biases may be defined as "behaviour that is influenced in an implicit manner" (De Houwer 2019 p. 836). Put differently, during implicit testing, the sources of behavioural control are not easily discriminable by the test-taker. Such a conceptualization does not necessarily require a retreat to mentalism, however. De Houwer (2019) suggested that defining implicit bias as behaviour may also offer benefits to cognitive psychologists by allowing for clarity between the to-be-explained phenomenon and the explanatory accounts of that phenomenon. In effect, so long as the nature of the "bias" being analyzed is understood at the behavioural level, the behaviour analyst can utilize and benefit from the same tools used by the social cognitivist.

A seminal study conducted by Watt et al. (1991) was probably the first to provide promise of a behaviour-analytic methodology for assessing socially established verbal relations, which in turn whetted the palette of behaviour-analysts to consider the experimental study of "attitudes" (see Roche et al., 2002). Watt et al. (1991) capitalized upon the stimulus equivalence phenomenon to examine how a sectarian social learning history in Northern Ireland in the 1990s might interfere with the emergence of new, incongruous stimulus relations. Specifically, they attempted to establish two three-member derived equivalence relations using a matching-to-sample (MTS) procedure[1], with the predicted equivalence classes containing a nonsense word, a Catholic name, and a Protestant symbol and a nonsense word. The configuration of the trained classes ran counter to verbal histories of learning in Northern Ireland at the time, wherein Protestant and Catholic names and symbols were usually exclusive rather than equivalent. The study assessed Catholics and Protestant participants from two countries: Northern Ireland and England. Classes were then tested in a standard MTS format, with a Protestant symbol as the sample and Catholic and Protestant names as comparison. According to training, the correct response in this instance was the Catholic name, and selecting this would have indicated the formation of trained stimulus relations. That is, choosing a Catholic name would have indicated that the experimental training had overridden the social learning history of Northern Irish participants. Conversely, selecting the Protestant name would prove that the pre-experimentally established conditional relations overpowered the experimental training. Indeed, equivalence classes emerged

[1] Within an MTS procedure, which can be used to train or test, stimulus relations; the participant is presented with three stimuli (e.g., A1, A2, B2). The so-called sample stimulus appears on its own, in the center (e.g., A1). The two comparison stimuli (e.g., A2, B2) appear beneath the sample. Of these two comparison stimuli, one shares a relation with the sample. This relation can be functional or semantic, depending on the demands of the experiment. The participant must select the comparison stimulus that matches the sample. Accuracy in selecting the related stimulus provides indication as to the participant's familiarity with the overall class and its individual exemplars. Where the procedure is used in a training format, feedback is generally offered, sometimes intermittently, whereas within a test format, feedback is not routinely offered. MTS procedures may be used to train and test for reflexive, symmetrical and transitive relations. They may also be employed to train/test equivalence classes by assessing the participant's ability to determine what stimuli *does not* belong to the class, in addition to determining what stimuli *do* belong in a class.

reliably only for the English participants, who were not socialized within the sectarian culture of the late 1980s in Northern Ireland. Researchers interested in stimulus equivalence were enthusiastic about this finding, and spoke of it as providing the foundation for a discreet and perhaps more reliable behaviour analytic test of verbal histories of learning than direct questioning (e.g., Kohlenberg et al., 1993; Leslie et al., 1993; Merwin & Wilson 2005).

Grey and Barnes (1996) explicitly attempted to provide a definition of the concept of "attitude" from a behavioural perspective and used the stimulus equivalence paradigm as the first port of call for assessing attitudes defined in their terms. They drew upon the process of transfer of function (Barnes & Keenan, 1993) in describing how words within verbal classes acquire affective functions that produce responses that might parallel an attitudinal response of preference or disfavor. For example, if one member of a particular ethnic group is associated directly with aversive stimuli, or directly trained relations are established in language between a small number of exemplars of that class and aversive stimuli (e.g., "Catholics are lazy"), it would be expected that other members of the verbal class might acquire some of the response functions of that aversive stimulus. An attitude, therefore, might be conceived as a generalized affective response to a verbal class of stimuli (i.e., an equivalence class).

Grey and Barnes (1996) tested this idea in an experiment designed to establish three three-member equivalence classes using an MTS procedure (i.e., A1-B1-C1, A2-B2-C2, A3-B3-C3) where all stimuli were nonsense syllables. The movie contents of video cassette tapes, labelled with A1 (sexually themed) or A2 (religiously themed), were shown to participants. Given the prevailing religious views at the time this study was conducted, and the sexual modesty these views promoted, evaluations toward religiosity and sexuality were expected to be positive and negative, respectively. After watching the A1 and A2 videotapes, participants were asked to categorize four more cassette tapes (labelled B1, C1, B2, C2) as

either ‘good’ or ‘bad’, without watching the content. In line with the transfer of functions effect, the video tapes were categorized in accordance with the relevant stimulus equivalence classes to which the original A1 and A2 video tapes belonged. In effect, the researchers had provided a primitive model of “attitudes” in terms of derived generalized evaluative responses. Importantly, the researchers also showed that apparent attitudes change as a function of the context in which the relevant stimuli are presented. Specifically, they found that when a sexual stimulus was presented alongside a ‘worse’ violently sexual stimulus (a video tape containing offensive sexual activity), the former became more acceptable and was sometimes categorized as good, in comparison to the novel stimulus. This finding provided some nuance to the embryonic behavioural approach to attitudes and aligns with contemporary views in cognitive psychology that attitudes must always be understood contextually (e.g., Castelli & Tomelleri, 2008; Jost, 2019).

In another study, Roche et al., (1997) examined resistance to change in stimulus relations established through different means prior to efforts to establish incompatible stimulus relations. The researchers established sexual functions for nonsense word stimuli A1 and C1 and non-sexual functions for A2 -C2 by pairing their brief presentation on a screen with sexual and non-sexual film clips, as appropriate. The establishment of the functional response classes A1-C1 and A2-C2, following the respondent conditioning procedure, was then tested with a simple matching test. Subsequently, the researchers attempted to reorganize the functional A1-C1/A2-C2 stimulus classes by exposing participants to a stimulus equivalence training procedure designed to produce the equivalence classes A1-B1-C2 and A2-B2-C1. A subsequent equivalence test was then administered to measure whether new classes had formed in accordance with the reorganized classes, or resistance to change had indeed prevented these new, trained classes from forming. However, performances on the second equivalence test corresponded with the initial respondent

conditioning, showing resistance to change towards current training and testing contingencies. Importantly, however, for participants that did not pass the matching test following the initial respondent conditioning, the laboratory programmed equivalence relations emerged more easily during the second round of testing.

In an often-overlooked study, Plaud (1995) examined how aversive stimulus functions shared by members of a class might interfere with the formation of arbitrary stimulus equivalence relations consisting of subsets of that class. A within-subjects approach was employed, so that each participant's performance in training and testing designed to lead to the formation of two stimulus equivalence classes, both consisting of images of snakes, was compared to their performance on an identical task involving flower images. Participants also filled out a fear of snakes questionnaire. Following training, a probe test was administered to test for the formation of equivalence classes. The test block contained 24 trials, and was cycled until full accuracy was obtained. Results showed that a higher reported fear of snakes was associated with requiring more repetitions of test blocks to reach criterion for equivalence class formation in the snake condition compared to the flower condition. It appeared reasonable to conclude, therefore, that the fear functions of the snake stimuli employed in the equivalence training procedure was the source of the delayed emergence of equivalence. However, other researchers suggested an alternative explanation.

Tyndall and colleagues (2004) assessed this "Plaud effect" more closely, suspecting that the effect was not due to the aversiveness of the stimuli *per se*, but rather to their shared functions and the relatedness of stimuli within the class (i.e., class "stickiness"). In their study, two functional classes of stimuli were established consisting of six S+ stimuli (responding towards was reinforced) and six S- stimuli (responding away from was reinforced). Two three-member stimulus classes were then trained using an MTS procedure. One of four S+/S- stimulus combinations were trained across each of five conditions (S+

only, S- only, S+/S- one approach and one avoid class, S+/S- functions mixed within class, and a no function condition). It was found that the formation of two 3-member distinct stimulus classes using 6 S+ stimuli (i.e., stimuli with same functions) required the most training trials to reach criterion. In both training and testing phases, the criterion was a minimum of 75% accuracy on each trial type (four trial types presented four times), in addition to maintaining successive correct responding on the final 12 trials of the block. Criteria was met quickest, indicating the quickest class formation, when stimulus equivalence classes corresponded with distinct functional response classes. These findings helped to identify features of learning contexts which impacted upon the acceleration and inhibition of stimulus equivalence class emergence. However, the manipulations across conditions in this study also inadvertently produced a methodology highly reminiscent of a procedure none other than the IAT. Put simply, a first "behavioural IAT" could have consisted of comparing the rate of acquisition of two different stimulus equivalence classes containing real-world stimuli. It would not have required prior training with arbitrary stimuli and yet would still have allowed researchers to identify the configuration of socially established stimulus relations.

In a pivotal experiment, which offered a critical process-level analysis of class formation and change, Hall et al., (2003) established laboratory-controlled stimulus relations involving shapes and colours. On a computer screen, a colour stimulus directly followed the presentation of a shape stimulus (i.e., shape A would be followed by red and shape B by green). The next stage involved establishing a directional response to the shape stimuli from stage 1. That is, when shape A was presented, a left positional keyboard press was reinforced. A right positional response was reinforced when shape B was presented. The final test stage of the experiment required participants to respond positionally to the colours from stage 1. The sample was split into two groups: consistent and inconsistent. That is, for the consistent

group, contingencies for correct responding were consistent with training, while for the inconsistent group they were not. A higher percentage of correct responses was recorded for the consistent group. The inconsistent group responded at chance levels. Indeed, Hall et al. (2003) explicitly acknowledged that the effects seen in their study likely paralleled those observed in the IAT.

Roche et al. (2005) suggested that IAT effects could be understood in terms of differences in fluency of responding to different verbal stimulus class configurations. Roche et al. specifically focused on the rate of *acquisition* of fluent responding to these different configurations. A lack of fluency in the acquisition of a specific configuration of response classes might be indicative of a previously established high rate of fluency in responding to the relevant stimuli according to the opposite pattern. Because such flexibility is established within a social context (i.e., the extent to which words can have multiple meanings and be categorized in different ways), Roche et al. concluded that IAT effects could be understood in terms of stimulus class configuration (in)compatibilities. The authors provided preliminary data to support this position, but this model was more rigorously tested by Gavin et al. (2008). Those researchers administered a training procedure designed to generate two 3-member equivalence relations using nonsense words as stimuli (i.e., A1-B1-C1, A2-B2-C2). A-B and B-C relations were directly trained while derived A-C relations were subsequently probed for in an MTS equivalence test. The idea was to administer a bare-bones IAT following such training to assess whether it would be sensitive to the trained stimulus class configurations.

Modifications to the IAT were reflective of several of the concerns outlined earlier. That is, corrective feedback followed all (not just incorrect) responses. Response windows were limited to 3000ms and missed responses (i.e., over 3000ms) were classified as incorrect. The rationale here was that the presence of the response window led to more errors under

whichever set of contingencies such errors were in principle more likely (cf. Bolsinova et al., 2016). In other words, the idea was to bring error rates under stimulus control directly within the procedure, eliminating the need for post-hoc data manipulations. Accordingly, response time was recorded as the time from trial onset until first emitted response. However, the usual IAT on-screen instructions describing the reinforcement contingencies for each block were present during each trial. The primary test score was calculated in terms of a difference in percentage response accuracy across the two test blocks, rather than in terms of an IAT D score.

In the consistent block of the modified IAT, a common positional response upon the presentation of A1 and C1, and a different common positional response upon the presentation of A2 and C2 stimuli, was reinforced. In the inconsistent block, A1 and C2 required a common response whereas A2 and C1 required an alternative common response for reinforcement. In effect, the functional response classes established in the inconsistent block of the modified IAT were incompatible with the equivalence relations established in phase 1. The test proved sensitive to the training history insofar as higher accuracy was recorded on the consistent block. This provided the first evidence that an explanation of IAT effects in terms of stimulus relation compatibilities was sufficient.

Later, Ridgeway et al. (2010) replicated this general effect. The authors exposed participants to MTS training designed to reorganize the previously established equivalence classes. These participants were subsequently re-exposed to the modified IAT. The performances on this second IAT reflected the modified equivalence relations based on response accuracy, but curiously not based on response times (although this was not highlighted in the paper). In other words, response accuracies proved to be a more sensitive measure of contingency change than response latency (interestingly, this is consistent with recent findings in other measures; cf. Cummins & De Houwer, 2022). Several other studies

then followed, employing a modified IAT to assess stimulus relations that had been established in the natural environment (e.g., sexual stimulus classes; see Gavin et al., 2012, and child-sex relations in the general population vs. child sex offenders; Roche et al. 2012).

Further modifications to the IAT procedure, involving a more behaviour analytically justifiable scoring metric and the removal of unnecessary or empirically unjustifiable methodological features of the IAT (e.g., persistent on-screen instructions, replaced by a response shaping procedure alone), seemed to eventually justify a new name for this procedure that reflected its behaviour-analytic orientation. The name chosen directly described the process that appeared to underlie the basic effect, differences in the speed of acquisition of functional response classes under different reinforcement contingencies. Thus, a new test format including additional features outlined below, was named the *Function Acquisition Speed Test* (FAST; as a convenient acronym and a nod to the speed of administration).

1.3.1 The Function Acquisition Speed Test: FAST

The novel FAST method was the natural product of increasing modifications to and conceptual analyses of the IAT methodology. One rather different feature, however, was an emphasis on a single target test format, which had been previously considered, but apparently abandoned in the IAT literature (e.g., Karpinski & Steinman, 2006). That is, the first FAST study (O'Reilly et al., 2012), explored the utility of assessing the speed of acquisition of functional response classes, for the purpose of indexing the strength of relations within a single class only. Specifically, after establishing two simple zero-node, two-member arbitrary relations involving nonsense words as stimuli, only one of these classes was targeted for indexing in terms of stimulus relatedness. The test involved instating reinforcement contingencies in the consistent block that required common responses to both members of a single class, and a second common response to two novel stimuli *not involved in prior*

training. The inconsistent block involved establishing a common response to one member of an established class and another novel stimulus, and a second positional response for the other member of the established class and yet a further novel nonsense stimulus. The idea was that such a procedure might provide a 'pure' index of individual relation strength, *not relative to* the strength of relations already established among members of a second stimulus class. The procedure was generally effective, successfully generating differences in class acquisition rates (measured as the number of trials required to produce ten consecutive correct responses) across the two test blocks. A further study (O'Reilly et al., 2013), extended the effect to assess the strength of relation amongst stimuli within 1-node derived relations. In this, and all FAST variations going forward, the order of block types was randomized rather than counterbalanced.

Although the idea of an absolute, single-target test was initially the goal, in-house research quickly indicated that FAST effects were generally stronger when two classes were being assessed simultaneously. It was reasoned that, using a relativistic (i.e., double target) procedure rather than an absolute one allowed the functional response classes being established to be accelerated by the already existing behavioural momentum (Nevin & Grace, 2000) established by relating two separate relations simultaneously. A similar conclusion was reached within the IAT literature but for different reasons (e.g., Robinson et al., 2005). Similarly, during an inconsistent block, both functional response classes would be incompatible with two established classes, rather than just one. Thus, a relativistic approach was adopted in the FAST going forward. In hindsight, an "absolute" measure of relatedness is inherently at odds with the contextualistic perspective of stimulus relations research, as others also concluded (see for example Hussey et al., 2016).

In addition to the methodological changes to the IAT which eventually led to the development of the FAST, changes to the scoring method were also made. That is, rather than

use raw response accuracy differences across test blocks as the metric of stimulus relatedness, the FAST drew upon a well-established method within the stimulus equivalence literature involving fluency criteria (i.e., satisfying a consecutive correct response criterion). Because the FAST was conceived as an assessment of relative speeds in the acquisition of functional response classes, it seemed appropriate to have a metric of learning speed that aligned with what was already used in the field, rather than a single-point datum extracted from summarized data based on steady state performances (i.e., as in the IAT and IRAP).

The introduction of the trials-to-criterion method was accompanied by the introduction of two single short baseline blocks (as opposed to practice blocks), one before and one after the two key blocks, and both involving different, arbitrarily chosen nonsense words. The rationale was that these would provide a baseline functional response class acquisition rate for that individual participant. A mean acquisition rate for these baseline blocks could be used to moderate the acquisition rate differential across the two key blocks. In other words, it would facilitate idiographic style standardization that would correct raw response rate differentials by the baseline rate of acquisition. To reflect these changes, a novel, fluency-based scoring metric that combined speed and accuracy, called the Strength of Relation (SoR) index was introduced (O'Reilly et al., 2012). The first iteration involved dividing the trials to criterion differential across blocks by the mean trial requirement on the baseline blocks for each participant. In the second iteration of the index (O'Reilly et al., 2013), the denominator was the natural logarithm of the mean baseline block trial requirement. One study conducted using the latter method (Cummins et al., 2019) used the FAST to measure the impact of behaviour-change focused health education interventions. Participants were health workers assigned to Positive or Negative messages regarding the use of condoms as disease prophylactics, or to a control (no message) condition. All participants then completed a FAST designed to assess relations between condoms and positive and

negative evaluative stimuli. Results showed that the FAST was sensitive to the content of these brief messages. That is, the performances of participants in the condom-positive message condition indicated stronger relations between condoms and positive evaluative terms relative to negative. This pattern was reversed for the negative message group. These results supported the idea that the FAST method was sensitive to verbal relations organized in a brief and naturalistic intervention.

Despite promising results for the native FAST, there were two shortcomings with this SoR scoring metric and associated baseline blocks. Firstly, after dozens of in-house experiments, it was concluded that baseline blocks showed the slowest overall acquisition rates; they were not generally slower than consistent and faster than inconsistent blocks, as initially anticipated. Having replicated this effect in-house with several different stimulus sets, it appeared that the novelty of the stimuli alone was the source of the slow acquisition rates during baseline blocks. Indeed, previous studies had found that the level of familiarity of stimuli (Holth & Arntzen, 1998), as well as the presence of salient emotive or conative stimulus functions for stimuli (Arntzen et al., 2018) was associated with an accelerated rate of stimulus class formation and reorganization. Thus, the use of baseline training blocks was abandoned.

Secondly, the trials-to-criterion component of the SoR index was problematic in its crudeness. That is, a single error on the 10th trial following a run of nine correct responses required the participant to be exposed to at least another 10 trials to satisfy the acquisition criterion. This caused enormous variations in response requirement criteria across blocks and across participants. In other words, the measure was inherently noisy. Thus, a finer metric was conceived in which regression lines were plotted for each block, creating learning curves on a cumulative record, and the difference in the slopes of these lines served as an index of the pre-existing relatedness of stimuli. For this purpose, the number of trials in a block

needed to be fixed, but such a method allowed for a more sensitive analysis of moment-to-moment change and a better functional understanding of the dynamics of the test performance. This would also allow for analyses of the trial-by-trial task performance dynamics. For example, in one study examining the strength of pre-existing verbal relations characteristic of gender stereotypes (Cartwright et al., 2016), in which a sample cumulative record was presented for readers, learning rates typically continue to differentiate as trials progressed through each of the blocks. The dynamics of the performance displayed in the moment-to-moment data corresponded with that of dozens of in-house experiments that showed that learning rates do not differentiate well across blocks within the first ten trials or so, and differentiation in learning trajectories across blocks appears to begin to minimize after 50 trials or so. Thus, while that research is unpublished, a block length of 50 trials was hit upon and appears to have served well in the interim.

Practice blocks were also considered and tested in dozens of in-house studies, but they made little difference to the outcome of the FAST. Importantly, however, the reader should understand that the FAST is conceived as an acquisition rate test, and so providing practice might function as a double-edged sword. That is, practice will serve the purpose of creating a steady state behaviour, as is achieved in the IAT and IRAP (see below) before response speed or fluency differences are assessed across the critical test blocks. However, within the behavioural tradition, behavioural variability is our very subject matter (cf. Sidman, 1960, Skinner, 1976). Therefore, if the contingency shifts across the two blocks are indeed the source of differences in performance, then this should be visible during acquisition itself, albeit with some noise. In other words, in both the IAT and IRAP, the very phenomenon of interest to behaviour analysts (i.e., behaviour *qua* behaviour) is being obfuscated through repeated practice before behavioural metrics are taken. Indeed, in both measures criteria are applied during practice to screen and eliminate participants who do not show such steady

state behaviour (Barnes-Holmes et al., 2010; Hussey et al., 2015). Thus, practice obscures the dynamics of the behavioural performance which should be of interest, even if it does achieve the purpose of eliminating a degree of noise in the data. The reader is reminded, however, that there is a balance to be struck between limiting one source of noise in the task which is not of interest (i.e., random variance) while also capturing another source of noise which *is* of interest (i.e., systematic variance). Striking this balance remains an issue, but contemporary FAST studies generally omit practice blocks.

While the behavioural model of the IAT was being developed, an unrelated research program in a different laboratory involved developing an alternative implicit measure based on a behaviour-analytic approach. Specifically, Barnes-Holmes et al. (2006) proposed what they called the Implicit Relational Association Test (IRAP), a measure that adopted a functional approach to the assessment of stimulus relations as a proxy for assessing implicit attitudes (although see Barnes-Holmes & Harte, 2022, for an arguably revisionist account of the initial impetus for the test). The test was topographically as close to the IAT as was possible, including the use of the same keyboard keys as operanda, and an almost identical scoring algorithm, instructions and stimulus presentation parameters. It was patently part of an effort to provide a behavioural alternative to the IAT given the title of the test, the titles of numerous subsequent papers and despite recent obfuscation on this matter (cf. Barnes-Holmes & Harte, 2022).

1.3.2 The Implicit Relational Association Procedure

The IRAP was developed originally as a measure of implicit beliefs and attitudes, rooted in the functional approach (Hughes et al., 2012). The test was designed specifically to provide nuanced information about the relations between stimuli, as opposed to a simple litmus test of stimulus relatedness as offered by the IAT. The rationale for the IRAP is also rooted in the work of Watt et al (1991). Like the FAST, it emerged from at the behavioural

analysis of derived stimulus relations and the realization that fluencies of relational responding might serve as a convenient proxy for attitudes, conceptualized within the field at the time in terms of networks of stimulus relations. However, in the case of the IRAP there was a greater interest in relations other than stimulus equivalence (e.g., opposition, comparison) and researchers hoped to develop a test that did not provide merely a relative measure of stimulus relatedness or relational bias. For these reasons, the test format first presented in 2006 was more heavily aligned with Relational Frame Theory (Hayes et a., 2001) paradigm.

A detailed analysis of RFT is beyond the scope of this thesis, but briefly, RFT may be described as a behaviour analytic account of language and cognition, which explains how relations between stimuli alter responses to those stimuli through the process of the transformation of stimulus functions in accordance with the nature of the relation in question. Thus, from a RFT perspective, stimulus relations are better conceived of as frames that can take on numerous forms, other than equivalence (e.g., sameness, opposition, distinction, hierarchy). Responding to stimuli in terms of these relations as well as in terms of equivalence relations, is likely established within the verbal community. Over time, complex relational networks of stimuli related to each other under various forms of contextual control become established. The IRAP approaches the assessment of attitudes in terms of an assessment of relations among stimuli in a complex social established relational network involving multiple stimulus relation types. For this reason, the test involves four blocks of tasks rather than just two. The additional blocks are used to assess the unrelatedness of stimuli along a particular dimension as well as the relatedness of stimuli along a dimension (see Hussey et al., 2015).

To illustrate the IRAP procedure, consider the following procedure employed to assess Irish attitudes toward city and rural living (Barnes-Holmes et al., 2009). The IRAP

employs four trial types within each of two task blocks. There are also practice blocks presented and the critical task blocks are repeated six times, although these methodological features are not relevant in the current context. The key aspect of the test, however, is that within each test block participants are required to respond in a particular manner to all four task types. Specifically, in one block they will be instructed to respond as if city stimuli are positive and rural exemplars as if they are negative. In the second block, they will be instructed to respond in the orthogonal manner. In the Barnes-Holmes et al. (2009) study, tasks were defined as assessing the following relational compatibilities; Dublin/Positive, Dublin/Negative, Country/positive and Country/Negative. More specifically, during each trial the category label (e.g., Dublin) was displayed at the top of the screen throughout the entire block. Beneath it, an evaluative concept (i.e., 'good') appeared. On-screen instructions directed the participant to press 'd' for similar (i.e., indicating that the category and evaluative concept were functionally similar) and 'k' for opposite (i.e., indicating the presented concepts were functionally opposite). Importantly, block instructions guided the responses of participants who were required to respond entirely on the basis of those instructions (i.e., either that Dublin is positive and rural is negative, or that Dublin is negative and rural is positive). These two different instructional contingencies define the consistency or inconsistency of the particular block with the social history of the participant. This more detailed and nuanced procedure allows researchers to calculate not only the relative relatedness of a target stimulus with positive rather than negative evaluative stimuli, but also the relative relatedness between the target stimulus and both positive and negative evaluative stimuli in isolation. In other words, the relatedness of the concept of city life to positive evaluative terms can be compared to the relationship that stimulus bears to negative evaluative terms providing a measure of the evaluation of city life irrespective of the

evaluation of rural life. It is in this sense that the measure has been claimed to be non-relative (Hughes et al., 2017).

As for the IAT, responding is assumed to be quicker on 'consistent' blocks, rather than on inconsistent blocks, and several papers have been written theorizing on the reasons for this performance beginning with the Relation Elaboration and Coherence model (REC; Hughes & Barnes-Holmes, 2013) and more recently the Multi-Dimensional Multi-Level framework (MDML, Barnes-Holmes et al., (2020a), formerly Hyper Dimensional Multi Level: HDML, Barnes-Holmes et al., (2020b).

There have been several commentaries on behaviour-analytic concerns about the IRAP procedure within the FAST literature, notwithstanding it's theoretical coherence under the rubric of RFT (see Ridgeway et al., 2010; O'Reilly, 2012; 2013; Gavin et al., 2012; Cartwright et al., 2016; Cummins et al 2018). These have not so far, but might include concerns regarding the top-down theoretical nature of the more recent and arguably practically untestable MMDL and HDML models. These have been devised based on post hoc theoretical revisions of data to date rather than from the ground up in purposeful prospective research. This does not completely nullify their conceptual contribution as guiding paradigms for research, but they run the risk of putting the cart before the horse in terms of directing research questions based on the theoretical model. The reversing of the usual direction within behaviour analysis of moving from data to theory rather than vice versa, jeopardizes the very behavioural approach itself and without sufficient justification. More specific methodological critiques have been provided within the FAST literature (e.g., Ridgeway et al., 2010), regarding the absence of limited hold response windows within both the IRAP and the IAT. Both rely on instructions and encouragement to ensure rapid responding and engage in post hoc data elimination to ensure that all responses are within a 3000-millisecond response window. Given the ease with which a response window could be

enforced, for instance with the instatement of contingencies rather than through instructions and post hoc data manipulation, such a strategy would appear very fitting for a behavioural test at least if not for the IAT. In addition, while the IAT response algorithm (Greenwald et al., 2003), eliminates overly rapid responding in order to normalize response time data for inferential analysis , the IRAP goes a step further in requiring a minimum average response latency of 2000ms, and removes participants who fail to meet this criterion (see Barnes-Holmes et al., 2010 a). The development of these and other procedures was not outlined in ongoing bottom up research assessed in public debate, but tends to appear in whole cloth following unpublished in-house research. For instance, the IRAP also requires a minimum response accuracy rate of 80%- a figure that has been arrived at without sufficient research and scrutiny or comparison of different methods across different studies. Given this, it is difficult to know to what extent the post-hoc scoring techniques are contriving to simulate increased stimulus control and thereby circumventing the need for a greater understanding of core process. Indeed, it seems that these researchers have prioritized the need for more ingenious methods of generating statistically significant test effects in the absence of those improved procedures. Put simply, a postdoc data analytic technique would not normally be included as an experimental methodology in the field of the experimental analysis of behaviour. Inextricably linking laboratory methods to scoring techniques is in essence the practice of psychometrics rather than experimental psychology.

The IRAP is also an entirely response-time based measure, rather than a true fluency-based measure and d-score algorithm drawn from the IAT scoring algorithm is employed to index the test effect based on response times alone. This is curious for a behavioural measure because it has not yet been satisfactorily established that response time alone is an index of relational fluency. Such a response-time dependent measure entirely eliminates the concept of behavioural probability, which is surely a first for behavioural measure in our field.

Finally, the IRAPs inheritance of the curious response correction procedure employed by the IAT, involves the punishment of incorrect responses only, in the absence of the reinforcement of correct responses. Within the social cognitive paradigm, feedback is conceived to work entirely on an instructional basis, rather than as a real tangible external contingency over behaviour (see De Houwer, 2009). However, it is curious for a behaviour analytic test to essentially adopt the position that reinforcement is not required for learning to occur, and is employed with the sole intention of teaching participants how to respond on each trial. It is therefore in essence not a learning task by definition despite being presented as such. Much has also been written within the FAST literature about the curious inclusion of the feedback presentation time (400ms, see Roddy et al., 2011) following incorrect responses in the response times for that trial, a procedure which serves only to considerably inflate the average response time on blocks in which more frequent incorrect responses are observed, thereby conflating fluency with speed (see Ridgeway, et al., 2010– Gavin et al 2012.)

Despite being offered as a behaviour analytic alternative to the IAT, the IRAP has arguably not in fact dealt with critiques of the IAT in a ground up research program that might have been expected of a behavioural analytic research agenda. Instead, the IRAP consists of a behaviour analytic interpretation of the IAT, presented in whole cloth with modifications over the years limited almost entirely to changes in instructions and scoring methods. What was sorely needed, however, was a ground up research program in which the test was developed in a public way and introduced across several successive, related studies. Such basic research involving the IRAP has never been produced using laboratory-controlled stimuli and the manipulation of all features of contingencies controlling performances in order to provide a better understanding of the phenomena being assessed by the test. In addition, criticism has been levied about the statistical power of such studies that involve multi-level analysis of variance with relatively small sample sizes conducted within a small

pool of researchers with insufficient replication across laboratories (see McLoughlin & Roche, 2022).

Despite the absence of basic research illustrating the emergence of the tool and the justification for all of its methodological features, researchers almost immediately began to employ the test for the assessment of real-world attitudes and biases. To illustrate, consider a Meat-eater/Vegetarian IRAP study (Barnes-Holmes et al., 2010 b). IRAPs involving meat and vegetable related target stimuli and positive and negative evaluative concepts were given to a cohort of vegetarians and meat-eaters. The IRAP in this study thus included the following four trial types: Meat/positive, Meat/negative, Vegetable/Positive and Vegetable/Negative. The IRAP was able to discriminate the direction of the bias for participants, and showed that while meat-eaters held a meat/positive and vegetable/positive bias, vegetarian participants showed anti-meat and pro-vegetable biases. Alternatively, the IAT that was also conducted as part of the study showed that both groups held implicit preferences for vegetables over meat, but that for vegetarians, this implicit preference was more pronounced. Being able to discern the direction of the bias, as opposed to just the presence of bias itself, was at the time, an exciting new offering of behavioural technology which no doubt added to the popularity and impact of the IRAP within the literature. However, a process-level based approach should remain the highest priority in the development of any tool to be used in psychological research, including implicit measures. In the absence of such a research program within the IRAP literature, it is difficult to assess the contribution it has actually made to our understanding of core process and the phenomena being measured by the test. In contrast, research into the FAST methodology began with simulations of the IAT effect and systematic and progressive modification of the procedure in published research (e.g., Gavin et al., 2008; Ridgeway et al., 2010) until a sufficiently distinct test format had been arrived at that it deserved its own moniker (O'Reilly et al., 2012), which

of course directly described core process rather than test format and purpose (i.e., a test for the relative acquisition rates of incompatible functional stimulus classes). In addition, the scoring methodology of the FAST endeavors to use as little abstraction as possible, beyond a fairly raw differential response fluency measure. Thus, the current research program focuses on the development of the FAST method, and builds on previous research to extend our understanding of the laboratory-controlled phenomena to which the test is sensitive. It should be apparent at this point, therefore, that the current research program will run parallel to the IRAP research program, and that at present there is no obvious grounds for collaboration between the two methods, given the top-down approach associated with the IRAP.

1.4 The FAST and Stimulus Relatedness

One of the most critical aspects of the behavioural account of implicit measures, and an aspect that is often assumed even in cognitive accounts, is the prediction that the magnitude of effects in implicit measures should be in proportion to the relatedness of the probed stimuli. In the behaviour-analytic field, research into stimulus equivalence yields have established that yield is functionally related to the fluency of the relevant baseline relations (e.g., Bortoloti et al., 2014; Fields et al., 1995;). Correspondingly, FAST scores should theoretically increase in tandem with increasing stimulus relatedness. While the same assumption has been made by IAT (and IRAP) researchers, however, this assumption has never been tested empirically and directly in laboratory-controlled research.

Fortunately, this question is surprisingly amenable to empirical investigation because relatedness can be conveniently objectively manipulated by overtraining (Bortoloti et al., 2013) or assessing relations of differing nodal distance (Moss-Lourenco & Fields, 2011). Cummins et al. (2018) and Cummins and Roche (2020) used both of these methods to assess the impact of relatedness on FAST scores. The 2018 study involved administering baseline MTS training across different periods of time and with different numbers of iterations across

experimental conditions. The study also involved a control condition, in which participants were exposed to a FAST consisting of stimuli that had not been presented during any prior phase, and a second control condition involving the FAST assessment of real word associations of standardized strengths based on the South Florida norms index (Nelson et al., 1998). In all conditions, except for the real word condition, stimuli consisted of nonsense syllables. The conditions involving training of arbitrary stimulus relations consisted of either 1 MTS session, 2 MTS sessions spread across 1 week, 3 MTS sessions spread across 2 weeks, or 3 MTS sessions all conducted in one sitting. A FAST to assess the strength of relations within and between the established equivalence relations was administered following the final sessions of each of these four training conditions.

FAST scores increased as a function of controlled stimulus relatedness, using the slope scoring method (Cartwright et al., 2016). Interestingly, the real word condition produced the strongest effects in terms of learning rate differentials, with the differential effect attributable to a degree of facilitated learning on the consistent block and impeded learning on the inconsistent block. This was the first evidence that scores on any implicit measure could be understood to be a function of the fluency of the relevant stimulus classes established prior to the test. In addition, it provided important information that even over-trained laboratory relations do not have the fluency of real-world verbal relations; thereby providing us with reference points for interpreting test scores (as opposed to the beginnings of standardization of test scores).

In a follow-up study, Cummins and Roche (2020) investigated the impact of varying nodal distances on FAST scores (note: the word ‘node’ is used to delineate the number of stimuli that separate two given exemplars within a stimulus class. For instance, an A1-B1-C1 class has a nodal distance of one). Two four-member equivalence classes (A1-B1-C1-D1, A2-B2-C2-D2) were established using an MTS procedure involving training each zero-node

pair to criterion in succession (i.e., first A1-B1 and A2-B2, then B1-C1 and B2-C2, etc.). Importantly, derived relations were not tested at this point. Three FAST tests were then administered to all participants in a counter-balanced order. The first was a zero-node FAST, in which the strength of A-B relations was tested. A 1-node FAST then probed for derived A-C relations, while the final 2-node FAST probed for A-D relations. An MTS test for all derived relations was then administered. At the group level, FAST scores decreased as nodal distance increased, as expected.

Interestingly, the block slope score for the inconsistent block significantly increased as a function of increasing nodal distance, while slope scores for the consistent block remained unaffected. These trends were visible at the group level and for most individual participants, although a large amount of variability in individual scores was also observed. This move away from group-level analyses and towards an individual level of analysis represents the most pressing next step for research using the FAST. Of course, individual variability is commonly seen on tests for derived relations, especially across differing assessment methods (e.g., Bentall et al., 1999). This alone, however, should not be grounds for a retreat to group-level statistics at the expense of individual-level analysis. Indeed, other implicit measures also exhibit a substantial degree of variability at the individual level (e.g., Klein, 2020; Hussey, 2020).

1.5 Outstanding Questions

Starting from equivalence training based methods in the tradition of Watt et al., (e.g., Roche et al., 2005), to modified implicit association tests (e.g., Gavin et al., 2008), to a native FAST (O'Reilly et al., 2012) and its most recent incarnation (e.g., Cummins et al., 2020), the FAST has been developed generally with an eye to focusing on laboratory-based studies using experimentally-controlled stimulus relations to examine the properties of test performances in a depth greater than typically seen in implicit measures research. However,

many empirical questions remain outstanding. For example, an investigation into the relationship between enforced response times windows (limited hold parameters) and response fluency is yet to be conducted. The relationship between these two variables is almost certainly complex, and the effect of response windows on response fluency is likely to differ at different points in the trajectory of learning.

Another issue yet to be explored relates to the reinforcement contingencies used in these tests. Specifically, a systematic analysis is required of simulated tests in which feedback is provided for correct responses only, or incorrect responses only, alongside an examination of the effect of a thinning of the reinforcement schedule on test scores. It may well be that a thinning of the schedule reduces the fluency on both blocks, or does so disproportionately across blocks, thereby enhancing the sensitivity of the contingencies to pre-experimental learning differences. Indeed, such an investigation could also encompass these same manipulations within the IAT procedure to gain a more detailed understanding of their impact across different procedures.

Yet another question relates to the optimal scoring metric for the test. For instance, rather than assess response fluency differentials across two single blocks of the test, an over-training approach could be taken in which the change in the fluency differential across blocks is accessed across multiple iterations of the test. Larger effects on the first iteration should persist across more iterations of the test than will weaker effects. Thus, a novel and more reliable metric of stimulus relatedness, might involve identifying the point at which learning rate differentials approach zero, or reach a half-way point between the differential on the first iteration and a zero-point differential (i.e., a half-life index). Further questions also remain regarding the optimal number of trials per block, the potential use of various instructions, and acceptable data standardization methods (e.g., log transformation).

In this vein, the utility of a new metric that deals with one potential confound of the learning slope differential method is currently being explored. Specifically, this method does not protect against fortuitous sequences of correct responding produced by rapid random responding. Simulations can trivially demonstrate that a high rate of random responding will produce block-slope scores that are not differentiable from medium-speed highly accurate responding, although the latter is clearly under greater stimulus control than the former. Ideally, learning rates would be corrected for by the attendant rate of incorrect responses per minute. A simple alternative, therefore, would be to calculate the difference between correct and incorrect responses per minute for each block, resulting in a fluency score for each block that reflects the proportionate rate of correct to incorrect responding. The overall FAST effect could then be calculated by subtracting the fluency score for the inconsistent block from that observed for the consistent block, producing what might be called a Response Fluency Differential (RFD) score. Notably, this metric gives primacy to accuracy, as a behaviour analyst would prefer (fast and inaccurate responses will result in very low scores compared to slow and accurate responses), while still accommodating for response times after accuracy has been achieved.

One important guiding principle for the future of FAST research is a stronger emphasis on individual participant effects. Studies using the FAST to date have typically focused on the group level of analysis (but see Cummins & Roche, 2020). Indeed, the same can be said for the IAT and IRAP (although see Finn et al., 2020). In effect, both the fields of social cognition and behaviour analysis are top-heavy with examinations of these measures at the group-level, with comparably little individual-level analysis. Indeed, even in those few studies which have examined individual-level data, they are limited in that the precision of individually estimated scores is rather poor (Klein, 2020; Hussey, 2020). What is needed now for the FAST (and indeed, other measures) to enable the production of translational research

findings, is a renewed focus on the individual level of analysis and improvement of the estimation of individual-level scores. Specifically, FAST researchers should seek to reduce unwanted random error variance while also more precisely estimating the systematic variance of interest (i.e., variance in scores due to stimulus relatedness). This is clearly a lofty challenge; measures of this sort are rarely developed in this manner. However, undertaking this direction of development will aid further in the FAST's development as a truly behaviour-analytic implicit measure, and indeed, will make its' use in practical settings more appropriate. This will be achieved only through a combination of methodological and statistical refinement, with an emphasis on both tight stimulus control and precise measurement.

Whatever the results of the interesting process-level research that will be conducted going forward, it is crucial to the aim of the behaviour-analytic research agenda, and in the interest of collegiality and openness within our science, that no particular methodological feature or scoring mechanism should ever be considered integral to the method, even where empirically supported. In other words, the FAST should be seen as a general methodological strategy linked to a very basic behavioural account of the core effect, in the same way in which Applied Behaviour Analysis represents a scientific strategy rather than a cook-book approach to treatment. All and any methodological and metric variables should be open to modification without claims of the bastardization of the general method. Measures that have achieved apparent proprietary status, with rigid methodological features, instructions, and scoring methods may serve to stagnate research, particularly if results garnished with novel methodologies are considered inadmissible under the umbrella term of the original methodology. If methodological differences are substantiated by sound measurement properties, they should be embraced rather than shunned. Of course, such a wide umbrella approach to methodology can open doors for the possibility of p-hacking (wherein multiple

criteria are employed in analysis until statistically significant results are found). However, the risk of this can be strongly mitigated by preregistration and open science practices, allowing researchers to make clear and transparent delineations between confirmatory and exploratory work (Nosek et al., 2020).

In effect, the FAST methodology is offered as a general starting point for assessing relations in a relatively indirect and convenient way and for indexing the strength of relations between stimuli within and across classes. In that sense, its status is no different to that of a wide variety of equivalence class training methods, such as matching to sample, card sorting, and a wide variety of fluency criteria applied during equivalence class training. These are merely the formats employed to harness well understood behavioural processes and they are not themselves the process. As it stands, the methodology, at its current stage of evolution is public domain and open source, and researchers are encouraged by developers of the method to (attempt to) replicate existing findings, explore new configurations and applications of the task, and push the measure's development forward.

In terms of assessing the relatedness of stimuli in a wide variety of stimulus relations, the FAST could now be considered part of the toolkit of behavioural researchers. Other novel methods have been explored in recent years including card sorting (Fields et al., 2014; Fields et al., 2012), although this indexes only the emergence or non-emergence of a whole class. While useful, card sorting is not a nuanced measure. In contrast, the advantage of the FAST method is that it can be administered (at least in principle) more than once during an equivalence training protocol and can be used as a measure of the increase in relatedness of stimuli within the class across time. It also allows for independent probing of symmetrical and transitive relations (see O'Reilly et al., 2012; 2013). However, one type of stimulus relation that has not yet been subject to indexing by the FAST are relations arising from associative conditioning procedures. That is, when a stimulus has its stimulus functions

changed, for example in an evaluative conditioning procedure, it's relatedness to relevant verbal evaluative terms should also be altered. For example, if in the course of an evaluative or naturalistic emotional learning experience, an individual who is bitten by a dog, might be expected to respond to the verbal relation between the word dog and negative evaluative terms more fluently insofar as they now share highly salient response functions. That is, the words 'dog' and 'pain' may be easier to establish as common members of a functional response class than the words 'dog' and 'pleasure' following the incident. Assessing whether or not this is the case remains one of the last outstanding challenges for the FAST as a measure of the relatedness of stimuli within stimulus relations of all kinds. Answering this final question for the applicability of the FAST is important in informing where and when the FAST, and other implicit measures, may be of use in applied settings.

A very small number of studies have suggested that the IAT may be a valid measure of evaluative conditioning. Gregg et al. (2006) established positive and negative evaluative functions for two imaginary social groups through instruction (rather than contiguous and contingent respondent conditioning) for half the participants, and through reading a short descriptive passage for the other half. An IAT was subsequently administered, employing names representing each fictional social group, and positive and negative evaluative words. In both conditions, IAT effects demonstrated a pattern of responding reflective of the evaluative "conditioning". That is, participants were faster to respond in the same way to stimuli representing positive evaluative words and the name of the fictitious social group for whom positive functions were established through instruction or read narratives.

Another such study conducted by Van Dessel et al. (2015) established differing evaluative functions for two social groups, through the use of a simple approach or avoidance training session, which was guided by instructions to approach or avoid stimulus exemplars on screen using a simple computer keyboard response operandum. This was done across two

conditions, using two fictitious (E.g., Niffites/Luupites) social groups in one condition, and two real groups with known valence (e.g., Blacks/Whites) in another. Using this procedure, appetitive approach functions were established for one fictitious social group the (e.g., Niffites/Blacks) and negative avoidance responses for the other (e.g., Luupites/Whites). This configuration was reversed for half of the participants in each condition, so that aversive functions were established for the Niffite/Black group and appetitive for the Luupite/White group. An IAT with exemplars representing the two social groups and positive and negative words was then administered. In the fictitious social group condition, participant performance on the IAT was in line with the approach and avoidance training, insofar as the arbitrarily created appetitive or aversive functions of the fictitious social groups influenced the speed of responding when responses shared positional keyboard properties with incompatible positive or negative evaluative terms. In the real group condition, the IAT failed to detect conditioning, suggesting that pre-existing racial evaluations for these participants counteracted attempts at establishing conditioning (e.g., establishing anti-White evaluations).

Findings from both Gregg et al. (2006) and Van Dessel et al. (2015) indicate the IAT is a useful measure of the relatedness between stimuli established as a result of associative conditioning procedures. A further study from Fazio and Olsen (2001) demonstrated the IAT to be an effective measure of evaluative functions using a more traditional evaluative conditioning procedure, wherein CSs were presented contingently and contiguously with USs over a series of trials. That is, one Pokémon character (CS+) was presented repeatedly with appetitive visual and verbal stimuli (US+), and another (CS-) with aversive visual and verbal stimuli (US-). This procedure was then followed by administration of an IAT, including both CS+ and CS- and positive/negative evaluative terms as stimuli. Performances on the test aligned with the conditioning contingencies. That is, participants were quicker to respond in

the same way to appetitive condition stimuli and positive evaluative terms, as well as aversive conditioned stimuli and negative evaluative terms.

One FAST study to date has already attempted to directly manipulate stimulus valence using an instructional-type procedure, rather than an explicit associative conditioning procedure. Specifically, in Cummins et al. (2019), health worker participants were exposed to one of the following interventions: (a) a positive message about condom use, (b) a negative message about condom use, or (c) no message. Following this, all participants completed two single-target style FASTs (see O'Reilly et 2012, 2013), employing images of condoms, positive words, images of the sky (neutral stimulus) or number words (neutral). The first test assessed relations between positive words and condoms (and the relation between the two neutral stimulus classes), and the second assessed the relatedness of negative words and condoms (and the relation between the two neutral stimulus classes). Results showed that the FAST score (in this case a Strength of Relations Index; see O'Reilly et al., 2012, 2013) differed significantly across conditions. That is, the FAST was sensitive to the evaluative functions established for condom stimuli through the initial laboratory procedure designed to establish positive or negative functions for these stimuli. This research outcome, considered together with the results of Cummins et al. (2018), suggest that the FAST should be able to index the strength of unconditioned, emotionally salient stimuli encountered in learning experiences, vis-à-vis the resulting change in relatedness between emerging conditioned stimuli and evaluative terms. The current research thesis will focus on examining this issue.

1.6 The Current Research

The current research aims to determine the utility of the FAST method for assessing the relatedness of stimuli resulting from evaluative/associative conditioning experiences. More specifically, this research aims to determine not only the existence of, but also the relative *strength* of evaluative functions that have been established through conditioning and

manipulated across conditions, as measured by the FAST. In this case, the evaluative conditioning procedures are conceptualized as laboratory analogues of everyday casual associative learning experiences (e.g., such as being bitten by a dog). The rationale for this is to produce a study that is as closely representative of realistic emotional experiences as possible. The benefit of this is that it will expedite future, more applied research, for example in investigations into whether the FAST may prove useful in clinical contexts where emotional experiences are of interest to inform patient treatment or diagnosis. Though the authors do not view the FAST as a diagnostic tool, nor is it used as such in this research, that is not to suggest that it could not provide some utility in such contexts. Indeed, this would be beneficial to investigate, though it is beyond the scope of the current research. To achieve the aforementioned laboratory analogue of associative learning, it was deemed appropriate to adopt some methodological approaches that may not be typical of behavioral research traditions. For example, some aspects of the evaluative conditioning procedure were altered to produce what is hoped to be a more broadly applicable body of research than would be if the research was conducted strictly within a behaviour analytic paradigm. Similarly, the researchers availed of naturally occurring linguistic categories (see below) to further achieve more realistic and transferrable results.

The main dependent variable, the FAST, has been discussed at length, and is joined by Likert scale stimulus ratings that were recorded for conditioned stimuli (see below). The independent variable was also maintained throughout all experiments, and is represented by the salience of the unconditioned visual stimuli used to instantiate evaluative functions for conditioned stimuli. This variable consisted of three levels of intensity, differentiated by the standardized ratings of salience and valence provided for each stimulus. Each participant was randomly assigned to one level (referred to as condition) at the beginning of the experiment. The purpose of this manipulation was to provide a quantitative difference in the level of

conditioning, that the FAST was later used to measure. As such, this research answered two questions, the first being whether the FAST could differentiate between basic appetitive and aversive conditioned stimuli. The second, more nuanced, question investigated was whether the magnitude of FAST scores varied as a result of the intensity of the unconditioned stimulus employed to establish stimulus functions.

Experiment 1 was conducted with an online, unsupervised participant sample, none of whom received remuneration but many of whom received course credit for participation. In this experiment, positive and negative evaluative functions were established for two separate innocuous stimulus classes (i.e., fruit and furniture), by pairing them contiguously and contingently with relevant emotionally salient visual stimuli. The emotional salience of the imagery was verified systematically across three conditions, with Condition 1 involving the most salient stimuli, and Condition 3 the least. Thereafter, a Likert rating scale was administered as a manipulation check, to ensure the conditioning was effective to a point where explicit acknowledgement of the conditioned aversive or appetitive functions of the CS was measurable. The FAST was then administered to all participants, employing the same conditioned stimuli as target stimuli, along with novel positive and negative evaluative terms as the evaluative stimuli. It was expected that FAST scores would reflect the conditioning contingencies to which participants were exposed insofar as more rapid acquisition of functional response classes involving appetitive stimuli and positive evaluative words, as well as aversive stimuli and negative evaluative words, would be observed compared to the orthogonal arrangement. It was also expected that the magnitude of the test effect size would vary as a function of the salience of the unconditioned stimuli used during the evaluative positioning procedure. In effect, approaching the FAST as a measure of stimulus relatedness (Cummins et al., 2018), this outcome would provide evidence that more salient unconditioned stimuli lead to higher levels of stimulus relatedness between conditioned and

unconditioned stimuli and that this increased relatedness is measurable using implicit style measures such as the FAST. Addressing this question was important, as it remains one of the last outstanding questions regarding the sensitivity of the FAST measure to stimulus relations of various kinds. However, it was also pressing because such an effect has already been shown for the IAT (Gregg et al., 2006; Van Dessel et al. 2015), and with the FAST (Cummins et Al., 2018), though different conditioning procedures were employed. The effects found in the first experiment were broadly supportive of the hypothesis, but a disappointing rate of necessary data exclusion inspired a replication in Experiment 2.

Experiment 2 involved the replication of the first experiment with a much larger, paid, online sample. This was done in an effort to identify a true effect, achieve impressive statistical power despite a large rate of data exclusion for non-adherence to the task among other issues, and to allow for more investigative post-hoc analysis in order to understand noise in the data if this were to arise again. A larger sample was also secured in an effort to conduct all data analyses while retaining outliers in the data set (other than those due to non-adherence to the task), which themselves represented some of the variance under analysis in this very research. Including outliers, however, masks true effects and must be compensated for in some cases by larger sample sizes. These results of this experiment appeared to be clearer than have been obtained in Experiment 1 and all hypotheses were supported. However, given the undesirability of resorting to sample size inflation in order to enhance experimental effects, it was decided to replicate the experiment once again but with a considerably smaller sample size. In this case, however, the effort was made to enhance adherence to the task through laboratory supervision that is traditional in this field of research.

Experiment 3 represented an in-person replication of Experiments 1 and 2 in a return to traditional behaviour analytic research approach, characterized by high experimental

control. It was hoped that the clear effects obtained an Experiment 2 would once again emerge with a small sample, but this was found not to be the case. Poor adherence to the task was still an issue, even for a supervised sample, and the effects, while broadly supporting the hypothesis, were not as clear as anticipated.

The experiments reported in this thesis have inadvertently formed a commentary on the viability of collecting data online using paid participants samples, versus the use of in person methodologies in a traditional university laboratory setting. This was not the initial goal of the research, but nevertheless needed to be pursued in light of problems with adherence to the task and large amounts of noise in the data. Conclusions are drawn in the general discussion and within each of the chapters to follow, regarding the relative merits of paid and unpaid participation by anonymous online research volunteers, versus the use of a typical undergraduate university participant population in a supervised setting. The general conclusion drawn is that there is no visible advantage to using the in person supervised training and testing format, and that the absence of an experimenter does not seem to increase disengagement from the task. More importantly, in relation to the original aims of the research, these were broadly supported across all three experiments.

Chapter 2

Assessing the FAST as a Measure of Emotionally Salient Experiences

Experiment 1

Experiment 1

2.1 Introduction

The current experiment first involved establishing emotional stimulus functions for everyday English words from the vernacular, which were assumed to have innocuous emotional functions. This was achieved using an associative evaluative conditioning procedure and a set of visual unconditioned stimuli which varied in positive or negative emotional intensity across three independent conditions. Participants were randomly assigned to a low, medium or high stimulus salience and arousal condition by the experimentation software. The top line purpose of the study was to assess the sensitivity of the FAST procedure to the difference in emotional stimulus functions across two vernacular stimulus classes established experimentally. A secondary purpose was to assess whether or not the strength of the emotional functions of the unconditioned stimuli employed impacted upon FAST scores.

As demonstrated in a small sample of studies outlined in the Introduction, the IAT may be a useful measure of laboratory established stimulus evaluation. However, in two of the three relevant studies, the laboratory created stimulus functions were established using narrative procedures. Thus, the relevant functions were established in accordance with complex verbal contingencies that may have involved a degree of derived relational responding. Only one IAT study (Fazio & Olsen, 2001) has employed unambiguous respondent conditioning procedures before the administration of an IAT to assess the resulting stimulus relatedness. Even the one FAST study that is relevant to addressing this issue did not employ a sufficiently unambiguous associative conditioning procedure to draw conclusions about the sensitivity of the FAST to respondently conditioned relations (Cummins et al., 2019). For this reason, it seems that research into the FAST procedure

should pursue assessing its' utility in measuring conditioned relations, but using more traditional laboratory methods for establishing respondently conditioned relations.

Pursuing this research question is an obvious next step for the field, given that it is a glaring knowledge gap arising from the production of several studies over the past number of years examining the sensitivity of the FAST to: simple two-member relations established through direct matching training (O'Reilly et al., 2012), three-member derived equivalence relations involving arbitrary stimuli (O'Reilly et al., 2013), naturalistic verbal category relations representative of stereotypes (Cartwright et al., 2016; Cummins et al., 2019), laboratory-created, three-member equivalence relations of varying levels of controlled relatedness (Cummins et al., 2018), two-node equivalence relations (Cummins et al., 2020) and natural verbal category compatibilities of known strength (Cummins et al., 2020). Of course, we acknowledge that it is a theoretical underpinning of the IAT that the relation types being assessed by that method are themselves direct associations and that the social cognitive research community has not speculated on the role of derived verbal relations in the generation of IAT effects. Nevertheless, this distinction is important from a behaviour analytic point of view because a method such as the IAT or its variants may be sensitive to one relation type and not another, or to one or another to different degrees. In addition, despite a lack of interest in the question, it may well be the case that most effects observed in administered IATs are in fact a result of relations established by means other than direct association, but such a question cannot be answered in the absence of even a single study addressing the issue of relation types and their effect on IAT scores.

Understanding the role of respondent conditioning in generating implicit test effects is especially relevant to our understanding of their potential use in applied settings (c.f., Roche et al., 2005; Gavin et al., 2008). Thus, the first experiment in this thesis will examine whether or not the FAST method is sensitive to laboratory conditioned relations as an analogue of the

emergence of simple associations following brief naturalistic associative learning experiences (i.e., as opposed to unconscious associations as assumed in the IAT model; Nosek et al., 2007). A corollary question, however, relates to the role played by the strength of the unconditioned stimulus in such processes. That is, it is almost an axiom of the respondent conditioning literature that the potency of the unconditioned stimulus increases the efficiency of the conditioning procedure and, relatedly, the potency of the conditioned stimulus (see Bevins et al., 1997). What is not yet known, however, is whether or not this increased potency of the conditioned stimulus enhances the relatedness of the CS and the US. The current research will address this question by manipulating the potency of the US across three conditions prior to the administration of a FAST procedure to assess CS-US relatedness.

For the current experiment, it was hypothesized that a FAST procedure employing target stimuli for whom emotional functions had been established using a respondent conditioning procedure, would produce FAST scores that were indicative of the experimentally manipulated emotional valence of those stimuli across test blocks (i.e., a main effect). It was also hypothesized that this block difference effect would vary significantly across three conditions in an interaction pointing to the measurable impact of the intensity of the unconditioned stimuli on the FAST effect.

2.2 Methodology

2.2.1 Participants

Eighty six participants volunteered for the study and participated online. Following application of necessary data exclusion criteria (see Results), data for a total of 62 participants were retained for analysis (age M= 26.34, SD= 13.51). Of these, 44 self-identified as female and 15 as male. The remainder (3) identified as non-binary. The participants in this study consisted largely of undergraduate psychology students seeking

course credit for participation in the study, but due to the anonymous nature of the study the precise proportion cannot be known. Exclusion criteria included: being under 18 years of age, not being fluent in English, and having an anxiety-related condition that would make viewing aversive images unadvisable.

2.2.2 Ethical Considerations

This study received ethical approval from the Maynooth University Research Ethics Committee. Participants were informed that they were free to cease their participation in the study at any point. Due to the anonymisation of data however, data retraction was not possible, and participants were informed of this at the outset of the study.

Due to the potentially distressing nature of the images that participants would be exposed to through the course of the study, participants were advised against participating if they had a history of anxiety related issues that would make viewing such images unadvisable. Participants were asked to use their discretion on this matter. Participants were also required to be over 18 years of age.

2.2.3 Apparatus

Participants accessed this study via an internet link and completed the experiment on a device and in an environment of their choosing. The entire procedure was delivered by the Inquisit software (Millisecond.com) platform.

The evaluative conditioning procedure (see below) involved the presentation of a series of visual images as unconditioned stimuli and a series of words as CS. . US were taken from the International Affective Picture System (IAPS; Lang et al., 2008). This is a set of images standardized in terms of their emotional valence and arousal coefficients. Three different sets of images were selected as US: one set for each of the three experimental conditions. Each set contained 8 stimuli, consisting of four aversive and for appetitive

stimuli. Sets differed from each other in terms of their standardized arousal and aversiveness coefficients (see Table 2.1 for ratings and image identity numbers).

As examples, one aversive image from Condition 1 contained prisoners tied up and blindfolded, another contained bloodied animal remains. In Condition 2, aversive images included a meat slicer containing cut up meat. The least aversive condition, Condition 3, contained pictures of fish being grilled, and a disheveled girl pouring wine. Appetitive imagery for Condition 1 included groups of smiling children and an attractive woman at the beach. Condition 2 images consisted of images of people smiling and pastries. Condition 3 included an image of an elderly couple, and another of a basket of fruit being carried. Condition 1 employed the most highly aversive/appetitive stimuli; Condition 2 employed moderately aversive/appetitive stimuli and Condition 3 used the least aversive/appetitive stimuli.

The CS consisted of two sets of four verbal stimuli. Aversive stimulus functions were established for one set while appetitive functions were established for the other set in a procedure in which each of the stimuli was employed several times. One of the verbal classes employed consisted of exemplars of fruit (i.e., Apple, Pear, Orange, Banana) and the other consisted of four verbal exemplars of furniture (i.e., Desk, ,Chair, Table, Sofa). Naturalistic linguistic categories (i.e., fruit, furniture) were taken advantage of to eliminate the need to train stimulus classes within the study (i.e., in order to expedite the evaluative conditioning of an entire verbal class for use within the FAST procedure). Using natural word categories also supports the move from basic to more applied research within the FAST research program, as it will indicate some of the considerations that are necessary when employing real world stimuli.

Table 2.1

Standardised valence and arousal ratings for each of the stimuli employed in each of the conditions as unconditioned stimuli (US). Image numbers refer to the IAPSs identifier for the relevant image.

	Aversive US				Appetitive US		
	Image	**Valence**	**Arousal**		**Image**	**Valence**	**Arousal**
Condition 1	7380	2.46	5.88	**Condition 1**	2345	7.41	5.42
	9400	2.50	5.99		4220	6.60	5.18
	9419	2.82	5.10		5260	7.34	5.71
	9500	2.42	5.82		7220	6.91	5.30
M		2.55	5.70	***M***		7.07	5.40
Condition 2	1280	3.66	4.93	**Condition 2**	2005	6.00	4.07
	6561	3.58	4.44		1947	5.85	4.35
	7361	3.10	5.09		4004	5.14	4.44
	9404	3.71	4.67		7402	5.98	5.05
M		3.51	4.78	***M***		5.74	4.48
Condition 3	1112	4.71	4.60	**Condition 3**	2214	5.01	3.46
	2752	4.07	4.84		2396	4.91	3.34
	6800	4.02	4.87		2980	5.61	3.09
	7484	4.99	4.24		7484	4.92	4.08
M		4.45	4.64	***M***		5.11	3.49

Note: M= Mean

A 7-point Likert scale was administered to record the evaluative functions of the stimuli following conditioning. The scale required participants to rate each of the CS on the seven-point scale, where 1= very aversive and 7= very pleasant (see Appendix I).

The Function Acquisition Speed Test (FAST) employed the 8 CS used in the conditioning procedure (e.g., names of furniture or fruit; see above) as target stimuli. Positive and negative evaluative stimuli were drawn from a range of those typically used in implicit testing research and consisted of the words Heaven, Love, Pleasure, Peace, Death, Filth, Murder and Sickness.

2.2.4 General Experimental Sequence

All phases of the study (evaluative conditioning, stimulus ratings and FAST) were administered online using the research tool, Inquisit (Millisecond Inc.). Participants accessed the study via a link which led them to the Millisecond Inc. server. Participants completed the study on their own devices, and in at a time/setting of their choosing. The information sheet (see Appendix II) provided to participants strongly suggested that they participate in the study in a quiet, distraction free environment and on a desktop computer rather than a mobile device, although this could not be controlled for remotely. After providing consent (see Appendix III), participants were exposed to a brief demographic questionnaire (Appendix IV), and were proceeded immediately to the evaluative conditioning phase. Participants then rated the subjective aversiveness of the conditioned stimuli on the 7-point Likert Scale (Appendix I). The FAST procedure was administered immediately after the rating scales, followed by a simple on-screen debrief (see Appendix V).

2.2.4.1 Evaluative Conditioning

Participants were instructed to simply look at the screen and to continue to pay attention to the sequence of events presented on screen. The importance of paying attention was emphasized in the instructions, by stressing that the information presented during this phase would be important in the subsequent phases.

There were 32 conditioning trials in total (i.e., 16 CS+ and 16 CS-), with trials presented in a quasi-random order so that no more than two successive trials consisted of a CS+ or CS-. The Inquisit software also controlled the random selection of appropriate CS and US stimuli on each trial. One second before the presentation of the US, a white fixation cross appeared in the centre of the screen. One of the 8 unconditioned fruit or furniture words was then presented to participants for one second, followed by a blank screen intertrial interval

(ITI), ranging randomly from 8-12 seconds. After the ITI, the relevant US appeared on the screen and remained for five seconds (e.g.., 'PEAR' followed by an appetitive image). Immediately after the US was removed from the screen, another ITI commenced. This conditioning paradigm can be described as a trace conditioning method.

The Inquisit software managed the random administration of conditions, and therefore which stimuli were employed in each administration of the evaluative conditioning protocol. That is, the three conditions of this study involved identical conditioning procedures but employed US of varying arousal and valence. As a reminder, Condition 1 employed the most aversive and appetitive unconditioned stimuli, Condition 2 employed moderately aversive and appetitive stimuli, whereas Condition 3 employed the least aversive and appetitive stimuli. This phase took 6 minutes approximately to complete.

2.2.4.2 Conditioned Stimulus Rating Scales

A series of Likert rating scales were administered to participants as a manipulation check for the conditioning procedure. Specifically, participants were asked to provide a rating on a Likert scale (1= very aversive, 7= very pleasant) for each of the CS established in the evaluative conditioning procedure. This procedure also provided dependent outcome data for the purpose of conducting a convergent validity analysis with FAST scores. Participants were instructed to indicate using the rating scale how much they associated the verbal stimuli employed throughout the evaluative conditioning phase with positive or negative thoughts or feelings. It was advised that ratings be made with little deliberation.

2.2.4.3 Function Acquisition Speed Test

The FAST was used to measure the strength of relations between the CS and semantically related verbal evaluative stimuli that were not previously employed in the study. Thus, the FAST was, in effect, used to assess the generalized effects of the conditioning

procedure on the relatedness of the conditioned stimuli to evaluative terms drawn from the vernacular.

Participants were issued brief on-screen instructions (see below) for completing the FAST before the procedure began. These informed the participant that they should respond as quickly and accurately as possible to stimuli as they appeared on the screen, and that they should use the feedback provided to them after every response to guide their subsequent responses. The importance of sustained attention on the task was emphasised. Attention was then directed toward the two operanda on the computer keyboard (e.g., Z and M keys). Participants were instructed to press the spacebar when they were ready to begin. The instructions presented to participants read as follows:

> "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."

The Inquisit software randomly determined which block (i.e., consistent, or inconsistent) would be presented first for each participant. Participants then completed a block of 52 trials in which exemplars from each of the two targets in these categories (fruit or furniture; positive or negative) were drawn randomly on a trial-by-trial basis. Thus, while there was no control over how many exemplars from each of the four stimulus classes (two conditioned stimulus classes and two novel evaluative word stimuli classes) was presented in each block, the 52-trial block length accommodated the possibility of an equal number of

stimulus presentations from each of the four classes (i.e., an equal number of conditioned appetitive, conditioned aversive, positive evaluative and negative evaluative stimuli).

Immediately upon completion of the first block, instructions for the completion of the second block were presented. These were different from the first set of instructions only by the indication that response contingencies may have changed but otherwise that the nature of the task would remain the same. No instructions were provided on screen during the trials in either block. The instructions for the second block were as follows:

> "Well Done! In the next part of this task, words will again begin to appear on the screen, one at a time. You must again learn to press either the 'Z' or the 'M' key, depending on what type of word appears, and based on the feedback that you are given after each response However, during this next part of the task, the rules of responding may have changed. Your task is to learn which type of key press wis required for each type of word. Try to respond AS QUICKLY AS POSSIBLE. When you're ready, press the spacebar to begin."

Immediately following the initiation of the procedure by the participant, the first intertrial interval (ITI) of 0.5 s was presented (i.e., a blank screen). After this initial ITI, the first stimulus (e.g., "PEAR") was presented in the centre of the screen, in large font . The participant was then required to respond by pressing either the "Z" or "M" key within a 3000ms response window. If a response was made within the 3000ms, the stimulus was removed from the screen and replaced with appropriate feedback (e.g., 'CORRECT appearing in red at the centre of the screen for 0.5 seconds). If no response was made within the 3000ms window, the stimulus was removed from the screen and corrective feedback for an incorrect response was given (i.e., WRONG appearing in red at the centre of the screen for 0.5 seconds). This reinforcement contingency can be described as a FR1 with a limited hold.

To help the reader to understand exactly the nature of the procedure, an example is provided below. This example is what a participant who received a fruit-positive, furniture-negative configuration within the conditioning procedure would have experienced. On the 'consistent' block (i.e., the block that outlines the correct pattern of responding as correspondent with the configuration of stimuli during the evaluative conditioning procedure), this participant would have been reinforced (e.g., CORRECT would appear on screen) for pressing the 'M' key in response to both fruit words and positive words. Similarly, reinforcement would have been given for pressing the 'Z' key in response to furniture and negative words on the consistent block. If, for instance, that same participant responded to a fruit exemplar with the 'Z' key (i.e., a response pattern *inconsistent* with conditioning), this would be recorded as an incorrect response, and WRONG would appear on screen. On the inconsistent block, wherein the response contingencies are reversed, that is, they no longer correspond with conditioning, the participant was reinforced for a pattern of responding similarly to fruit and negative words (e.g., 'M' key), and similarly to furniture and negative words (e.g., 'Z' key). If a response pattern consistent with conditioning emerged (e.g., responding with the 'M' key to both fruit and positive exemplars), negative feedback was provided to the participant following their response.

The Inquisit software recorded the number of correct and incorrect responses on each block of the FAST, as well as the time taken to make each response and complete the block. A fluency differential rate was then calculated for each block by dividing the difference between the correct and incorrect response total by the time in seconds taken to complete the block. A total FAST score was calculated by subtracting the fluency differential score of the inconsistent block from that of the consistent block, and multiplying the result by 60,000 to produce a per-minute Response Fluency Differential (RFD) score. This number reflects the degree to which the response fluency on the consistent block (corrected for by the

inconsistent response rate) is greater than that observed on the inconsistent block. Equation 2.1 displays the formula used to compute the RFD.

Equation 2.1

RFD or FAST score calculation formula

$$\text{RFD} = \left(\left(\frac{TCC-TIC}{TTC}\right)\right) - \left(\left(\frac{TCI-TII}{TTI}\right)\right) \times 60000$$

TC = Total Correct, TI = Total Incorrect, TT = Total Time, 'C' terminate indicates the consistent block, 'I' terminate indicates the inconsistent block.

Note: This formula is to be used when only the accuracy rates per block are recorded. For the purpose of the experiments in the current research, it is important to note that the InQuisit software calculated the rate of accuracy of each block *per minute*, therefore eliminating the need to apply the final step of multiplying by 60,000 in this instance. The RFD score is a slight deviation from the previously used block learning rate differential method. However, the RFD scoring method has advantages that make its' use more preferable (see Cummins et al., 2018 and Discussion section).

2.3 Results

2.3.1 Missing Data and Excluded Cases

A small number of participants (n=4) were excluded from analysis due to producing no responses at all on one or both FAST blocks. For a task in which there are only two possible response options, a 50% correct response rate can be achieved without paying any attention to the task requirements at all. It was therefore also necessary to exclude participants who failed to achieve above-chance response accuracy. Because the FAST method involves an interest in the speed of responding as well as accuracy, it was decided that such a criterion would not be based on accuracy alone. While recognizing that any such

exclusion criterion will be arbitrary, the current study settled on a data exclusion criterion that focused on what would approximate a typical chance-level rate of responding per unit of time. More specifically, it was identified that a rate of >10 incorrect responses per minute amounts to approximately 25 incorrect responses across the 52-trial block over approximately 2 1/2 minutes (i.e., around chance level responding). The application of this very conservative exclusion criterion led to the elimination of the datasets for 11 participants. Following this, datasets for participants who failed to demonstrate evidence of conditioning according to conditioning contingencies through their stimulus ratings were also excluded. Specifically, participants who failed to rate the conditioned aversive stimulus as more aversive than the conditioned positive stimulus, regardless of the size of the difference, were excluded from further analysis (n=13). The final sample size used in this analysis was n= 62.

2.3.2 Descriptive Statistics

The means and confidence intervals for FAST scores (RFD), individual task block fluencies and stimulus ratings are provided in Table 2.2. It is important to note that the figures in this section are purely descriptive of numerical trends in the data, and inferential analyses concerning the statistical significance of any differences between scores can be found in subsequent sections. A positive mean RFD score was observed for all conditions as expected, indicating that in all conditions, response fluencies were higher (on average) for the consistent block relative to the inconsistent block. However, participants in Condition 1 (C1) did not produce the highest FAST scores as expected, but rather produced scores intermediately (M= 3.73, 95% CI: -.78 – 7.79) relative to the other conditions. The highest FAST scores were recorded from Condition 2 (C2; M= 4.93, 95% CI: 1.43 – 8.27). As expected, the lowest FAST scores were calculated for Condition 3 (C3; lowest emotional salience USs; M= 3.70, 95% CI: 1.72 – 5.69). These figures suggest that the expected trend

of decreasing FAST scores as a function of decreasing CS salience was not observed at the group level.

Table 2.2

Means and 95% Confidence Intervals for RFD scores, individual block fluency scores and CS ratings for each condition and the combined cohort.

	C1		**C2**		**C3**		**Cohort**	
N	19		18		25		62	
	M	95% CI	M	95% CI	M	95% CI	M	95% CI
RFD	3.73	-.78 – 7.79	4.93	1.43 – 8.27	3.70	1.72 – 5.69	3.07	1.33-5.33
ConFluency	20.32	17.77 – 22.65	20.25	18.06 – 22.25	18.95	17.04 – 20.53	19.75	18.42-20.98
InconFluency	16.58	13.93 – 19.18	15.33	12.80 – 17.7	15.24	13.53 – 16.94	15.68	14.53-17.00
AvgCSPos	5.48	5.11 – 5.87	4.95	4.57 – 5.37	4.59	4.32 – 4.90	4.97	4.73-5.20
AvgCSNeg	2.20	1.70 – 2.69	3.18	2.76 – 3.58	3.26	2.91 – 3.59	2.91	2.66-3.20

Note: C: Condition. RFD: Reaction Fluency Differential (FAST score). ConFluency: correct - incorrect responses per minute on the consistent block, InconFluency: correct - incorrect responses per minute on the inconsistent block. AvgCSPos/Neg: Average Conditioned Stimulus ratings for positive/negative conditioned stimuli.

An average conditioned stimulus class evaluative rating was calculated for each participant by collapsing the ratings of the four individual class exemplars, providing us with an index of stimulus class appetitiveness (i.e., higher scores are indicative of a lower aversiveness/higher appetitiveness). The use of this procedure primarily functioned as a manipulation check for the conditioning phase and was the basis of data exclusions for the current analyses (see above). As expected, the aversive conditioned stimuli were rated as most aversive / least appetitive by participants in C1 (M= 2.20, 95% CI: 1.70 – 2.69), moderately aversive and appetitive by C2 participants (M= 3.18, 95% CI: 2.76 – 3.58) and least aversive and most appetitive by C3 participants (M=3.26, 95% CI: 2.91 – 3.59). As expected, the appetitively conditioned stimuli were rated in accordance with the opposite pattern across conditions. That is, the highest ratings of appetitiveness were recorded for C1

participants (M= 5.48, 95% CI: 5.11 – 5.87), C2 participants produced moderate ratings (M=4.95, 95% CI: 4.57 – 5.37), while ratings recorded for C3 participants were indicative of near neutral evaluations of the stimuli (M= 4.59, 95% CI: 4.32- 4.90). These ratings indicate that at face value, the evaluative conditioning phase was successful in establishing the intended stimulus functions and that these varied as intended across conditions prior to the administration of the FAST procedure.

2.3.3 Correlations

In an attempt to assess the convergent validity of the FAST, the relationships between RFD (FAST) scores, CSpos and CSneg ratings were investigated using the Pearson product moment correlation coefficient (see Table 2.3). The assumptions of linearity, normality and homoscedasticity were not violated. As expected, the relationship between CSpos and CSneg ratings was strong and negative ($r = -.500$, $p < .001$). A positive, medium correlation between CSpos and RFD scores ($r = .438$, $p > .001$) was found, indicating that higher CS appetitiveness ratings are associated with higher RFD scores. Similarly, a negative, medium correlation was found between RFD scores and CSneg ratings ($r = -.457$, $p > .001$), indicating that more negatively rated aversive CS are associated with higher RFD scores.

Table 2.3

Correlations matrix displaying relationships between RFD scores and average ratings for aversive and appetitively conditioned stimuli.

		1.	2.
1.	**RFD**	1	
2.	**CSPos**	.438**	1.
3.	**CSNeg**	-.457**	-.500**

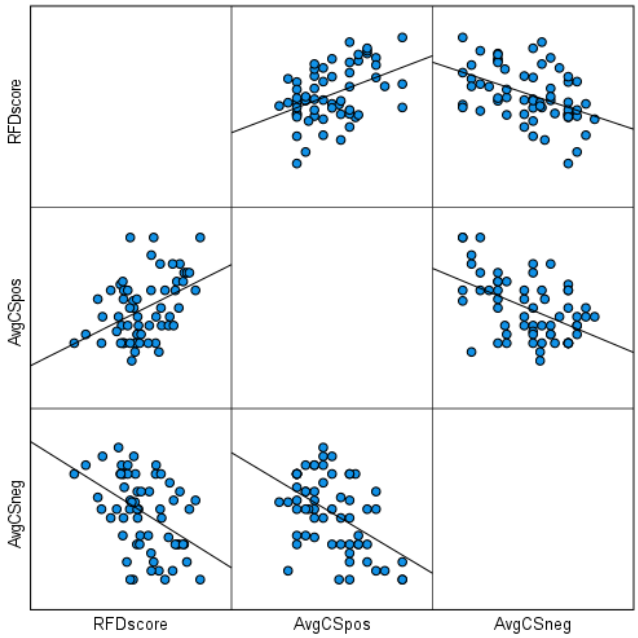
** $p < .001$

These trends are in line with the expected outcomes of the conditioning contingencies in terms of establishing evaluative responses to the conditioned stimuli. These evaluations

also coincided with the standardized arousal and valence ratings of the IAPS stimuli. Figure 2.1 shows a matrix of the scatterplots with regression lines for each of these correlations.

Figure 2.1

Scatterplot matrix indicating the relationship between stimulus ratings and RFD scores.



The scatter plots clearly illustrate the expected negative and positive correlations between variables as indicated by the clustering of scores around the regression line. These outcomes suggest that the FAST was sensitive to the subjective evaluations of stimuli created by the conditioning procedure.

2.3.4 Quantifying the Effect of Stimulus Function Assignment on Conditioning

A 2x2 mixed factorial ANOVA was conducted to assess the impact of stimulus function assignment (i.e., whether or not fruit or furniture stimuli were established as aversive or appetitive conditioned stimuli) on the fluency scores for each block. There was a significant interaction between stimulus function assignment and block fluency scores [$F(1,60)= 8.376$, $p=.005$], with a medium effect size ($\eta 2= .123$). There was also a main effect for block [$F(1,60)= 14.515$, $p <.001$], with a large effect size ($\eta 2= .195$). This result indicates that there was an overall group level FAST effect (i.e., response fluency differential across blocks in the predicted direction) for the cohort as a whole irrespective of the randomized

functions of the fruit and furniture stimuli. A review of the mean block fluency scores recorded in each condition (see table) also indicates that the effect was descriptively lower for those for whom furniture was established as a CSpos.

Table 2.4

Table displaying mean consistent and inconsistent block fluency scores for each stimulus function assignment condition separately.

Stimulus Function Assignment	Mean Consistent Fluency	Mean Inconsistent Fluency
Fruit-Positive	21.14	15.03
Furniture-Positive	17.54	16.70

2.3.5 Quantifying Block Sensitivity to Conditioned Stimulus Valence: Block Fluency Scores

A 2x3 mixed factorial analysis of variance was conducted to explore the impact of condition on individual block fluency score differences (i.e., FAST effect). There was no significant interaction effect between individual block fluency differences and condition [$F(2,59)= .165$, $p= .848$]. This suggests that the statistically significant difference in response fluencies across blocks (i.e., the FAST effect) did not vary significantly by condition. The main effect for block was significant, [$F(1,59)= 18.311$, $p < .001$], with a large effect size ($\eta 2= .237$). In other words, for all three conditions combined, there was a significant difference in response fluency across the blocks (i.e., a significant FAST effect). However, the variation in scores across conditions was not statistically significant. Figure 2.2 illustrates the estimated marginal means of the individual block fluencies for each condition. It is noteworthy that for this particular analysis, the main effect of condition was not a matter of concern, or indeed a psychologically meaningful variable, and therefore will not be reported in this or subsequent iterations of this analysis.

Figure 2.2

Line graph depicting mean differences in block fluencies across conditions.

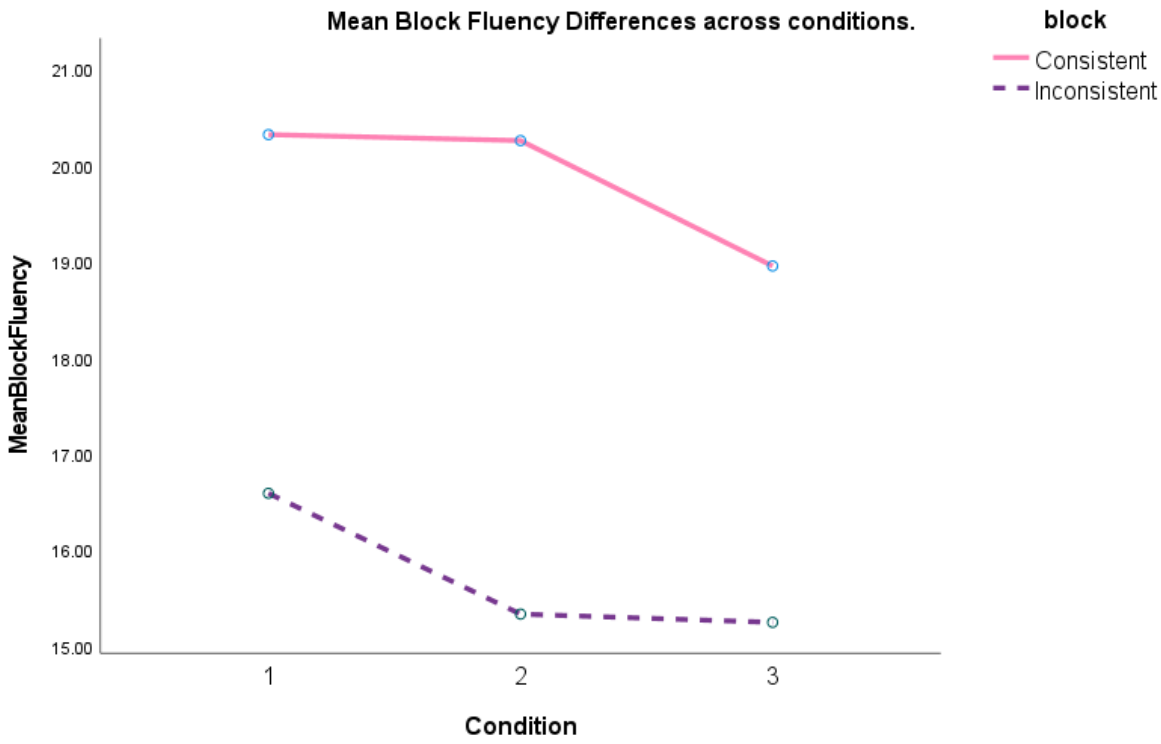


Figure 2.2 demonstrates a visible difference in fluency across the two blocks, albeit this is not reflected by a statistically significant difference in magnitude across conditions. A general trend toward the expected decrease in fluency as governed by condition was seen here.

2.3.6 Quantifying FAST Sensitivity to Conditioned Stimulus Valence: RFD Scores

A one-way analysis of variance was conducted to assess whether the magnitude of change in RFD scores (as opposed to block fluency scores) across conditions was statistically significant. Results showed no significant difference in mean scores across conditions [$F(2,59)= 0.165$, $p= .848$]. These findings support the conclusion that in the current

experiment the FAST was not sensitive to the varied emotional salience of conditioned stimuli.

2.3.7 Planned Comparisons

Planned comparisons were conducted to determine whether or not significant FAST effects (i.e., response fluency differentials across the conditions) were observed within any of the conditions considered separately. Three separate paired samples t-tests were conducted, with the Bonferroni adjustment applied to the alpha accordingly (i.e., p value adjusted to $p > .017$). There was no significant block fluency differential calculated for C1 across either consistent (M= 20.32, 95% CI: 17.77-22.65) or inconsistent (M=16.58 95% CI: 13.93-19.18) blocks; $t(18)=1.688$, $p=.109$. In C2, a significant difference was recorded between consistent (M= 20.25, 95% CI: 18.06-22.25) and inconsistent blocks (M= 15.33, 95% CI: 12.80-17.70); $t(17)=2.742$, $p=.014$. Fluency scores (Consistent M= 18.95, 95% CI: 17.04-20.53, Inconsistent M= 15.24, 13.53-16.94) in C3 also differed from one another significantly; $t(24)= 3.519$, $p= .002$.

2.4 Discussion

The current experiment was a pilot study to test the hypothesis that the FAST would be a sensitive measure of the intensity of simple emotional experiences of an associative type. Three experimental conditions differed only in terms of the salience of the US employed to simulate that associative emotional experience. The self-report stimulus evaluation ratings acted as a manipulation check to determine the effectiveness of the evaluative conditioning and to provide an explicit measure of conditioned stimulus salience. The FAST then acted as the implicit measure of stimulus evaluation. It appears that, while the FAST was sensitive to the conditioning contingency, it was not sensitive to the salience of the stimuli employed in that procedure. In other words, the FAST scores recorded for participants did not vary significantly by the salience of the stimuli used during the evaluative conditioning. This

suggests that the FAST is a sensitive measure of evaluations established in an extremely brief procedure simulating an everyday emotional experience (e.g., exposure to a salient news story involving graphic images or aversive words). However, these data do not suggest that the FAST is capable of distinguishing between groups who have been exposed to stimuli of different levels of emotional intensity.

It is important to consider, however, that the exploratory correlational analysis examining the relationship between subjective evaluations of the condition stimuli and the outcome FAST scores suggest otherwise. That is, when FAST scores are considered in terms of how the conditioned stimuli were rated subjectively by participants, rather than on the basis of condition membership, robust and highly significant correlations are observed. In other words, it may be that participants' own pre-experimental response probabilities to the unconditioned stimuli varied sufficiently with that of the sample used in the standardization process for the IAPS images to be more reliable than the arousal and valence scores provided by the producers of that stimulus set.

The significant effect of stimulus function assignment during evaluative conditioning on overall block fluency differences is a cause for concern. This analysis indicated that the FAST effects were more prominent for the participants who received fruit-positive CS-US pairings as opposed to those who received a furniture-positive configuration. This effect likely occurred due to pre-existing evaluative differences between fruit and furniture that were not assessed in this study. In hindsight, this was a methodological weakness of the study that had arisen due to the assumption that any randomly chosen classes of everyday words referring to everyday objects should be more or less equal in valence. However, evidence for a likely pre-existing valence difference across these two-word classes for the current participants is suggested by differences in valence recorded for these stimuli by Warriner et al. (2013). These researchers catalogued the valence, arousal and dominance norms of

14,000 words in the English vernacular along a 9-point scale. For the fruit words used as CSs in the current study, an average valence of 6.71 was established, while an average valence of 5.8 was established for the furniture words. Therefore, the ratings from Warriner et al. (2013) suggest that there was indeed a pre-existing imbalance in the evaluative functions of the CSs employed in Experiment 1. Indeed, recent research has confirmed that stimulus associations are more easily formed between cues and target words when the cue and target word are closer in emotional valence (Buades-Sitjar et al., 2021). This is consistent with behaviour analytic research showing that stimulus equivalence relations are more easily formed amongst stimuli that are discriminable from non-class members, on the basis of shared emotional valence (see Plaud 1995; see also Tyndall et al, 2004). Thus, there is a distinct possibility that these pre-existing evaluative functions competed with the conditioning contingencies for those participants for whom fruit was established as an aversive conditioned stimulus. Of course, this experiment nevertheless succeeded in generating a differential in conditioned stimulus evaluations that was measured successfully by the FAST procedure.

To summarize, the FAST was sensitive to the evaluative conditioning contingencies for the entire cohort considered as a whole. Thus, the basic effect of interest has been established. However, overall, the FAST's sensitivity to the conditioning histories of the participants across the three experimental groups in terms of FAST score magnitude is not clear. Indeed, stimulus control proved to be an issue with the experimental design, however the rate of attrition observed in the current experiment may also be related to the lack of environmental control the researchers were capable of exercising, given the remote data collection strategy. As such, Experiment 2 will seek to establish in principle cross-condition effects, using a larger, sample to prevent and reduce the impact of data loss. The sample in the following experiment will also be offered financial renumeration, in the hopes that this

will increase motivation and therefore improve data quality. As an aside, this experiment will also include a commentary on the impact of participant renumeration.

Chapter 3

Establishing In-Principle Effects with a Larger, Renumerated Sample

Experiment 2

3. Experiment 2

3.1 Introduction

It was concluded at the end of the previous study that the data was of poor quality, due to the lack of experimental control, possibly due to the environment in which the experiment was conducted for most participants (i.e., remotely, unsupervised and unremunerated, although several did receive unconditional course credit for participation). The post-hoc participant exclusion procedure that needed to be employed to remediate the problem of poor quality data resulted in an unacceptably high attrition rate. This experiment was conducted for the purpose of replicating Experiment 1 with a large sample of participants to counter the problem of attrition due to poor adherence to the task. These participants were also recruited through a professional subject participant recruitment service on the assumption that remuneration might increase adherence to the learning task. There is evidence that suggests participants who are paid, and aware that payment is conditional on the basis of data quality, are more likely to produce high quality data (Palan & Schitter, 2017).

Furthermore, despite issuing instructions to participants in Experiment 1 to conduct the task on a desktop device seated comfortably, it became apparent anecdotally that many participants had used mobile devices, although this number could not be confirmed. Because the FAST is a response fluency task requiring full engagement and excellent stimulus control, the quality of data recruited from performances on mobile devices is at the very least suspect. This is not just because of the motor nature of the task but due to the likelihood of social distraction in the types of environments in which people may use such mobile devices. Thus, Experiment 2 involved recruiting a larger number of participants, that were remunerated appropriately for their time. Importantly, participants were also selected on the basis of

primarily using a desktop computer for their research participation via the professional participant recruitment service Prolific, which includes participant device usage habits in the selection criteria for researchers.

It is important at this point to acknowledge that while large sample sizes can help to address natural variability and avert type 1 errors, this approach does not address poor stimulus control inherent in study designs. However, despite the non-equivalent effect of stimulus assignment to the roles of CS+ and CS- observed in Experiment 1, it was deemed unwise to alter any relevant design feature of the experimental procedure in Experiment 2, in the interest of creating a systematic replication and in trying to pinpoint the source of poor performance adherence to the task.

The use of a larger sample in Experiment 2 will also address a concern that has been levied against behavioural research in the implicit testing field regarding low statistical power. Specifically, many studies using the Implicit Relational Assessment Procedure may have adequate sample sizes for simple planned comparisons conducted at the analysis stage. However, they are underpowered to a point of concern when multiple post-hoc analyses are undertaken without Bonferroni correction. This concern deepens when multiple hypotheses are tested on the same small data set (McLoughlin & Roche, 2022).

While it is characteristic of the behavioural tradition to use small sample sizes and compensate for this with higher experimental control, the more frequent use of hypothesis testing in the field, particularly around translational research of the current kind, requires that larger samples are gathered. It is no longer acceptable for low-n statistical rationale to be applied where group designs have been employed. It is also not acceptable to base sample sizes and assess statistical power only in terms of the quantification of main effects, when additional post-hoc analyses could be predictably expected. Thus, Experiment 2 will employ

a large number of participants beyond what was initially envisaged for Experiment 1 in order to account for attrition, and to facilitate ample correlational and post-hoc testing, as may be required.

3.2 Methodology

3.2.1 Participants

A total of 519 participants were recruited on a gender balanced basis as paid volunteers via the Prolific platform. Prolific is a web-based participant pool for online research. Participants affiliated with this site participate in studies in return for monetary compensation. After applying the appropriate exclusion criteria (see Results), the sample consisted of n= 217. The mean age in years for the final sample was 23.72 (SD= 2.81). Females accounted for 52.1% of the sample (n=113), while Males made up 47.5% (n=103). One person (0.5%) identified as non-binary in this dataset.

3.2.2 Procedure

The apparatus and procedure followed within this experiment were identical to those used in Experiment 1, aside from the participant recruitment strategy. However, in contrast to the previous study, the web link via which participants participated was posted only to the professional research participant recruitment service Prolific.

3.3 Results

3.3.1 Excluded Cases and Missing Data

The original sample size collected was n=519. To be included for analysis, participants were required to have completed all stages of the experiment. Several individuals failed to complete all aspects of the study (e.g., failure to respond on a single trial of either block of the FAST, n= 56), and data from these participants were removed from further

analysis. The data from another 14 participants were removed due to a coding error on the InQuisit server, leading to the duplication of participant numbers. A further 62 participants were excluded from analysis due to responding with near chance levels of accuracy (i.e., defined here as >10 incorrect responses per minute on any one block). 169 participants were then removed due to showing no evidence of evaluative conditioning through their stimulus ratings (i.e., the CS+ ratings were equal to or lower than the CS- ratings). This was the same approach used in Experiment 1 and left a final sample size of 217 for analysis.

3.3.2 Descriptive Statistics

Descriptive statistics are presented in Table 3.1 for RFD (FAST) scores, individual block fluency scores, and average evaluative ratings of the appetitive and aversive stimuli. As in the previous chapter, the figures in this section are representative only of numerical trends in the data. They do not indicate the presence of meaningful or statistically significant differences between scores, which themselves are discussed in following sections. For all participants, a conditioned stimulus rating differential score was calculated by subtracting the mean rating for the aversive conditioned stimulus from that for the appetitive conditioned stimulus (where larger ratings indicate more positive evaluations). This provided a stimulus rating difference score (+/-), which indicated the degree of correspondence between conditioned stimulus ratings and the intended evaluative conditioning outcome. Positive rating differential scores indicated differentials in the subjective evaluations of conditioned stimuli in the expected direction (see Table 3.1).

As expected, Condition 1 (C1; lowest valence/highest aversiveness) participants rated the aversive conditioned stimuli the most negatively (M=2.73, 95% CI= 2.41 – 3.07). In addition, this group rated the appetitive conditioned stimuli most positively (M=5.00, 95% CI= 4.66 – 5.31). Also meeting expectations, the least differentiated CS ratings were recorded

for participants in Condition 3 (C3; appetitive; M=4.60, 95% CI= 4.39 – 4.79, aversive; M=4.20, 95% CI= 3.99 – 4.42). Finally, Condition 2 (C2) participants rated the aversive (M= 3.53, 95% CI= 3.20 – 3.86) and appetitive conditioned stimuli (M= 4.88, 95% CI= 4.58 – 5.21) moderately compared to C1 and C3. These ratings suggest that, at the group level, the evaluative conditioning did establish emotional stimulus functions for the conditioned stimuli to varying degrees across conditions, as intended. This trend was additionally reflected in the RFD scores, which "implicitly" indexed the stimulus evaluation differential.

Table 3.1

Means and 95% Confidence Intervals for RFD scores, individual block fluency scores and CS ratings for each condition and the combined cohort.

	C1		**C2**		**C3**		**Cohort**	
N	66		60		91		217	
	M	95% CI	M	95% CI	M	95% CI	M	95% CI
RFD	3.48	2.16 - 4.84	2.55	1.12 – 3.90	.68	-.59 – 2.00	2.05	1.19 – 2.80
ConFluency	21.43	20.34 – 22.49	19.35	18.01 – 20.64	19.20	18.02 – 20.27	19.92	19.22- 20.56
InconFluency	17.95	17.03 – 18.84	16.80	15.61 – 18.01	18.52	17.46 – 19.58	17.87	17.25- 18.49
AvgCSPos	5.00	4.66 – 5.31	4.88	4.58 – 5.21	4.60	4.39 – 4.79	4.80	4.63 - 4.95
AvgCSNeg	2.73	2.41 – 3.07	3.53	3.20 – 3.86	4.20	3.99 – 4.42	3.57	3.39 - 3.77

Note: *C: Condition. RFD: Reaction Fluency Differential (FAST score). ConFluency: correct - incorrect responses per minute on the consistent block, InconFluency: correct - incorrect responses per minute on the inconsistent block. AvgCSPos/Neg: Average Conditioned Stimulus ratings for appetitive/aversive conditioned stimuli.*

C1 participants generally produced the highest RFD scores (M=3.48, 95% CI= 2.16 - 4.84) compared to the other conditions. As expected, the lowest RFD scores were recorded for participants in C3 (M= .68, 95% CI= -.59 – 2.00). The scores recorded for C3 participants are indicative of little or no functional difference between the CS+ and the CS-. As expected,

the RFD scores for participants in C2 fell roughly midway between those recorded for the other two conditions (M= 2.55, 95% CI= 1.12 – 3.90).

These results suggest that the RFD scores varied as a function of the emotional salience of the USs employed during the conditioning phase, and were therefore sensitive to the salience of these stimuli.

3.3.3 Correlations

The relationships between RFD scores and Likert scale ratings for the appetitive and aversive conditioned stimuli were investigated using the Pearson product-moment correlation coefficient (see Table 3.2). The assumptions of normality and linearity were both violated (see Figure 3.1). A medium, positive correlation between RFD scores and average CSPos ratings was found (r= .308, p <.001) as expected. There was also a moderate, inverse relationship between RFD scores and average CSNeg ratings (r= -.348, p<.001). The relationship between ratings for the two conditioned stimulus was moderate and negative (r= -.207, p<.001).

In summary, higher FAST scores were found to be associated with more positive ratings of appetitive conditioned stimuli. Higher FAST scores were also associated with more negative ratings of the aversive conditioned stimuli.

Table 3.2

Correlations table displaying the relationships between RFD scores and average ratings for aversive and appetitive conditioned stimuli.

	1.	**2.**
1.RFD	1.	
2.CSPos	.308**	1.
3.CSNeg	-.348**	-.207**

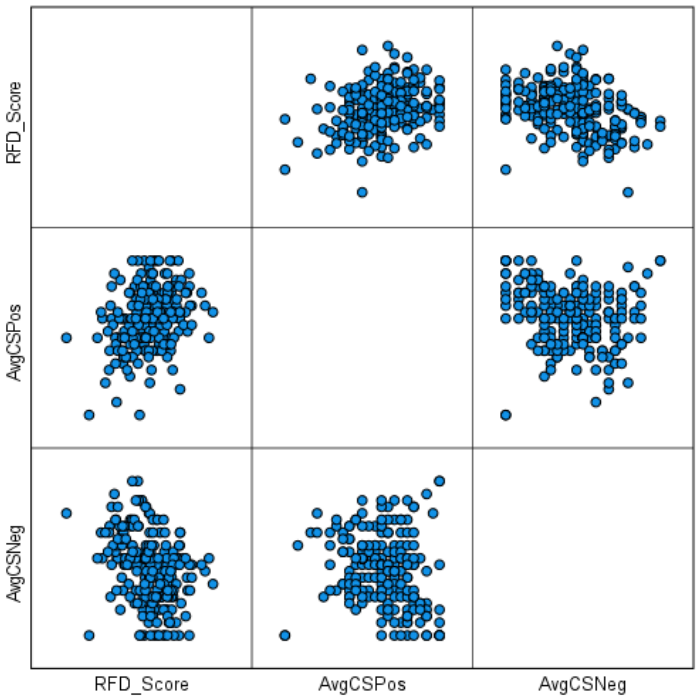
**p<.001

Figure 3.1 below displays a scatterplot matrix with the relationships between RFD scores and average ratings provided for aversive and appetitive CS. As evident in the figure,

the variables do not share a strong linear relationship, as suggested by the violation of assumptions.

Figure 3.1

Scatterplot matrix indicating the relationship between stimulus ratings and RFD scores.



3.3.4 Quantifying the effect of Stimulus Function Assignment on Conditioning

Prior to the main analysis, a 2x2 mixed factorial ANOVA was conducted to assess the potential impact of stimulus function assignment during respondent conditioning (i.e., Fruit-Positive/Furniture-Negative vs Fruit-Negative/Furniture-Positive) on individual block fluency scores. A significant interaction between block and stimulus configuration was found [$F(1,215)= 89.519$, $p <.001$], with a large effect size ($\eta 2= .294$). The main effect for block was not significant [$F(1,215)= 1.159$, $p= .283$], as was the main effect for direction [$F(1,215)= 0.155$, $p=.694$]. These results indicate that stimulus configuration had a significant impact on block fluency scores. Descriptively, the mean fluency scores for each block suggest that there was a reduced effect for those with a furniture positive configuration, and moreover, the effect for this portion of the sample was the reversal of what would be expected given conditioning.

Table 3.3

Table displaying mean consistent and inconsistent block fluency scores for each stimulus function configuration.

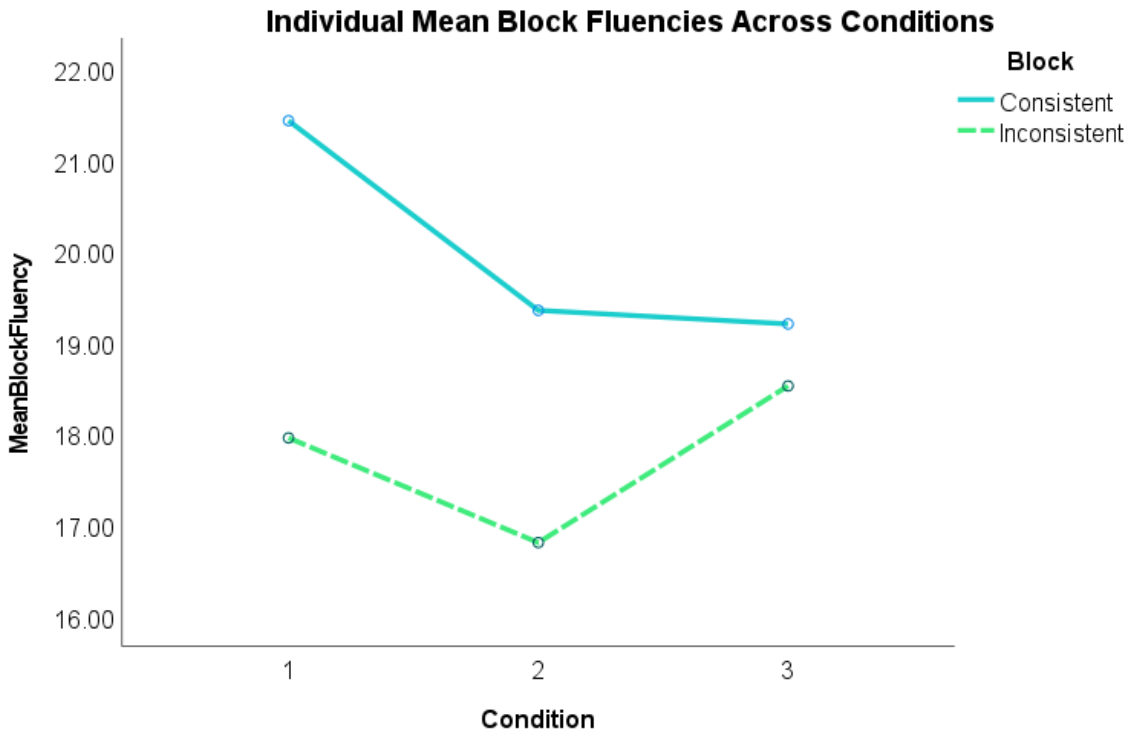
Stimulus Function Assignment	Mean Consistent Fluency	Mean Inconsistent Fluency
Fruit-Positive	20.86	16.80
Furniture-Positive	17.45	20.86

3.3.5 Quantifying FAST Sensitivity to Conditioned Stimulus Valence: Block Fluency Scores

A mixed factorial analysis of variance was conducted to quantify the significance of the observed difference in block fluencies across the three conditions.-Results indicated a significant interaction between block and condition [$F(2,214)= 4.539$, $p= .012$], with a small effect size ($\eta^2= .041$). A main effect for block was also found to be significant [$F(1,214)= 29.761$, $p <.001$], with a moderate effect size ($\eta^2= .122$). This indicates that the FAST was sensitive to a respondently conditioned stimulus relation, as well as the increasing salience of the US stimuli employed in three different versions of that conditioning procedure across conditions. Figure 3.2 shows the noticeable difference in performance across blocks in the expected direction in C1, and that this differential generally diminishes as expected towards C3. The graph also indicated that the overall FAST effect (i.e., fluency differential) is weakening across conditions, irrespective of the absolute magnitude of individual block fluencies.

Figure 3.2

Line graph depicting mean block fluencies across condition

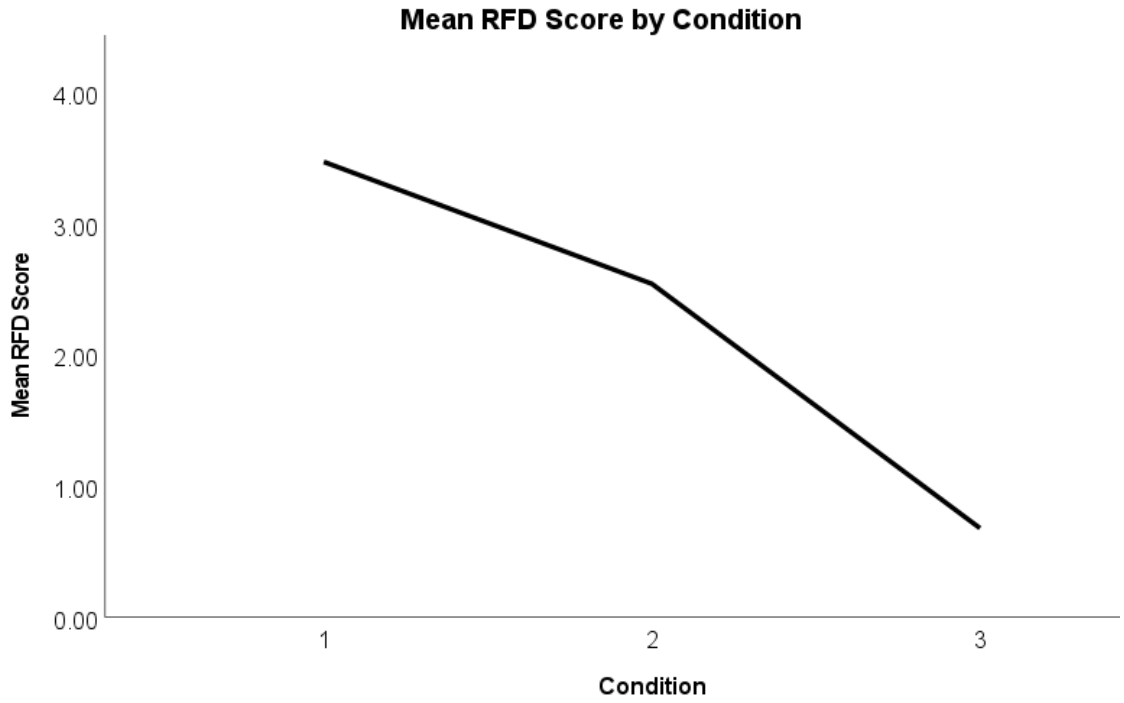


3.3.6 Quantifying FAST Sensitivity to Conditioned Stimulus Valence: RFD Scores

To further assess the decreasing FAST effect across conditions, a one way analysis of variance was conducted to determine which conditions were significantly different in terms of overall RFD scores. As expected given the previous analysis, a significant difference in RFD scores across conditions was identified [$F(2,216)= 4.359$, $p= .012$], with a small effect size ($\eta 2= .041$). Post hoc Tukey HSD tests demonstrated that there was a significant difference in mean RFD scores between C1 (M=3.47, 95% CI: 2.09-4.86) and C3 (M=.68, 95% CI: -.66-2.02); p= .011. Figure 3.3 displays the magnitude of decrease in mean RFD score across conditions as a function of decreasing US salience during conditioning.

Figure 3.3

Line graph depicting mean RFD differences across conditions



3.3.7 Planned Comparisons

To assess whether individual conditions produced statistically significant FAST effects considered in their own right, three paired samples t-tests were conducted, to compare mean fluency scores across the two blocks. For all analyses, the alpha was adjusted to $p<0.017$. Results showed that an overall FAST effect was recorded for C1, with lower levels of fluency observed on the inconsistent (M= 17.95, SD= 3.77) compared to the consistent block (M=21.43, SD= 4.28, t(65)=5.005, p <.001). A similar effect was observed in C2, in which inconsistent fluency scores (M=16.80, SD= 4.96) were significantly lower than those on the consistent block (M=19.35, SD= 5.16); t(59)= 3.597, p <.001). As expected, in C3 there was no significant difference observed across the inconsistent (M=18.52, SD=5.30); and consistent (M=19.20, SD= 5.40) block fluency scores t(90)=1.011, p=.315.

3.4 Discussion

The current study was a replication of Experiment 1, but used a larger and remunerated sample of participants. All methodological features were held constant, as were all data exclusion and analytic methods. Results showed that in the current experiment, the FAST was sensitive to both the conditioning contingency and unconditioned stimulus salience. That is, the FAST was sensitive to the evaluative conditioning established through the initial conditioning procedure. FAST scores also varied significantly as a function of the salience of unconditioned stimuli used throughout the evaluative conditioning procedure (i.e., it was sensitive to experimental condition). Only Condition 1 and 2 participants showed a significant FAST effect when considered alone. The absence of a FAST effect for Condition 3 participants suggests that the conditioned stimuli were not sufficiently functionally distinct in that condition to be differenced by the FAST score, as expected.

It is noteworthy that FAST scores differentiated between conditions in this experiment, whereas they had failed to do so in Experiment 1. In other words, this experiment established the expected effects more clearly then did Experiment 1. This is likely to do with the very healthy sample size and increased statistical power in Experiment 2. To assess the degree of statistical power necessary to detect main effects with the current sample size, a post-hoc power analysis was considered. However, calculating post-hoc observed power is a relatively meaningless exercise that inevitably leads to near 100% power achievement where effects are significant with low p values (see Laekens, 2022). A more meaningful practice may be to calculate the power that would be achieved, given the sample size, for various levels of effect size that might be observed in the design employed here (i.e., a sensitivity analysis). In order to conduct such an analysis, firstly, the $\eta 2$ effect values observed for the main analysis (main effects and interaction) was converted to Cohen's f effect sizes (using $f = \sqrt{\eta 2 / (1 - \eta 2)}$). This allowed for the calculation of required power for given *f* effect sizes using the

software package G*Power (v. 3.1.9.6). This calculation was conducted for the Cohen's *f* effect size statistics of 0.2 (small) and 0.4 (borderline large) for the interaction and main effects, respectively, according to Cohen (1988; see also Rosenthal, 1996). The analysis indicated that, given the sample size and the above effect sizes, we would have 100% and 49% power to detect the respective effect sizes. Given the power associated with the analyses, we can conclude with reasonable confidence that our expectation that the FAST would be a sensitive measure of conditioned stimulus functions was met. That is, FAST block fluency scores differed across blocks for the cohort as a whole, and the magnitude of this difference varied as a function of unconditioned stimulus salience during the initial respondent conditioning learning experience. FAST (RFD) scores also correlated with subjective conditioned stimulus ratings, as expected.

The observed correlation between participants' conditioned stimulus ratings and overall FAST effects were somewhat weaker than Experiment 1, though in both cases relationships were significant. This may be partly due to the fact that in this experiment, assumptions of normality and linearity in the stimulus rating data set were violated, likely due to uncontrolled variance in the data and the unreliability of the rating measure. While this might not be expected to have a very noticeable difference on correlation coefficients or the significance of correlations (see Havlieck et al., 1997), it is worth remembering that effect of abnormal data distributions on correlation test outcomes increases with increasing sample size (see Brown & Lathrop, 1971). Indeed, the sample size in the current study is considerably larger than that used to demonstrate this very principle in the original research reported by Brown and Lathrop (1971). Thus, it would be unwise to speculate very much on the meaning of any differences in the strengths of the correlations observed across the two experiments.

It is apparent that the steadiest decrease in block fluency scores across conditions as unconditioned stimulus salience decreased, was observed for performances on the consistent block. In fact, fluency scores increased somewhat from Condition 2 to Condition 3 on the inconsistent block, even while fluency differentials across the blocks decreased within conditions. In effect, the decrease in RFD scores across conditions seems to be driven by a steady decrease in performance on the consistent block, and relatively unchanging inconsistent block performances. The stable decrease in consistent block performance fluency across conditions suggests that the consistent block was becoming more "difficult" with decreasing stimulus salience, rather than the inconsistent block becoming easier. In other words, the main driver of the cross-condition effect appears to be the ease with which functional response classes are established involving compatible stimuli, an effect that appears to increase with conditioned stimulus salience. While increasing stimulus salience also generally decreases the acquisition rates of functional response classes consisting of incompatible stimuli (i.e., performance on the inconsistent block), this effect is less linear and apparent.

The foregoing observation parallels that made by Cummins et al. (2018), who found that fluency scores on the consistent block increased with increasing stimulus relatedness (experimentally controlled). In contrast, fluency scores observed on the inconsistent block of the test were more varied with varying controlled stimulus relatedness. Thus, the increasing magnitude of difference in fluency scores across the two test blocks and across conditions is accounted for more by increases in fluency on the consistent block than decreases on the inconsistent block. This observation may provide some pointers to implicit test researchers regarding the nature of the phenomenon being measured by these tests, namely, that consistent block performances may be more sensitive to stimulus evaluations than are inconsistent block performances. The functional approach to implicit testing has already

suggested that these tests might be simply considered to be measures of stimulus relatedness within classes and stimulus unrelatedness across classes (Roche et al., 2005; Cummins et al., 2018). However, the current data now suggest that these test scores index stimulus compatibilities to a greater extent than stimulus incompatibilities.

The post-hoc planned comparisons showed that consistently, participants had higher levels of performance fluency on the consistent block (i.e. responding stimuli according to the conditioned stimulus relations) than on the inconsistent block. However, within condition effects were only observed for Condition 1 and Condition 2, as might have been expected. That is, considered alone, these two conditions managed to generate a response fluency differential across blocks which is indicative of an established "implicit" test effect based on conditioned stimulus relations alone. This is an important achievement for the FAST procedure, as the effects of such relations on implicit test outcomes has not yet been conducted within the behavioural field and has been explored only minimally in the social cognitive literature. Specifically, using an evaluative conditioning procedure similar to the current study, Olson, and Fazio (2001) showed that the CS+ was evaluated more positively than CS- on an implicit measure (IAT). The conditioning procedure employed by Olson and Fazio (2001) was more intensive, involving 40 CS-US pairings presented compared to 32 presentations in the current study. Following conditioning, explicit CS evaluative ratings were recorded and an IAT employing the CS was administered. It is notable that both the implicit (IAT) effects and the implicit-explicit correlations observed in Olsen and Fazio (2001) were weaker than what was recorded in the present study. While this outcome might lead one to conclude that the FAST is more sensitive a measure than the IAT in the assessment of evaluatively conditioned relations, such a conclusion would be premature given procedural differences across these studies, as well as differences in scoring metrics used within the two different procedures.

An interesting interpretation of the current findings may come from considering the results of the real social group condition reported by Van Dessel et al. (2015; see Introduction). This study attempted to alter stimulus functions using an approach and avoidance task designed to establish general approach (appetitive) and general avoidance (aversive) functions for two arbitrary classes of stimuli, counterbalanced across participant cohorts. While the IAT was successful in detecting laboratory manipulated stimulus relations where the conditioned stimuli were fictional social classes, it was unsuccessful where they consisted of real, valenced social classes of black and white individuals. Interestingly, however, ratings of the black and white stimulus categories following the avoidance and approach procedure were indicative of an impending IAT effect. That is, they appeared to reflect a successful conditioning procedure. However, the IAT was not sensitive to efforts by the researchers to manipulate racial bias using the avoidance and approach procedure. One explanation for this finding was that the pre-existing implicit racial bias against Black people was so strong that the expected IAT effect was not present, even for those for whom Approach functions were established for Black person exemplars. Read this way, the IAT may be viewed as particularly useful for assessing entrenched biases even following subjective verbal reports of changes in attitude. At the same time, however, it is important to note that this study did not advance our knowledge of the sensitivity of the implicit tests to conditioned relations. In the current study, however, the functions of the arbitrary stimulus classes chosen as CS were clearly and successfully manipulated. More importantly, while pre-existing stimulus functions had a confounding effect in both the Van Dessel et al., (2015) and the current Experiment 2, this experiment recorded test outcomes that were a function of the laboratory conditioning procedure, despite the challenge of pre-existing stimulus function confounds. Future studies on the IAT and FAST should more carefully control pre-existing stimulus functions, and more carefully assess the outcomes of conditioning using a more

varied set of measures not limited to subjective ratings and reports but perhaps including behavioural and physiological measures of appetitive and aversive stimulus functions.

A limitation of the current experiment was the disappointingly high rate of participant attrition due to poor data quality, despite the use of renumeration. This outcome was surprising because Prolific participants are aware that payment is contingent upon the data they provide being of an acceptable standard to the experimenter (i.e., payment can be withheld for poor task adherence, although this was not done in this study). Importantly, although a filter had been applied to select volunteer participants who indicated that their primary digital device for research participation was a desktop computer, this could not be enforced. In addition, due to a technical error, operating systems of users were not tracked on the Inquisit server, therefore preventing the exclusion of participants who had used a mobile device. Thus, it would be reasonable to conclude that even amongst participants who remained within the final participant cohort for analysis, many may have used mobile devices and have been distracted while completing the study. In future, researchers should ensure and enforce the use of desktop devices and the conduct of the study in quiet appropriate settings. Nevertheless, the extra noise that was potentially added to the current data due to less than perfect attention to the task may well have been more than offset by the larger sample size and increase statistical power in terms of identifying an in-principle effect.

Of course, the ultimate aim of research of this kind is to achieve a level of experimental control, robustness, and clarity of effect that it can be replicated reliably with small samples. While group level effects are interesting scientifically, they are not sufficiently satisfactory to the behaviour analyst to allow conclusions to be drawn that can be directly linked to behavioural principles. This study forms part of the ongoing ground-up research effort to develop an implicit-style test based on behavioural principles, in which all test outcomes can be easily traced back to those same principles at an individual or a small

sample level. Thus, Experiment 3 will consider a return to an in-person laboratory setting in order to gain greater control over these observed effects. Instead of dealing with the problem of poor adherence by increasing sample size, it may be worth exploring controlling for this with a more formal research participation context and more in-person monitoring of the performance of individuals under laboratory conditions.

Chapter 4

A Return to Behaviour-Analytic Roots: Reproducing In-Principle Findings with a Smaller, In -Person Sample

Experiment 3

4. Experiment 3

4.1 Introduction

As discussed in Chapter 3, collecting data remotely is advantageous in that a larger and more diverse sample size can be achieved with more ease and in less time (Palan & Schitter, 2017). This approach has the added benefit of convenience, in that scheduling conflicts and organising participants to come into the laboratory are not a concern. However, despite these benefits, doubts remain as to the quality of online data. Online data quality can be affected by the software used (de Leeuw & Motz, 2016). Some research indicates that online participants provide high quality data when they do not encounter distractions or technological failures (Reinecke & Gajos, 2015). However, these are unavoidable at times and can interfere with data quality (Reinecke & Gajos, 2015; Reimers & Stewart, 2008). These extenuating factors can be addressed by employing procedural attention checks and data cleaning techniques (Anwyl-Irvine et al., 2021; Gough et al., 2012), although these solutions necessitate the introduction of additional tasks into the midst of the experimental procedure, and/or the elimination of participants according to relatively arbitrary criteria. In other words, it would be preferable if no data at all had to be deleted from the data set and sufficient experimental control was exerted over the behaviour of participants that all data produced was of use and interest to the researcher.

For the previous two experiments (conducted remotely), data cleaning was necessitated due to various issues with data quality, relating to technological unsuitability (i.e., using inappropriate devices) and suspected inattention to the task (i.e., responding at below chance levels of accuracy). As a final attempt to enforce adherence to the task, it was reasoned that the experimental procedure should return to the experimental context in which all of the foundational principles for the FAST were developed, namely to individual

participation in experimental procedures in a supervised laboratory setting. In essence, such procedures were assumed to enhance "demand effects" (Orne, 1962). That is, the knowledge that the experimenter was on the other side of the laboratory cubicle room door was expected to motivate participants to attend and respond to the task to the best of their ability. This phenomenon is widely documented; and described by Orne (1962) as the participant's eagerness to be a "good" participant. Demand effects have been outlined as a causal factor for participants of experiments completing seemingly meaningless tasks for extended periods with a high level of performance. Moreover, Weber and Cook (1972) outlined that when a participant is not made aware of the hypothesis of the study, they are more likely to follow instructions given by the experimenter. Given that participants were not informed of the hypothesis of the current study until after their participation was complete, it was expected that the repeated instructions to pay continuous attention to the task would be adhered to in the current study.

This study was conducted as a systematic replication of Experiments 1 and 2, with the important difference that data were collected from in-person participants who were under close laboratory supervision. This was done to determine whether or not clearer experimental effects would be observed when a smaller number of participants are exposed to the experimental procedures under close supervision in a traditional University research laboratory setting, compared to when participants are recruited online and offered remuneration in an unsupervised setting.

It was also expected that, as a corollary of increased attention to the learning task, attrition rates would be lowered and the need for the exclusion of data sets would approach zero. It was hoped that if these outcomes were met, the data recorded from each participant would be easier to generalize to the rest of the population and be a fair representation of the phenomena under analysis.

4.2 Methodology

4.2.1 Participants

A total of 87 participants were recruited for this study. All participants were recruited as volunteers from the student body of Maynooth University. After applying the exclusion criteria (see Results), and the removal of one participant due to a fault on the Inquisit server, a total of 56 participants remained for analysis (age in years M=20.89, SD=6.80). Of these, 35 identified as female, 19 as male and 2 as non-binary/other. Participants were not renumerated in this experiment; however, a small portion of the final sample (n=8) did receive unconditional course credit for their cooperation.

4.2.2 Procedure

The apparatus and procedure in this study were identical to Experiment 1, the only differences being the method of recruitment used and the location in which the study was conducted. Recruitment involved a snowball sampling and direct approach method. Willing participants sat comfortably at a desktop computer in a small laboratory cubicle in the MU Psychology Department. The Inquisit study link was set up prior to participant arrival so that once participants were briefed on the experiment, they were able to begin.

Subjects sat comfortably in a chair in a 2x2m enclosed laboratory cubicle, at a standard computer desk. The computer screen was approx.70cm from the participants' face and was set at eye level. On the desk was a 15inch monitor, with a mouse and full keyboard. Participants were brought to the cubicle and were oriented to the computer and relevant operanda. The broad instructions for the experiment as well as what participants could expect were verbally dictated to participants, after which the chance to ask any questions was offered. Concluding this, participant attention was directed toward the screen where the official information page (see Appendix II) was displayed. The participant was then informed

that to continue onto the task they would have to indicate their consent to participating in the study by supplying their age and other basic demographic details (see Appendix IV). At this point, the experimenter left the cubicle and shut the door. They remained outside the door in the case of any questions or technical failures that required intervention. The participant, given no questions, was then left alone to complete the experimental procedure and was instructed to seek the experimenter upon completion, whereby they were subsequently debriefed and again offered the opportunity to ask any questions relating to the experiment.

4.3 Results

4.3.1 Excluded Cases and Missing Data

The original sample collected consisted of 87 participants. One participant had to be excluded due to an error on the InQuisit server which caused the experiment script execution to fail. As in previous experiments, to be included in the analysis, participants were required to have completed all phases. No participants failed to engage with any one stage of the experiment. In keeping with the previous studies, participants were also excluded for responding with near chance levels of accuracy (i.e., defined as >10 incorrect responses per minute on any one block, n=15). Finally, data for participants who failed to show evidence of conditioning according to the contingencies of their evaluative conditioning (Calculated in terms of a rating differential score as outlined in section x of Experiment 1) were also excluded (n= 15). The final sample size was 56, 8 of which participated for course credit.

4.3.2 Descriptive Statistics

Means and confidence intervals are presented in Table 4.1 for RFD (FAST) score, fluency scores for consistent and inconsistent blocks and average subjective ratings of appetitive and aversive conditioned stimuli, which followed the expected trend. That is, the average ratings recorded from Condition 1 (C1) participants for appetitive conditioned

stimuli were most positive (M= 5.19, 95% CI: 4.76 – 5.57). Those recorded from Condition 3 (C3) were least positive (M= 4.63, 95% CI: 4.21 – 5.06), while participants in Condition 2 (C2) rated the appetitive stimuli moderately (M= 5.13, 95% CI: 4.68 – 5.51). The cross-condition difference in ratings was relatively small, albeit in the predicted direction. With regards to the aversive conditioned stimulus ratings, the expected trend was not observed. That is, while C1 participants rated the aversive stimuli most negatively as expected (M= 2.42, 95% CI: 2.05 – 2.77), those in C2 rated them moderately (M= 2.53, 95% CI: 2.06 – 3.08), while the lowest aversiveness ratings were provided by C participants (M= 3.23, 95% CI: 2.82 – 3.16). These ratings suggest that at both group and condition level, the evaluative conditioning phase effectively established differential evaluative functions across the two conditioned stimulus sets. Again, it is important to remember that these differences, and those discussed throughout this section, are merely descriptive of numerical trends in the data, and should not be taken to represent any statistically significant differences, which themselves will be examined below.

Descriptive statistical analysis of FAST scores did not indicate the expected trend of decreasing scores in line with increasing image valence. More specifically, for C1, a mean RFD score of 1.89 (95% CI: -.90 – 4.64) was recorded. As expected, a much lower score was recorded for participants in C3 (M= .24, 95% CI: -2.16 – 2.32). In C2, however, the mean RFD score recorded was not intermediate, as expected, but rather the largest observed across the three conditions at 4.34 (95% CI: .44 – 7.96). Closer inspection showed an abnormally large confidence interval for the mean RFD score recorded in C2.

Given the wide confidence interval observed for C2 RFD scores, the data set was inspected for outliers, in an attempt to understand the degree of noise in this data set compared to the others. It was found that the five highest RFD scores recorded were for C2 participants. One C2 participant had an RFD score of 22.59 (2.97 SD above mean). In

comparison to the cohort mean, this case is an obvious outlier and worthy of consideration when assessing these descriptive statistics. When this single extreme outlier is removed, the mean RFD score for C2 is reduced to 3.26 (95% CI: -.32 – 6.69). However, even then, the mean C2 RFD score does not follow in the expected trend, suggesting that in this instance, RFD scores did not vary as a function of the salience of USs used within the conditioning phase.

Table 4.1

Means and 95% Confidence Intervals for RFD scores, individual block fluency scores and average CS ratings for each condition and the combined cohort.

	C1		C2		C3		Cohort	
N	21		18		17		56	
	M	95% CI	M	95% CI	M	95% CI	M	95% CI
RFD	1.89	-.90 – 4.64	4.34	.44 – 7.96	.24	-2.16 – 2.32	2.17	.46 – 3.94
Con Fluency	20.01	17.84 – 21.97	18.10	15.39 – 20.94	17.60	15.36 – 19.77	18.67	17.14 – 20.07
Incon Fluency	18.12	15.91 – 20.31	13.77	11.15 – 16.40	17.34	15.27 – 19.44	16.49	15.05 – 17.96
AvgCSPos	5.19	4.76 – 5.57	5.13	4.68 – 5.51	4.63	4.21 – 5.06	5.00	4.76 – 5.23
AvgCSNeg	2.42	2.05 – 2.77	2.53	2.06 – 3.08	3.23	2.82 – 3.16	2.70	2.46 – 2.97

Note: *C: Condition. RFD: Reaction Fluency Differential (FAST score). ConFluency: correct - incorrect responses per minute on the consistent block, InconFluency: correct - incorrect responses per minute on the inconsistent block. AvgCSPos/Neg: Average Conditioned Stimulus ratings for appetitive/aversive conditioned stimuli.*

4.3.3 Correlations

A Pearson-product moment correlational analysis was carried out to assess the extent to which participants' RFD scores were representative of their recorded subjective ratings of aversive and appetitive conditioned stimuli. The assumptions of homoscedasticity and linearity were violated for this analysis. The only statistically significant relationship was the expected inverse relationship between average CSPos and CSNeg stimulus ratings (r= -.368, p =.005). The correlations between RFD and CSPos, and RFD and CSNeg were not

significant. The relationship between RFD score and the average CSPos ratings were in the expected direction, but were not significant. The relationship between RFD score and average CSNeg was not significant or in the expected direction.

Table 4.2

Correlations table displaying relationships between RFD scores and average ratings for aversive and appetitively conditioned stimuli.

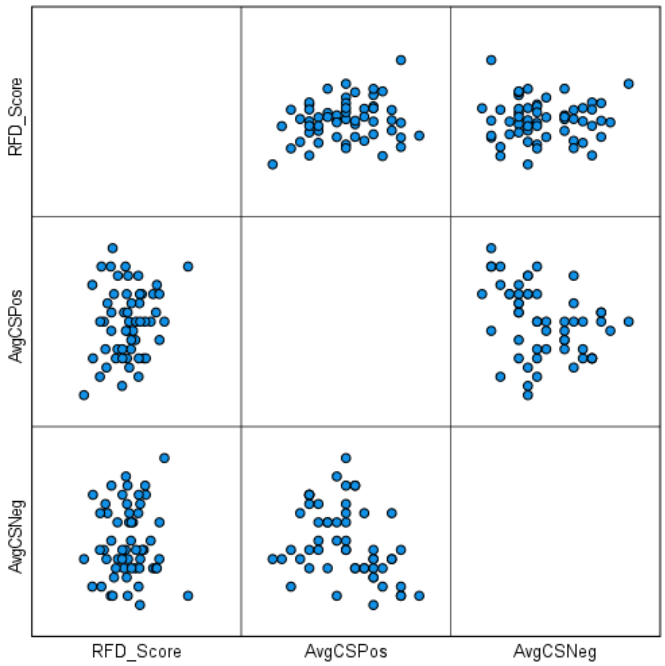
	1.	2.
1.RFD		
2.CSPos	.201	
3.CSNeg	.016	-.368**

** p <.05

Figure 4.1 displays the relationship between the variables included in the correlational analysis. The graph supports the suggestion from the correlational analysis, and the violation of assumptions, that the relationships between the variables are weak and non-linear.

Figure 4.1

Scatterplot matrix indicating the relationship between stimulus ratings and RFD scores.



4.3.4 Quantifying the Effect of Stimulus Function Assignment on Conditioning

A 2x2 mixed factorial ANOVA was conducted to evaluate the impact of stimulus function assignment (i.e., whether fruit or furniture was established as the aversive conditioned stimuli or vice versa) on block fluency scores. A significant interaction was found between block fluency scores and the stimulus function assignment [$F(1,54)=15.576$, $p <.001$], with a large effect size ($\eta 2= .224$). There was also a significant main effect found for block [$F(1,54)= 4.971$, $p= .03$] with a small effect size ($\eta 2 = .084$). This result indicates a successfully created FAST effect at the group level, irrespective of stimulus function assignment, but that the effect trended lower for those who received a furniture-positive CS-US pairing during evaluative conditioning (see Table 4.3)

Table 4.3

Table displaying the mean consistent and inconsistent block fluency scores for each stimulus function configuration.

Stimulus Function Assignment	Mean Consistent Fluency	Mean Inconsistent Fluency
Fruit-Positive	20.18	15.12
Furniture-Positive	16.79	18.20

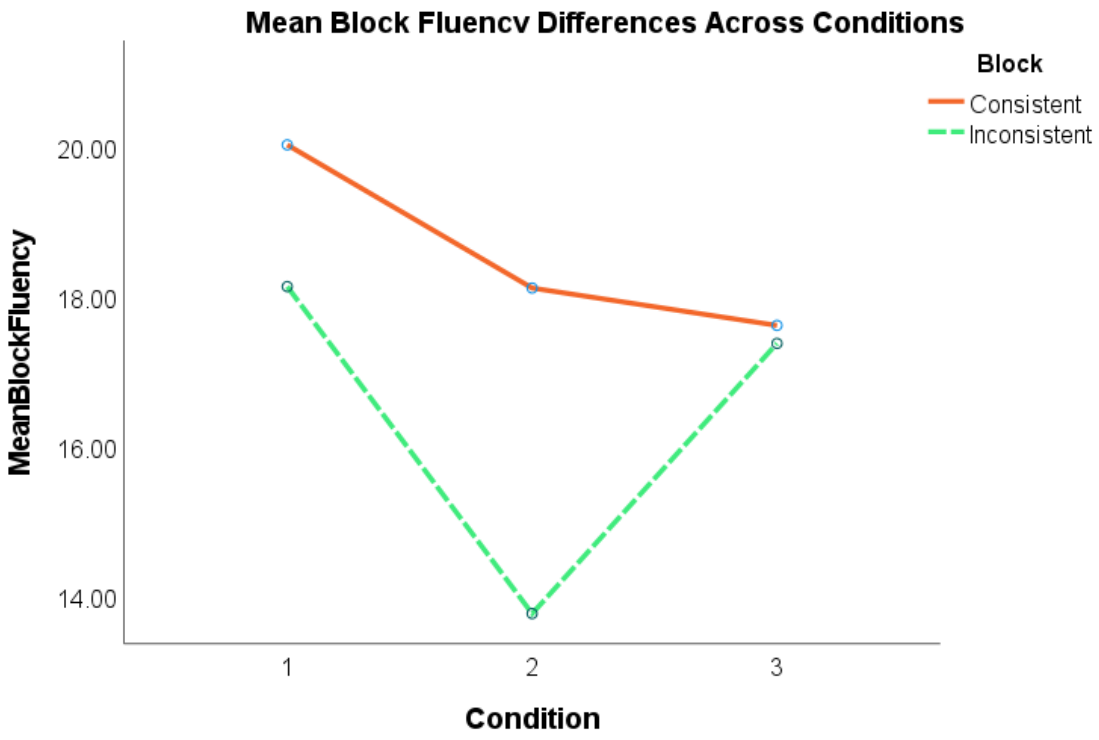
4.3.5 Quantifying FAST Sensitivity to Conditioned Stimulus Valence: Block Fluency Scores

A mixed factorial analysis of variance was conducted to explore the impact of condition on individual block fluency score differences. The interaction between block and condition was not significant [$F(2,53)= 1.624$, $p= .207$]. However, there was a significant main effect for block [$F(1,53)= 5.604$, $p= .022$], with a small effect size ($\eta 2= .096$).

These results indicate that for the combined participant sample, there was an overall difference in fluency from the consistent to the inconsistent block. However, this difference was not easily apparent when conditions were considered individually. Figure 4.2 illustrates the estimated marginal means of the individual block fluencies for each condition and depicts the steady decline in fluency from C1 to C3 on the consistent block, as expected. The inconsistent block fluency decreases from Condition 1 to 2, but increases from Condition 2 to 3.

Figure 4.2

Line Graph depicting mean differences in block fluencies across conditions.



4.3.5 Quantifying FAST Sensitivity to Conditioned Stimulus Valence: RFD Scores

A one way analysis of variance was conducted to assess whether the magnitude of change in RFD scores across conditions was statistically significant. The results showed that

there were no significant differences between any of the groups [$F(2,53)= 1.624$, p= .207]. These findings thereby indicate that in this instance, the FAST was not a sensitive measure of varied stimulus salience.

4.3.6 Planned Comparisons

Planned comparisons were conducted to assess whether significant FAST effects (i.e., response fluency differentials) were observed within any of the conditions considered separately. Three separate paired samples t-tests were conducted to determine this. The Bonferroni adjustment was applied such that only alphas of p> .017 were accepted as significant. There were no significant differences in consistent and inconsistent block fluencies in C1; t(20)= 1.363, p= .188, C2; t(17)= 2.127, p= .048 or C3; t(16)= .208, p= .838.

Separately, to assess whether conditioned appetitive stimulus ratings were significantly different across stimulus function assignment, an independent samples t-test was conducted. Results showed no significant differences in ratings of the appetitive conditioned stimulus across to the two cohorts of participants who were exposed to the different conditioning configuration: t(54)= 1.031, p=.307. The same outcome arose for the aversive stimulus ratings; t(54)=.619, p=.538.

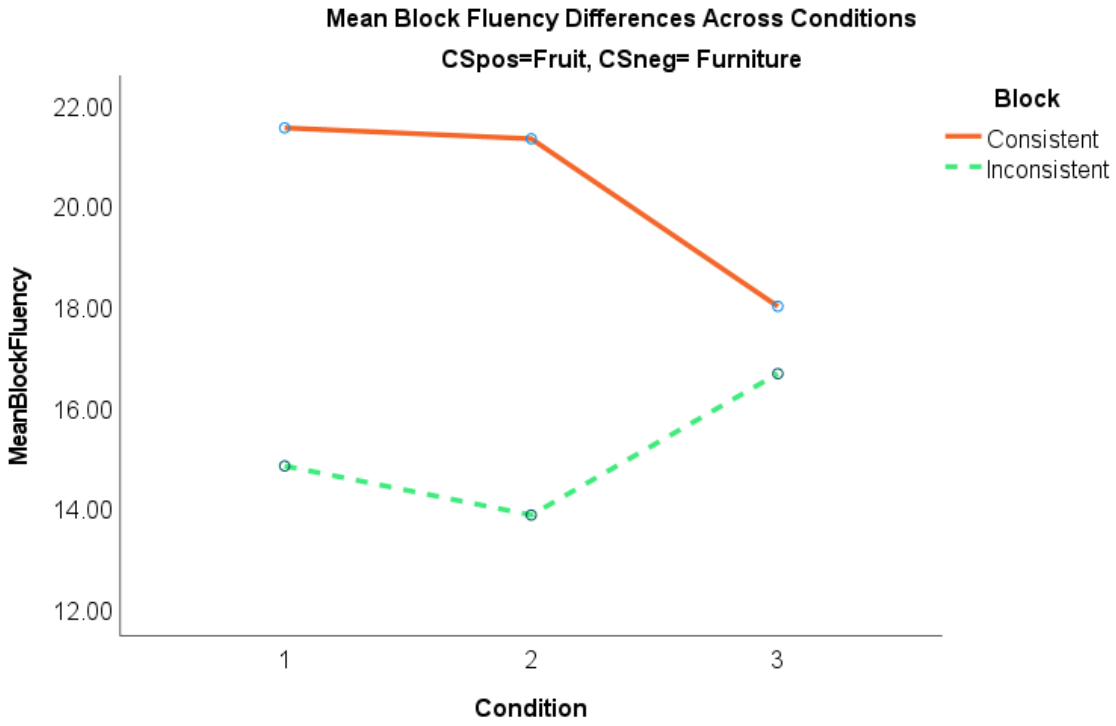
4.3.7 Post Hoc Analyses

Given the interaction between stimulus configuration and block fluency scores, it is clear that the conditioned stimuli had pre-existing functions that confounded the conditioning efforts, and also the test score outcomes. It was therefore decided to conduct the main analyses again separately for each group of participants, separated on the basis of which stimulus function assignment configuration was employed. Although this analysis would not indicate the sources of any differing pre-existing stimulus functions, it would demonstrate the impact these hypothesised pre-existing relations had on fluency scores across blocks.

This analysis showed that for participants who received a fruit-positive and furniture-negative CS-US pairing, there was a significant interaction effect between block and condition [F(2,28)=3.445, p= .046], with a large effect size (ɳ2= .197). Similarly, the main effect for block was significant [F(1,28)= 23.493, p< .001], and this had a very large effect size (ɳ2= .456; see Figure 4.2).

Figure 4.3

Line Graph depicting mean differences in block fluencies across conditions for participants receiving Fruit-Positive Furniture-Negative CS-US pairings.

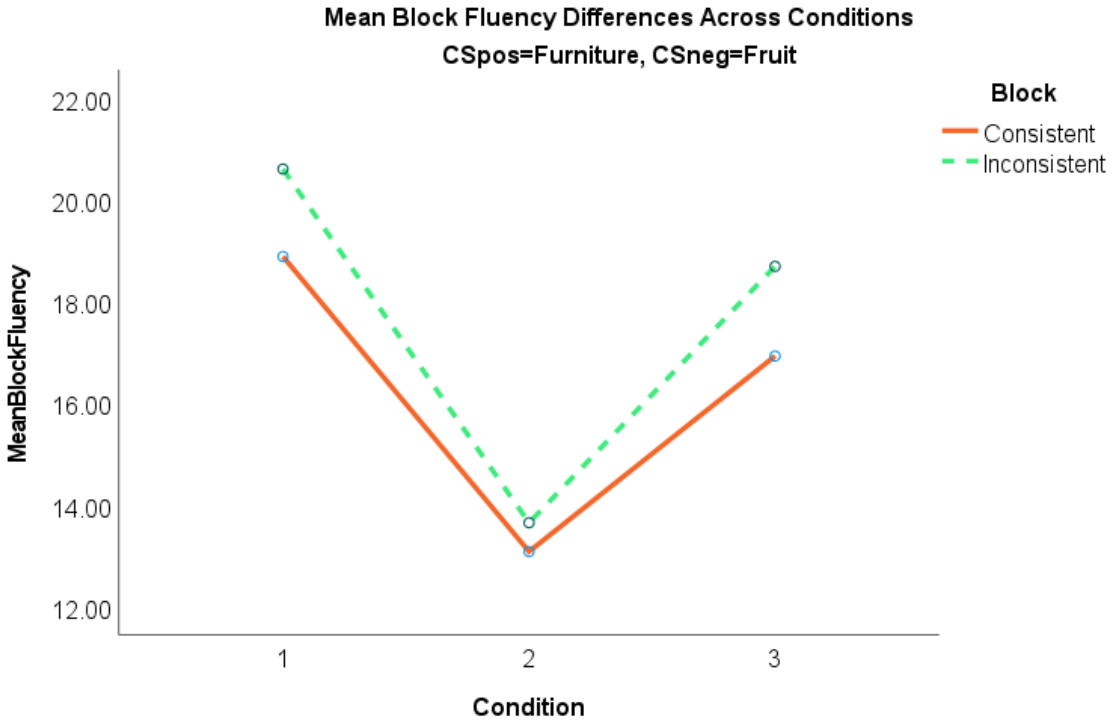


Assessing the portion of the sample who received a furniture-positive and fruit-negative stimulus function assignment, there was no significant interaction effect found between block and condition [F(2,22)= .096, p= .908]. In addition, there was no main effect found for block [F(1,22)= 1.168, p= .291]. The line graph for this group (see Figure 4.3) clearly demonstrates a block fluency pattern in the reverse of what may have been expected given the conditioning procedure. This outcome clearly suggests that the expected effects were not at all apparent for those participants for whom furniture was established as a

positive affective stimulus, even though subjective ratings corresponded to this conditioning contingency.

Figure 4.4

Line Graph depicting mean differences in block fluencies across conditions for participants receiving Furniture-Positive and Fruit-Negative CS-US pairings.



4.4 Discussion

This study was conducted for the purpose of replicating the previous two experiments with an in-person sample. It was expected that due to higher levels of experimental control, results from this study would be more reliable than the online iterations, in more accurately quantifying any true effect. Moreover, it was anticipated that fewer participant data sets would need to be excluded from analysis due to poor task adherence, technological failure, or inappropriate software usage. This was in fact not the case. Although no participants were excluded due to inappropriate device usage, or for failure to engage with any one task of the

experiment, the proportion of participants eliminated for poor task adherence was higher than was required in Experiments 1 or 2. This will be discussed further in the General Discussion.

It is noteworthy that there was an extreme outlier in the present data set, that was retained for all analyses. While standard practice generally recommends treating outliers with data cleaning methods (i.e., removal, replacing the outlier with the sample mean, or with 'possible values' Cousineau & Chartier, 2010), it was determined that no such techniques would be employed in the current study, in the interest of retaining behavioural variability. More specifically, from the behaviour-analytic perspective, behavioural variance is the very phenomenon we are trying to understand and study (Sidman 1960; Skinner 1976). Thus, active attempts to reduce or indeed eliminate these differences is at the least questionable. Furthermore, it is quite possible that inter-individual score variability is associated with varying abilities to complete the FAST test itself. That is, some individuals may be able to respond or learn more quickly than others. Indeed, this has been observed in previous research (c.f. O'Reilly et al. 2013). In any case, the variability in this sample can be typical of smaller datasets, which further highlights the benefits of larger sample sizes, such as that collected for Experiment 2.

One analysis that may have been notably compromised as a result of the large variance in scores was the correlational analysis investigating the relationships between the RFD score, and the average appetitive and aversive conditioned stimulus ratings. The observed correlations were weaker than expected, given the very strong relationship observed in the previous experiments between RFD scores and stimulus ratings. It is important to note, however, that the violation of the assumptions of linearity and homoscedasticity were documented during the analysis and were likely the result of outliers included in the data set. Importantly, however, the sample size was also relatively small.

The results of the repeated measures analysis of variance were similar to those found in Experiment 1, in that there was no interaction between the variables found. This suggests that the FAST was not sensitive to the varying degrees of CS salience as established by the evaluative conditioning procedure. However, a significant main effect for test block was observed here as in the previous two experiments, demonstrating that while the FAST failed to provide an index of conditioned stimulus salience, it was sensitive to the evaluative functions established for conditioned stimuli.

A clear linear decrease in consistent block fluency is evident in the main analysis of variance which compared mean block fluency scores across conditions, (see Figure 4.3). That is, fluency decreases as a function of CS valence as expected. Scores on the inconsistent block, however, do not demonstrate the same trend. In similar fashion to what was observed in Experiment 2, once again it is the consistent block fluency score that seems to be more sensitive to the changing stimulus valence during the conditioning phase. This is an important observation, as it is now apparent across all three experiments of this thesis. This finding gives further credence to the idea that stimulus relations may be, at least in principle, assessed by examining consistent block fluencies alone (see General Discussion).

The absence of an expected decrease in the magnitude of overall fluency differential (RFD) scores in line with stimulus salience is quite possibly related to the pre-existing functions of the designated conditioned stimuli. That is, the assigned function of the conditioned stimuli as appetitive or aversive evidently confounded the overall effects observed, although not enough to completely eliminate them. That is, the conditioning procedure successfully produced a cohort-level FAST effect, irrespective of the function assignment to each conditioned stimulus. In addition, it should be remembered that only participants who rated conditioned stimuli in line with their conditioning contingencies were included in the final participant cohort. Thus, despite any reduced conditioning effects for

one large portion of participants for whom the experimenter attempted to establish fruit as an aversive stimulus and furniture as an appetitive stimulus class, the basic effect of conditioned evaluative functions on FAST outcomes has been observed.

That observation notwithstanding, such a variance in the effect sizes that might be observed across participants for whom differing evaluative functions were established for each conditioned stimulus class, could easily overshadow the variance across conditions in block fluency differentials that arose as a result of the salience of the unconditioned stimuli. In effect, the simultaneous efforts of increasing US salience *and* unwittingly altering conditioned stimulus' pre-existing salience may have cancelled each other out in this case, at least in terms of inferential statistical effect significance. Although the researchers were aware of the stimulus control issue by the end of Experiment 1, it was deemed more important to retain the same stimuli in order to achieve a systematic reproduction, than to increase stimulus control by reassigning conditioned stimulus exemplars. As discussed in Chapter 2, the population-normed valences calculated for the fruit and furniture exemplars (c.f. Warriner et al., 2013) employed in the experimental procedure were not equal, which further suggests that pre-existing CS functions likely facilitated or impeded the conditioning procedure used. It is very difficult to speculate on how real-world stimulus functions (i.e., valence or salience) might affect performance in a learning task like the FAST in the absence of good stimulus control, but resources like that developed by Warriner et al. (2013) can help to narrow the range of socially established influence functions for laboratory stimuli.

Another issue encountered in the current experiment was low adherence to the task or general experimental / laboratory control. This experiment reproduced the procedures of the previous experiments with an in-person sample under highly supervised laboratory conditions, for the sole purpose of producing more valid and reliable findings. In ways, this aim was met; every participant completed the experiment in an identical environment, using

identical devices. Attentiveness was also improved to some extent insofar as no participant was eliminated from the sample as a result of not having engaged at all with any one phase of the experiment. However, it remained necessary to eliminate a small proportion of participants due to responding at chance levels of accuracy, which of course indicates an unacceptable level of inattentiveness. Thus, while the expected demand characteristics did prevent some participant attrition, the current experiment was not successful in completely removing the need to eliminate participants (see General Discussion).

The lack of adherence to the task observed for so many participants may not be surprising when one considers that the majority of participants were not remunerated, with only a small number receiving course credit for participation (14.28% of final sample). Indeed, one study has found that paid participants recruited from *Prolific* generally produce more high quality and reliable data than university undergraduate samples (Peer et al., 2017). While it would be inappropriate to speculate about the statistical significance of the differences in data exclusion rates across the experiments, at the very least, we can say that the concerted effort in Experiment 3 to establish tight laboratory control and exert demand effects over the behaviour of participants failed. Future research would appear, at present, to not be disadvantaged by the use of paid online participant samples although, as stated above, it is essential that stimulus control issues be rectified first before more firm conclusions can be drawn on this issue.

To investigate the possibility that significant interaction effects may have been observed had different stimuli been employed as conditioned stimuli, the sample was split according to the functions assigned to the conditioned stimuli (i.e., CS+ or CS-). It was found that response fluencies were higher when the contingencies of the FAST required similar responses for fruit and positive exemplars, and a different similar response was required for to furniture and negative stimuli, compared to when the configuration was reversed. This

response pattern was the case regardless of whether or not this configuration represented the evaluative conditioning. This increased fluency, across both blocks combined, does not speak to the differential in fluency across the two blocks, which was still present and in the predicted direction at the cohort level. However, it does suggest that responding similarly to fruit and positive was on the whole "easier", clearly indicative from higher fluency in both relating fruit terms to positive evaluative stimuli in the consistent block, and *even the reverse* on the inconsistent block. This is indicative of pre-existing greater relational flexibility with fruit related terms compared to furniture related terms in the repertoires of the participants. While it is true that all participants included in the analysis showed stimulus ratings that were in line with the conditioning contingencies, it does not follow that this alone is sufficient to shift the nonverbal evaluative functions of stimuli, nor verbally established evaluations responded to under response time pressure.

In summary, the in-person supervised participant sample employed in Experiment 3 did not lead to clearer experimental effects as expected. In hindsight, this may not be surprising given the findings of Peer et al., (2017) regarding the superior quality of data recorded using paid professional participants. Nonetheless, those participants deemed to have performed to a minimal standard in terms of adherence (i.e., above chance level responding on the FAST and completing all phases), successfully demonstrated a cross-block performance bias within the FAST procedure that directly and significantly reflected the occurrence of the conditioning. While pre-existing stimulus functions likely compromised the clarity of this effect, a significant FAST effect was achieved for the sample considered as a whole, irrespective of the functions assigned to the conditioned stimuli. This latter observation demonstrates an in-principle effect and a second systematic replication of that same basic effect.

Experiment 3 failed to support the hypothesis that the salience of the unconditioned stimuli used in the evaluative conditioning procedure would significantly alter the magnitude of the FAST effect. However, we can conclude with some confidence now, for the third time in this thesis, that the FAST procedure is sensitive to laboratory controlled evaluative stimulus functions arising as a result of an associative evaluative conditioning procedure.

Chapter 5

General Discussion

5.1 Introduction

This Discussion will consist of an outline of the main findings of each experiment in sequence, accompanied by a review of the most salient implications of the research along with points worthy of consideration regarding methodological weaknesses or alternative explanations for the observed effects. A review of each of the chapters will then be followed by a discussion of global issues pertaining to the research program as a whole, including reflections on the use of remote participants in online research, issues of stimulus control and recommendations for future research.

5.2 Experiment 1

The findings from Experiment 1 produced tentative evidence of the FAST's ability to detect differences in the salience of emotional events, as observed by the mean RFD score for each condition. The difference in fluency scores increased in tandem with the increased salience of the unconditioned stimulus during a simple associative learning experience. However, this latter trend was not found to be significant. Nonetheless, the FAST did successfully differentiate the acquired stimulus functions of the aversive and appetitive conditioned stimuli, indicating effective conditioning, and sensitivity of the FAST to conditioned evaluative functions.

One possible reason that the experimental conditions did not interact with the main FAST effect may be due to random error as a result of small sample size. An *a-priori* power analysis with a defined power of 0.9 and alpha set at 0.05 was conducted, and the suggested sample size of 54 participants was met (final n=62). However, given the division of the sample into 6 cells for the 2x3 ANOVA, the final analysis for Experiment 1 was conducted with about 10 participants in each cell. According to Simmons et al. (2011), however, a minimum cell size of 20 data points is necessary to detect most effects. This point of concern

was addressed in Experiment 2, in which a much larger sample size was recruited to meet the recommendations of Simmons et al. (2011). The significant cross-condition effects found in Experiment 2, taken together with the previous finding that the FAST is sensitive to even weak stimulus relations (Cummins et al., 2018), support this low sample size interpretation of outcomes. Further support for this interpretation is found in the analysis of conditioned stimulus ratings, which also differed across conditions, though this observation is purely descriptive. This suggests that levels of US salience (intensity) were sufficiently different from one another to produce varying evaluations of conditioned stimuli, at least at the explicit level. These ratings served as a manipulation check of the conditioning procedure and suggest that any failure for differences in FAST effects across conditions to precipitate subsequently, is due to noise in the data. This issue is further discussed below.

As argued above, the excessive noisiness of the data likely resulted from the lower than ideal sample sizes employed. In addition to the low sample size, it is probable that at least part of the noise in the data was caused by the remote data collection strategy. That is, while numerous benefits to remote data collection have been identified, including increased diversity (Reinecke & Gajos, 2015), and larger sample sizes (Peer et al., 2017), there are also concerns over the quality of data resulting from such research methods. The data for Experiment 1 were collected from an unsupervised, remote, online sample. Indeed, it was emphasized to participants at several times during the experimental procedure that their attention was required throughout the entirety of the experiment. However, given the lack of experimental control associated with the remote data collection, it is likely that at least some participants failed to give the task their undivided attention, as evidenced by the significant performance-related attrition recorded in Experiment 1. That is not to say that participants were intentionally sabotaging the experiment by not paying attention, but rather it speaks to the impact of environmental distractions that are sometimes unavoidable with remote

research (see Reinecke & Gajos, 2015 for a qualitative analysis). Indeed, environmental distractions have demonstrated significant negative impacts on attention based task performance (Varao-Sousa et al., 2018). In effect, it is important for researchers to understand the potential trade-off between data diversity and quantity, and data quality.

One specific source of environmental distraction that became apparent post-data collection was that despite highlighting numerous times throughout the recruitment, information, and consent phases of the experiment the importance of completing the study on a laptop or desktop computer, several participants participated using mobile phones. Evidence for this was merely anecdotal, based on volunteered self-reports of some participants, and so the exact number of such participants cannot be known. In fact, Inquisitweb software does allow for the logging of user operating systems but this facility had not been employed at the time of the running of Experiment 1 or 2. While some research has shown promise for mobile phone based experimental research (e.g., Reimers & Stewart, 2008; Rachuri et al., 2010), it is important to understand that the FAST had not been optimized for phone screens, and some phases, such as the instructional screens, may have been so small on some devices as to be unreadable. Moreover, mobile phones are designed to be used in a multitude of environments. Thus, not only is participation on a mobile phone not suitable for the motor movement requirements of the FAST format itself, but it also heightens the possibility that participation may occur in less than suitable environments (i.e., noisy public areas). In addition, the mobile phone does not allow for the level of dexterity required for a response fluency measure, given its cumbersome nature and small on-screen operanda. Indeed, one study (Reimers & Stewart, 2008) reported that reaction times recorded during procedures administered via mobile phones are significantly longer when compared to reaction times recorded using the same procedure on a desktop or laptop computer. Given that an unknowable portion of the Experiment 1 sample participated in that experiment using

a mobile device, serious doubt must be cast over the likelihood of having seen any nuanced effects across conditions under such circumstances. At the same time, however, the difference in fluency across blocks was robust enough to be reliably visible across this plus the subsequent Experiment 2, which also potentially suffered from the same methodological compromises. It is an interesting additional consideration, however, that if *all* participants in Experiment 1 had used a mobile phone, the resulting scores would all be proportionately affected by the response time inflation associated with mobile phone usage. Of course, usage of these devices would still have produced noise, but it would have manifested as decreased inter-participant variability in fluency. However, it could be argued that the usage of a mixture of devices by unknown numbers of participants led to even more potential for spurious data effects to emerge. As such, researchers choosing to collect data remotely in future should consider adding device related restrictions to participation.

5.3 Experiment 2

The purpose of replicating Experiment 1 with a larger and remunerated sample was to address the concern that the results arose partly from low data quality, poor task adherence and excessive data variability that is typically characteristic of low sample sizes. Thus, Experiment 2 aimed to establish in-principal effects in support of the hypothesis that the FAST is an effective measure of the emotional salience of conditioned stimuli. The results of Experiment 2 showed that block fluency scores differed significantly from one another, and that the magnitude of this difference varied across conditions, as expected. Individual fluency measures for the consistent block steadily decreased from Condition 1 through to Condition 3. However, this trend was not very apparent for inconsistent block fluency scores. Nevertheless, overall block fluency (i.e., FAST scores) in each condition followed the expected trajectory, with the fluency of responding decreasing across conditions from Condition 1. Finally, significant FAST effects (i.e., a significant difference in fluency across

blocks) were found for Conditions 1 and 2, but not Condition 3, broadly in line with expectations. In effect, Experiment 2 appears to have confirmed the expectations that the FAST would be sensitive to conditioned stimulus relations, whose relatedness, in turn, would be affected by the salience of an unconditioned stimulus used to establish those relations in the first instance.

It is important to highlight, however, that the approach taken in Experiment 2 was not viewed as a solution to the excessive data noise and poor task adherence observed in Experiment 1. Resorting to larger sample sizes and increasing participant motivation through remuneration does not satisfy the requirement to increase understanding of core behavioural process and eventually arrive at principles and processes generalizable to each individual within a study. Of course, there is always going to be a risk of greater data variability with very small sample sizes, but the arguably excessive sample sizes employed in Experiment 2 serve merely to chase down the behavioural effect of interest in an in-principle manner. They do not fill the scientist with confidence that there is as of yet sufficient stimulus control over the phenomenon of interest. On the other hand, a small-N oriented study is only acceptable in cases in which the aim is merely to demonstrate already established behavioural phenomena (e.g., McGlinchey et al., 2000). Smaller sample sizes are also perfectly permissible where the experimental procedure allows complete behavioural control, sufficiently high as to eliminate variability (McLoughlin & Roche, 2022). However, where the research is aimed at testing specific hypotheses (see Hughes et al., 2017), the use of smaller sample sizes normally associated with behavioural research becomes questionable and the ability to generalise findings from such studies becomes limited (McLoughlin & Roche, 2022).

The IRAP research body has recently been the subject of criticism along the foregoing lines. More specifically, several IRAP studies (e.g., Power et al., 2017) have been

considerably underpowered, given an excessive number of post-hoc data analyses and overly complex interactive designs consisting of a large number of independent variables but with a relatively small sample size. In many cases, Bonferroni correction is not applied by IRAP researchers conducting multiple post-hoc tests. Such a criticism can certainly not be levelled against Experiment 2 of the current thesis. Ironically, it is the potentially *excessive power* that may raise cause for concern, because seeking it represents a deviation from the normal behaviour-analytic focus on enhanced behavioural control and healthy effect sizes, over the establishment of effects extracted inferentially using statistical models. Nevertheless, it is important to point out that effect sizes remained respectable, in addition to the increased statistical significance of the interaction effects found in Experiment 2. Indeed, in Experiment 2, the effect size for the main interaction effect between block fluency scores and condition was much larger than that observed in Experiment 1. The main effect for block was marginally smaller in Experiment 2, though this is likely attributable to the stronger impact of conditioned stimulus assignment on fluency scores, as suggested by the larger effect size for that analysis. Despite the impact of stimulus function assignment on block fluency scores, Experiment 2 nonetheless found significant FAST effects across the board, with respectable effect sizes, indicating that the phenomenon was real and robust when considered across a large number of participants, as opposed to being merely statistically probable (i.e., low p values) given the larger sample size.

It is worth acknowledging at this point, that despite addressing the problem of sample sizes in Experiment 2, no modifications were made to the original Experiment 1 procedure regarding the impact of stimulus function assignment on FAST fluency scores. The reader may recall a brief discussion based on the findings of Warriner et al., (2013) that allowed the author to conclude that the conditioned fruit stimuli likely had more positive pre-experimental functions than furniture stimuli for most participants. This likely compromised

the effectiveness of the evaluative conditioning procedure. Indeed, this issue could have been easily addressed with minimal procedural alterations involving the substitution of the conditioned stimuli with more neutral and equally balanced stimuli. However, such alterations to the experimental procedure would have prevented the *systematic* replication of Experiment 1, which is an important part of the investigative experimental procedure for identifying underlying processes and phenomena (see Schmidt, 2009). Indeed, it has been argued by several researchers that replication of experimental outcomes is a vital aspect of knowledge production (Shapin & Schafer, 1985). Thus, even aside from the issue of addressing concerns over poor task adherence and sample size, a close replication of Experiment 1 was merited in any case. Should the results have turned out to be the same with the increased and remunerated sample in Experiment 2, just as much would have been learned about the nature of the phenomena under analysis although the conclusions would have been different (see also Popper, 1959). Thus, Experiment 2 purposefully retained all aspects of the experimental procedure, even those that constituted limitations, in the interest of determining whether the lack of effects observed in Experiment 1 were caused jointly by sample size constraints and / or poor task adherence.

It is unfortunate that not all null effects are further investigated in research. However, research novelty is rewarded over systematic replication through publication, which contributes to why it is practiced so infrequently (Schmidt, 2009). This holds true even when replication may offer worthwhile contributions, for instance, in determining whether a given finding truly exists, or is attributable to a Type 1 error. Even failed replication attempts offer value in that they outline the necessary conditions for the phenomenon of interest to occur (Schmidt, 2009; see also Sidman, 1960). Given the increasing appreciation of the burgeoning replication crisis currently plaguing much of psychological research (see Ioannidis, 2005), it is important that researchers are encouraged to reproduce, replicate, and measure the

reliability of research findings. Indeed, this is especially important for behavioural researchers choosing to adopt inferential statistical approaches, considering that small sample sizes are notoriously associated with lower power (Ioannidis, 2005). While a critique of publishers who demand novelty in research is far beyond the scope of this thesis, the trickle down impact of this on research integrity must be realised. Experiment 2 of the current study therefore partly represents the current researcher's effort to assess the replicability of findings as well as establish in-principal effects in the first instance.

Another issue worth considering is that the robust effects observed in Experiment 2 may be partly the result of the renumeration of participants. The offer of financial renumeration might have increased participant motivation, especially given that within the modus operandi of Prolific, payment for participation in studies is withheld until engagement with the task is confirmed by researchers. In effect, while payment is not conditional on any particular *type* of performance, it is in fact conditional upon engagement with the task to satisfactory standards. This feature of Experiment 2 distinguishes it from Experiments 1 and 3, although in both of those cases, some participants received course credit unconditional to performance. It is also worth noting, however, that while effects were more robust and in the expected direction in all cases for Experiment 2, there was not a sufficiently impressive decrease in non-engagement rates across participants to suggest that remuneration was in fact increasing the quality of the data, although the observation of poor quality data due to non-adherence to the task is not incommensurate with improved performance among a particular cohort of participants, due to the effect of increased motivation.

5.4 Experiment 3

The purpose of replicating the second experiment was to reproduce the effects that were established in-principle with a large sample, but with a small laboratory supervised

sample. The rationale was that if the effects observed in Experiment 2 were real and not spurious, they should be replicable on a small scale, where adherence to the task is improved through direct supervision. However, the results of Experiment 3 did not successfully replicate those from Experiment 2. The interaction effect between task block and stimulus salience did not prove to be significant. While there was an overall main effect for block, the effect size was also reduced, relative to Experiment 2. The dataset in Experiment 3 also suffered from considerable variability. That is, there were some noteworthy outlying scores for the RFD variable, which may have impacted the outcome of analysis. It appears that as a result, the correlation between RFD scores and average stimulus ratings was not significant, whereas it had been in both previous experiments, as expected. Moreover, there were no significant differences in fluency between the consistent and inconsistent blocks for any condition considered alone. Given such a deviation from the expected results, it was deemed appropriate to directly assess the magnitude of the impact of stimulus function assignment on FAST scores. This split sample post-hoc analysis displayed some interesting trends, which will be discussed in terms of pre-existing stimulus functions. First, however, a brief discussion on the main effect of block is worth considering, along with issues related to data removal prior to analysis.

The homogeneity of FAST scores across conditions in Experiment 3 seemed to suggest a loss of experimental control during at least one of the experimental phases (i.e., during conditioning or during the FAST). Nonetheless, it is important to remember that a main effect, however small, was found for block fluency scores, irrespective of condition (i.e., main effect). That is, there was a sample-wide functional difference between aversive and appetitive conditioned stimuli reflected in FAST performances, even though no functional differences in conditioned stimulus salience were observed. The general sensitivity of the FAST to the conditioned evaluative functions therefore might suggest that the issue of

large variance in FAST data lies not with the FAST *per se*, but with the conditioning procedure itself, which was designed to change the pre-experimental functions of the conditioned stimuli to varying degrees. this suggestion is corroborated by the observation that there was large variance in the post-conditioning stimulus ratings provided for the conditioned stimuli. The issue of the effectiveness of the conditioning procedure will be more fully addressed in a subsequent section. For now, however, other possible sources of data variation related to data cleaning procedures employed in this research will be explored.

As mentioned previously, the conservative approach to data cleaning in Experiment 3 may have interacted negatively with some of the predicted statistical analysis outcomes. Data from all participants who met the predefined exclusion criteria were retained in the interest of preserving behavioural variability within the dataset. Within the IRAP, data reduction is common practice and is comparable to IAT standards in terms of procedural opacity (see Barnes-Holmes et al., 2010b; Greenwald et al., 2003). The use of such data reduction methods for statistical purposes might be considered to be antithetical to a functional approach. That is, given that our field is based on an interest in behavioural variability (Sidman, 1960), it is odd to remove variability post-hoc in order to create the impression of steady-state behaviour for the purpose of statistical analysis. Thus, the extension of a commitment to a functional approach within the FAST research program to include an eschewing of data reduction wherever possible, may in fact have resulted in the weakening of statistically abstracted effects.

It is understandable, of course, that with more and more reliance on inferential statistics necessary to answer important questions about the applicability of implicit tests, it is at times essential to eliminate extreme instances of variability for the purpose of conducting the statistical analysis of interest (Cousineau & Chartier, 2010). As has been mentioned, and will be discussed further below, a degree of data reduction was unfortunately necessary

throughout the current study. However, the current researcher refrained from implementing any data reduction strategies insofar as was reasonable. Given the very small number of extreme outliers in the dataset and given the fact that the correlational analysis in question was supplementary, rather than core to the aims of the study, it was felt that the lesser of two evils was to leave the dataset in its' original state for the purpose of the correlational analysis despite the risk to the significance of outcomes (see Osborne & Overbay, 2004).

It is perfectly reasonable, therefore, to conclude that the correlational analysis in Experiment 3 was not significant due to the inclusion of outliers in the data set. However, it is also worth considering grounds against the removal of outliers in such situations, at least within the behavioural tradition. One study by Amd and Passarelli (2020), involved recording implicit and explicit responses to conditioned stimuli following evaluative conditioning of those stimuli. They found that implicit evaluations were always altered by the conditioning, whereas change in explicit evaluation was more variable. It seems then, that explicit and implicit evaluations may be two functionally different phenomena, and that variability in responding to stimuli will be increased to a greater or lesser extent, depending on the measure employed to quantify the effect of conditioning. Indeed, this effect has also been demonstrated by Noel et al. (2019), who successfully altered implicit, but not explicit evaluations of heavy drinking behaviours following a brief evaluative conditioning procedure. This distinction between implicit and explicit processes is well established in the cognitive field, since the introduction of the implicit attitude concept (Greenwald & Banaji, 1995). Within the behavioural field, there have been some attempts at explicating this phenomenon. Recall the Multi-Dimensional Multi-Level and Hyper-Dimensional Multi-Level Frameworks, offered by Barnes-Holmes et al. (2020a, 2020b). While these theoretical models have the potential to offer some valuable insight into the implicit-explicit distinction from a behavioural perspective, there is as of yet no *prospective* empirical support for these

frameworks. While they may be useful as a heuristic for understanding behavioural phenomena, it would be as of yet unwise to draw on them as fully explanatory given their post-hoc, theoretical nature. Those issues notwithstanding, however, the general observation that implicit and explicit measures of the same stimuli following conditioning may diverge in unexpected ways may go some way towards explaining the lack of correlation between the explicit and implicit ratings of stimuli in the current case.

While there was a clear cross-block fluency difference observed in Experiment 3, it is interesting to note that the effect of the stimulus salience variable most notably impacted performance on the consistent block. Interestingly, a similar pattern was observed in the other two experiments. This finding indicates that response fluency on a given block is greatest when the contingency of reinforcement in place for that block is consistent with the conditioning contingencies. In contrast, response fluency on a block for which the response contingencies are inconsistent with the conditioning contingencies is not as notably diminished. Put simply, the enhancement of fluency is achieved more readily through conditioning procedures than is the retardation of response fluency. As a result, differences in overall FAST effects across the blocks was driven by changes in fluency during the consistent task blocks, for the greatest part. However, there is at present no way to reliably conclude with confidence whether the increased fluency on the consistent block is best characterized as increased S+ control or decreased S- control. That is, it is not easy to assess whether the conditioning had the effect of increasing the appetitiveness of the CSpos or the aversiveness of the CSneg.

While it does not address the foregoing issue directly, the use of stimulus databases that quantify verbal stimulus valence (e.g., Warriner et al., 2013) would certainly help to eliminate imbalances in valence across conditioned stimulus classes, that would at least attenuate uncontrolled for variation in the functions of the aversive and appetitive conditioned

stimuli. In addition, it may well be that stimuli with pre-existing functions of a particular type somehow engender greater levels of response fluency, independent of conditioned functions. This is an empirical matter, because at the very least the valence of conditioned stimuli should be controlled, and ideally balanced from the outset. Indeed, such is the apparent effect of pre-existing stimulus functions in Experiment 3 on FAST effects and / or the efficacy of the evaluative conditioning procedure, that post hoc analysis of FAST effects demonstrated the interactions occurred as expected for participants exposed to one stimulus assignment configuration and not the other. This finding supports the widely held assumption that the relatedness of two stimuli increases with stimulus potency. The FAST, as a measure of stimulus relatedness, should therefore be, and indeed was, sensitive to increased US, and consequently, CS salience, albeit for half of the sample. Future research should bear in mind, therefore, the risks of using stimuli that have not been assessed for pre-existing functions.

The foregoing discussion regarding the impact of salience manipulations on specific aspects of the FAST performance suggests a very interesting possibility regarding the FAST index itself. That is, if consistent block performances are sensitive to the salience of stimuli, then, in principle, the fluency observed in performance on the consistent block alone may serve as an index of the relatedness of stimuli within both stimulus pairs. While it is very useful to contrast such relatedness indices against a measure of resistance to change in the functional class structure, this nevertheless presents itself as a theoretically viable possibility.

One way in which the current FAST procedure could potentially offer insight into the dynamics of the underlying bias being assessed would be to examine more closely performances on each individual task block. To do this, the strength of the relations under assessment would need to be manipulated, as the current experiments have done, with the impact of this manipulation on individual block scores being the measure of interest. This suggestion follows the observation that the same trend, whereby stimulus relatedness appears

to affect fluency on the consistent block relative to the inconsistent block, has also been observed by Cummins et al. (2018). In effect, it might be suggested here that the most sensitive aspect of the FAST performance to the structure of verbal relations is intra-class stimulus relatedness, rather than the inter-class unrelatedness. This insight is important as it represents a genuine contribution to our understanding of the core process underlying a behaviourally conceived implicit test.

The suggestion that stimulus relatedness can be assessed by only examining performance on the consistent block does not negate the utility of the inconsistent block, however. The inconsistent block likely serves an important function itself as a contrast task against which to compare performances on the consistent block. Nevertheless, the reliable, predicted variance in performance on the consistent block suggests an important theoretical conclusion regarding the dynamics of implicit test performances. It is important to reiterate that this conclusion could not be made in the absence of basic laboratory research with artificially created stimulus relations, manipulated across various continua, that contribute to the underlying relations under assessment. In other words, conclusions of this kind could only be arrived at with process-based research, and much theoretical inference using real-world stimulus classes, and in the absence of the manipulation of key variables affecting stimulus relatedness. This has been the Achilles heel of implicit testing research to date, which has been regrettably slow to adopt a basic research approach that can allow a light to be shone on the core processes with minimal inference and theoretical speculation. Nonetheless, this issue is worthy of consideration in future research.

5.5 Global Considerations

5.5.1 Online Data Collection

One aim of the current study was to compare differences in data quality when collected online with unsupervised samples, versus in the laboratory with supervision.

Attrition rates between samples were compared, after which it became evident that the largest attrition was recorded in Experiment 3, in spite of full engagement with each experimental task from all participants in that cohort not observed elsewhere. In this experiment, 17.24% of the original sample were removed due to failure to respond on the FAST at above chance levels of accuracy. The levels of attrition for this same criterion in Experiment 1 and 2 were 12.2% and 11.94% of the original samples, respectively. This indicates that the Experiment 2 cohort was most attentive and receptive to the response feedback offered in the FAST. All participants in Experiments 1 and 2 received renumeration in the form of unconditional course credit or conditional financial payment, respectively. Recall that in Experiment 2, participants were informed at the outset that they could only expect to receive renumeration for their participation given satisfactory completion and engagement with the task, whereas Experiment 1 participants received their renumeration irrespective of these factors. Notably, only a small portion of Experiment 3 participants received unconditional renumeration. Thus, in support of evidence offered by Palan and Schitter (2017), the current study demonstrates that unsupervised, conditionally renumerated participants recruited via Prolific provided the highest quality data (defined here in terms of attentiveness). This study also corroborates the findings of Peer et al. (2017), which suggested that participants recruited from a participant pool, particularly university students, tend to provide the lowest quality data when compared with online paid and unpaid samples. The current findings do not suggest that demand characteristics (Orne 1962) had any effect on improving performance on the FAST for in-person participants. As such, when quality is associated with attentiveness, it seems that conditional financial renumeration produces the best quality data.

Another aspect of data quality not previously discussed, but that also was cause for concern, was the number of participants who failed to show evidence of effective evaluative conditioning (i.e., lack of difference between average CS- and CS+ ratings). Weber and

Cook (1972) suggested that in conditioning experiments, participants aware of conditioning contingencies may actively respond in opposition to these contingencies in the interest of seeming independent. Indeed, Page and Lumia (1968) directly assessed levels of participant cooperation with demand characteristics. Participants who were aware of the experimental aim admitted to being intentionally uncooperative with the experimental demands (Page & Lumia, 1962). This finding directly opposes Orne's (1962) description of the good subject, who alters their behaviour to help confirm the experimental hypothesis when they become aware of it. Despite the widely accepted concept of the good subject, this only truly applies to research conducted in a laboratory setting, whereas much of the current data was collected remotely. Even the data collected in the laboratory has less than satisfactory levels of attrition associated with failure to show evident conditioning (17.24% of sample). Thus, the above research indicates that participants who independently realised the aim of the conditioning procedure (i.e., to establish emotional functions for innocuous stimuli) may have actively provided ratings that opposed the associative contingencies.

While there may be some truth to such claims, assertions about a participants' awareness, even those taken from the participant through post-experimental interviews, are not infallible (Weber & Cook, 1972). Conclusions based upon such assertions are therefore hypothetical and weakly supported. Thus, an empirically supported alternative is preferable. For instance, Varao-Sousa et al. (2018) compared levels of attentiveness across different environments, and found higher inattentiveness levels for participants in everyday settings compared to those in a laboratory, and concluded the cause of this was environmental distraction. Failure to show evidence of conditioning through CS ratings in the current study is arguably associated with inattentiveness during the conditioning phase of the experiment. Indeed, Experiment 2, which used an unsupervised but renumerated sample, saw the largest attrition associated with this failure (32.56%). However, the environmental distractions

argument (Varao-Sousa et al., 2018) does not explain the lowest attrition rates being recorded in Experiment 1, also a remotely collected sample (14.44%; Experiment 3: 17.24%). Possible causes for the observed attrition rates in the current experiments will be discussed further below, addressing two areas of concern: the effectiveness of the conditioning procedure itself, and the degree of stimulus control attained within the experimental procedure.

5.5.2 Conditioning Procedure Artefacts

To address the first concern, the efficacy of the conditioning procedure used in the current study, which attempted to establish evaluative functions for innocuous verbal stimuli through respondent conditioning will be examined. As per a century of laboratory tradition, effective respondent conditioning requires a period of negative stimulation between (a) the response consequence and (b) the commencement of the next trial. That is, a latent period where no stimulus is presented is required between trials to effectively establish conditioning. Importantly, the period of negative stimulation must be much longer than the presentation of the unconditioned stimulus, to facilitate stimulus discrimination (see Dinsmoor, 1995). Recall that the latent period between trials (i.e., the intertrial interval; ITI) in the current study was 8-12 seconds. Similarly, the presentation of the unconditioned stimulus itself remained on screen for only 5 seconds. Thus, the ITI exceeded the US presentation duration by as little as 3s, and a maximum of 7s. Even more concerning in hindsight is the fact that the entire CS-US trial required a minimum of 6s, reducing the salience of the CS-US contingency even further with respect to the intertrial interval and the onset of the subsequent trial. In contrast, several other studies within the behaviour analytic fields have used considerably longer ITIs. For example, Plaud & Martini (1999) employed an interval of 120 seconds, while Dougher et al., (1994) used an ITI of between 90 and 120 seconds. Even within the associative conditioning literature, in which shorter intertrial intervals are common, researchers have

used intertrial intervals in many cases at least as long as what was employed here in the current study (e.g., Baeyens et al., 1992; 12s).

While the intertrial intervals may have been sufficiently short in the current study as to compromise the quality of conditioning, it is important to understand that as a partly translational research effort, an evaluative conditioning procedure was used that more closely resembled that used in the evaluative conditioning literature, generally running parallel to the behavioural analytic research field. Within that literature, robust conditioning effects are often reported with surprisingly low ITIs. For instance, ITIs employed by Glaser and Kuchenbrandt (2017), Kattner (2014), and Olson and Fazio (2006) ranged between 1-2.5s. Part of the reason why shorter intervals are employed by cognitivists is related to their view that contingency awareness helps to establish evaluative conditioning (Hoffman et al., 2010). That is, while the behavioural tradition operates on the assumption that the CS-US contingency is the most important part of the conditioning process, the cognitive tradition attributes effective conditioning to mediating mental processes, namely, conscious propositional knowledge (De Houwer, 2006). Consequently, cognitive researchers typically place less emphasis on the relative length of ITI to US presentation. In effect, it may not follow automatically that the short intertrial intervals observed in the current study are solely responsible for the less than optimal conditioning effects and the requirement to eliminate participants whose ratings of the conditioned stimuli had not been as predicted following the conditioning procedure.

Interestingly, one study has argued that attributing poor conditioning outcomes to attentional factors, rather than lack of contingency awareness, may be more accurate (Field & Moore, 2005). That is, stronger stimulus conditioning is achieved when attention to the procedure is heightened, irrespective of contingency awareness. By introducing a distractor task in one condition, the researchers found that poor conditioning outcomes were not as

explicable by a failure to identify the CS-US contingency as they were by the presence or absence of a distractor task. In other words, the distractor task negatively impacted conditioning even when the participant correctly identified the contingency (Field & Moore, 2005). Thus, identifying the source of poor conditioning outcomes is a rather complex matter and would involve too much speculation post hoc in the current case. The conservative thing to do therefore, may well be to resort to robust behaviour analytic conditioning procedures, rather than to rely on what appear to be sensitive and risky evaluative conditioning procedures.

5.5.3 Stimulus Control

The second source of weaker than expected FAST effect difference across conditions, especially in Experiments 1 and 3, may relate to an important stimulus control issue. That is, the current FAST procedure was not so much designed to measure the relation established between a CS and a US, but to measure the generalized effect of the aversive and appetitive functions established for two conditioned stimuli. More specifically, during the conditioning procedure, emotionally salient visual images were used as unconditioned stimuli. Conditioned stimuli consisted of randomly chosen innocuous verbal categories. Importantly, however, the unconditioned stimuli were never used in the FAST procedure as target stimuli. Such a procedure would have involved attempting to establish functional response classes between compatible conditioned and unconditioned stimuli in one block, and incompatible conditioned and unconditioned stimuli in the other block. Instead, novel evaluative terms were used in place of the unconditioned stimuli. It was assumed that the conditioned functions for the conditioned stimuli would result in a generalized expansion of the conditioned stimulus classes to include these evaluative terms. Indeed, the results of all three experiments confirmed this assumption. That is, classes including appetitive conditioned stimuli were found to have expanded to include positive evaluative terms and for such classes to show

resistance to change during the inconsistent block. Similarly, classes with aversive conditioned stimuli expanded to include negative evaluative stimuli and for this larger class, showed resistance to change under the reinforcement contingencies of the inconsistent block. This procedure, in hindsight, might have yielded lower FAST scores than would have been achieved had the unconditioned stimuli been used in place of the generalized evaluative terms. Future research should compare FAST scores resulting from measurements of directly conditioned relations, as opposed to generalized ones as was achieved here. Indeed, while not a perfect analog of the current research, one study involving both an operant avoidance and elicited respondent fear response, reported lower probability of avoidance and fear with increasing semantic relatedness of the probe stimulus to the conditioned stimulus. In other words, the aversiveness of a conditioned stimulus appears to decrease with increasing semantic distance from the original conditioned stimulus such that generalized fear responses are weaker than directly conditioned ones (see Boyle et al., 2016). In addition, it has been documented that FAST effects decrease as a function of stimulus relatedness (Cummins et al., 2018, 2020). Putting all this together, the unreliable cross-condition effects observed in the current study may be related to the fact that the FAST was attempting to measure a generalized outcome of conditioning rather than the relatedness of the conditioned and unconditioned stimuli themselves. There are two ways in which this might be addressed should similar research be completed again in future; firstly, one might employ the same stimuli, conditioned and unconditioned throughout all phases of the experiment. Secondly, and what may be considered a limitation of the current study, a within-subjects approach may be employed, wherein the same participants receive conditioning at all three levels of salience, completing the FAST in between each conditioning phase. Such an approach would also be useful in addressing some sample-size related issues that were experienced in Experiments 1

and 3. It would also provide greater control over unwanted, confounding between subjects effects, namely, varying levels of attention and task motivation.

5.5.4 Stimulus Function Assignment

The stimulus function assignment was found to have an unfortunate, unexpected impact on FAST scores. A significant interaction between stimulus function assignment and block fluency scores was present in all three experiments. More specifically, when Experiment 3 block fluency scores were examined separately, on the basis of stimulus function assignment to the fruit and furniture verbal classes, it was found that participants consistently demonstrated higher response fluency *on both test blocks* when fruit and positive exemplars shared a functional response, and furniture and negative words shared a different functional response, than when these response contingencies were reversed. Although it is hard to determine definitively whether this response pattern came from pre-existing positive evaluations of fruit, or conversely, negative evaluations of furniture, population-normed ratings on word valence and salience (Warriner et al., 2013), may help clarify the issue.

Warriner et al., (2013) indicated that at a population level, evaluations of fruit tend to be more positively valenced than furniture. This suggests, therefore, that fruit likely had pre-existing positive conative functions for the entire cohort, resulting in weaker than expected main effects across all three experiments, and the unexpected failure to observe interactions between main effects and unconditioned stimulus salience across Experiments 1 and 3. Future research would surely benefit from reference to databases like that supplied by Warriner et al. (2013; but see also Buchanan et al., 2016, McRae et al., 2005 and Vinson & Vigliocco, 2008) in selecting conditioned stimulus sets, to control for potentially confounding pre-experimental stimulus functions.

The confound outlined above was unanticipated, because conditioned stimulus classes were chosen on the basis of assumed equal salience and valence characteristics. Assuming this salience symmetry across the conditioned stimulus classes, the unconditioned stimuli were selected carefully from a standardized database (the IAPS) that assured matched arousal and orthogonal valence across the aversive and appetitive stimuli. This level of stimulus control was sought for the unconditioned stimuli on the basis of research demonstrating a confounding impact of salience *asymmetry* on implicit test effects. As one example, Perkins and Forehand (2006) assessed the impact of stimulus valence on self-concept IAT effects. Attribute categories included self-representative descriptors (e.g., ambitious) and words of a similar valence but opposite semantic meaning (e.g., easy going). Concept categories included self- and other-representative stimuli, resulting in idiographic IATs for all participants. Perkins and Forehand (2006) isolated valence and held semantic meaning constant; such that valence of attribute categories was manipulated (e.g., positive valence= ambitious, negative valence= cutthroat). IAT responses were quicker when the self-concept and actual attributes were assigned the same response key, and slower when the self-concept was assigned a similar response to an attribute of similar meaning but reversed valence. The findings of this study suggested stimulus valence can impact significantly on IAT effects. Rothermund and Wentura (2004) found similar results in their study, wherein salience asymmetry and valence were isolated separately to quantify the impact of each on IAT effects. Stimulus valence was manipulated across two groups by including (a) clearly positive and (b) clearly negative celebrity names alongside unknown names as target variables, with positive and negative words as attribute variables. Salience symmetry was manipulated by the presentation of known celebrity (salient) or unknown (non-salient) names. Irrespective of group membership, stimulus salience was the main driver of the effect. According to Rothermund and Wentura's (2004) salience asymmetry account, negative and novel stimuli

are more salient and easier to categorize together than apart. Similarly, positive, and familiar stimuli will be less salient and also easier to categorize together. Indeed, this account held true within their experiment, as participants more easily categorized celebrity names with positive stimuli, regardless of whether the names represented positive or negative celebrities.

Together, Perkins and Forehand (2006) and Rothermund and Wentura (2004) demonstrated that asymmetrical target stimulus valence and salience both confound IAT effects. This underlined the importance of controlling these variables in the current research. Retrospectively, we can reasonably conclude that the conditioned stimuli were themselves already functionally asymmetrical, which consequently rendered any attempts at establishing stimulus functions that were orthogonal across verbal categories, but for which arousal indices were equal, was unlikely to be very successful. Indeed, using novel stimuli as conditioned stimuli in place of the chosen fruit and furniture categories would have eliminated such asymmetries, however in doing so, a training procedure to form stimulus classes would have been necessary prior to the conditioning procedure, unnecessarily elongating the process. Also recall that as part of the translational effort of this research, using natural language categories within the FAST represents a concentrated effort to engage in more applied research.

Attempts at evaluating the degree to which stimulus asymmetries, and separately, stimulus relatedness, contribute toward implicit test scores is a very complex matter. More specifically, stimuli can elicit multiple functions. With respect to the current study, it is possible that conditioned stimuli simultaneously held appetitive and aversive functions. This is a problem for tests like the IAT and the FAST because they are 'relative' measures that indicate only the degree of relatedness between stimuli relative to the relatedness of others. That is, they do not allow one to ascertain the nature of that relationship. For example, if an individual shows a racial bias towards White people on the IAT, we do not know if they have

a pro-White (relative to Black) or anti-Black (relative to white) bias. These two very different phenomena will lead to the same test outcome. This matter is not unrelated to the stimulus asymmetry issue, insofar as the results of a relative implicit test do not facilitate conclusions about the absolute appetitive or aversiveness of stimuli. They are designed only to indicate the *relative* aversiveness or repetitiveness of stimuli. Thus, it can often be not apparent that a test effect is the result of the relative degrees of aversiveness of two stimuli, when those two stimuli are aversive, rather than the result of relative degrees of aversiveness when one of the stimuli is aversive and the other is appetitive. In other words, uncontrolled stimulus asymmetries along any stimulus dimension will not always be obvious in the outcomes of these tests, as they are not designed to test this feature of relatedness among stimuli.

IRAP researchers however, claim that their test procedure allows one to assess, in a more absolute way, the non-relative properties of stimuli. For example, Hughes et al. (2017) employed the more nuanced IRAP procedure (not directly comparable to the IAT or FAST in that it involves four separate trial types rather than two) to assess the degree to which two historically opposed groups (i.e., northern Irish Catholics and Protestant) responded on an IRAP with a pattern that represented in-group favouritism and/or out-group degradation. In-group exemplars were assumed to hold positive valence for each social group, and the comparison group was assumed to be negatively valenced (i.e., for Catholic participants, protestant exemplars were the assumed negative target stimuli). Neither social group responded with out-group degradation, but both groups displayed in-group preferences. However, Hughes et al. (2017) were only able to draw these conclusions due to the non-relative assessment style of the IRAP. That is, the IRAP assesses relational equivalences between stimuli but also indexes the relational non-equivalences on a separate trial type. Together, these two trial types provide a more nuanced picture of the degree of relatedness between any two verbal classes. Moreover, the IRAP can assess relations other than

equivalence that might obtain between any two verbal classes. While this methodology is highly elaborate and promising, it was not the focus of the current research, which attempted to further develop a more basic implicit test methodology from the ground up based on first behavioural principles. The reader is re-directed to Chapter 1 for a more elaborated argument regarding the merits of a methodical principle-based approach, as adopted by the FAST research tradition, but not by the IRAP research program. The matter of interest is that a deeper initial understanding of the various functions held by class exemplars would simplify research. Specifically, it would negate the need for such elaborate processes generally associated with social psychological research questions, and allow researchers to draw more reliable conclusions given their understanding of the variety of stimulus functions elicited by their stimuli during the design phase. For instance, stimulus classes could be created ab initio using abstract characters, and extended conditioning procedures in which all members of classes had similar functions established for them. This is a tedious procedure that was considered for the current study, but later abandoned on the basis of expediency, but also in developing an ecologically valid demonstration of the FAST's propensity for assessing the emotional / evaluative potency of conditioned stimuli.

Another viable way in which the stimulus control challenge could have been addressed was to simply introduce a measure of stimulus evaluation prior to the conditioning procedure. Specifically, a conditioned stimulus rating scale administered at the very beginning of the experiment before any conditioning had taken place would have provided an indication of noteworthy differences in stimulus functions valence and arousal (i.e., salience asymmetries) at baseline. Indeed, such baseline measures would have offered statistical control for post-conditioning effects. Amd and Passarelli (2020) used pre and post conditioning stimulus ratings to achieve a measure of the effectiveness of their conditioning procedure. Stimuli were ranked implicitly and explicitly to indicate the affective reactions

they had toward the stimuli. Conditioning then took place, and after this phase was concluded, rankings were recorded once again. The difference between pre and post conditioning rankings indicated the influence of the conditioning procedure on stimulus evaluations. The absence of pre-conditioning stimulus ratings, like that employed by Amd and Passarelli (2020), is a limitation that can be easily addressed in future research.

On that note, another issue with the present conditioning procedure that may have contributed to the weak cross-condition effects is that the unconditioned stimuli used to establish evaluative functions for each condition were not sufficiently different to one another in their salience to cause measurable differences in implicit test effects. The unconditioned stimuli for each condition were selected from the IAPS on the basis of scaled differences in salience and valence ratings. This allowed systematic control over the degree of variance in stimulus salience and valence across conditions. Within the IAPS, ratings range from 1 (very weak arousal and low valence) to 9 (very strong arousal and high valence). For ethical reasons (i.e., in the absence of evidence that selecting the most aversive images as unconditioned stimuli was actually required), ratings for all images selected for the current study did not include the highest or lowest rated images. For instance, IAPS ratings ranged from a low of 2.46 to a high of 4.99 for aversive stimulus valence across three conditions. This left an average of around one point difference in valence between each condition. The same is true for arousal ratings and indeed for appetitive stimuli also. In short, the variance in image valence and arousal between conditions may not have been pronounced enough to result in significant differences in FAST scores between conditions.

While the images employed as unconditioned stimuli could have been better differentiated based on standardized ratings, it was felt that the current procedure would at least establish an in-principle effect, which it appears to have done at least in Experiments 1 and 2. Experiment 3 is anomalous in that the cross-condition effect was in fact not even in

the expected direction at the descriptive level, although the prima facie analysis of the descriptive data suggests that the trend is chaotic enough and condition differences small enough to not represent a pattern based on experimental manipulations, but rather random error.

Given the ethical constraints placed voluntarily on the experimental procedure, a small methodological adjustment may also have been required, at least with the benefit of hindsight. That is, given the reduced differentiation of unconditioned stimuli across conditions, it may have been feasible to exclude Condition 2 and to include a control condition that was not exposed to any conditioning at all. This group would not have required exposure to a conditioning procedure and should have shown very weak to no effects on the FAST procedure. In addition, a control condition would have provided a baseline level of bias towards fruit related stimuli which could have functioned to provide a factor by which to adjust effects for those participants for whom fruit functioned as a CS+.

One final consideration with regards to the weak cross-condition differences in effect sizes, relates to the suitability of implicit tests in general for assessing the magnitudes of effect in a linear way. That is, it may well be that these test measures are more binary in their functioning then would be hoped for, and while they may identify the existence of a stimulus relation, may not be perfect indices of the strength of that relatedness. Admittedly, Cummins et al. (2018, 2020), have previously shown that the FAST is sensitive to controlled stimulus relatedness, amongst both laboratory created stimulus equivalence classes and socially established word equivalences. However, that research dealt only with complex verbal relations involving either derived relations of known nodal distance or naturally occurring verbal relations with an unknown genesis. In other words, it does not follow that because the FAST is sensitive to relations established through operant processes, that it is also necessarily sensitive to the relatedness of stimuli established through associative conditioning processes.

Moreover, participants' ability to respond to conditioned stimuli in line with the conditioning contingency does not imply that an implicit test would be sensitive to the differences in those ratings produced verbally and post-hoc in a deliberative manner without time limits. Such ratings may constitute a different behavioural phenomena than the automatic responding to the immediate and most dominant stimulus functions of the stimuli presented within the FAST procedure. Indeed, the entire implicit testing field is based on the observation that supposed 'conscious' propositions regarding the relations between stimuli are different from automatic immediate responses (c.f., Ciarrochi et al., 2016). Both reflect valid and real aspects of a stimulus which can produce semantically incompatible responses (e.g., a conscious preference and an automatic dislike; see Greenwald et al., 1998). That is, the commonly less than perfect correlation between automatic (implicit) and deliberated (explicit) responses is outlined by Greenwald et al. (1998) as evidence of the divergence of the two constructs. Behavioural and cognitive communities agree that both automatic and more deliberated responses exist separately and are therefore not expected to correlate (Hughes et al., 2011). In fact, within the field of behaviour analysis, a conceptual model describing the process by which these two response forms can deviate has been developed in the context of IRAP research (i.e., the Relational Elaboration Coherence Model; Barnes-Holmes et al., 2010).

5.5.5 Response Fluency Differential (RFD) Metric

As mentioned in the methodology, this study employed the novel Response Fluency Differential (RFD) metric to quantify FAST effects. This metric has only been employed in in-house research to date, and as of yet has not been used in published studies. The initial scoring metric proposed for the FAST by O'Reilly et al. (2012), the Strength of Relations (SoR) index, was based on a "trials to criterion" measure, which itself was based on a simple percentage correct score system, common in stimulus equivalence research up to that point.

Specifically, rather than use a finite test block as had been employed by seminal researchers such as Watt et al. (1991), the FAST procedure presented a potentially infinite number of trials on each test block and recorded the number of trials required for participants to produce at least nine correct responses in any series of ten trials. The SoR method also involved the use of baseline blocks, which employed nonsense stimuli to indicate the individual's baseline speed of functional response class acquisition, in addition to that for the critical test blocks. These baseline blocks contextualized critical block performances in that they quantified the degree of bias toward or against a given stimulus configuration, relative to the baseline speed of learning on such tasks.

While it was more representative of the behaviour analytic roots from which the FAST was developed than social cognitivist response time-based metrics, the SoR scoring method was somewhat rudimentary. Researchers concluded after numerous in-house experiments that the baseline blocks did not serve their intended purpose. The novel stimuli employed in baseline blocks resulted in slower acquisition than on critical blocks involving familiar stimuli with established or assumed salient stimulus functions. In contrast, it had been expected that functional response class acquisition rates on baseline blocks would fall somewhere between that of the consistent and inconsistent critical block. This was consistently not the case. Furthermore, the SoR index was not refined enough to distinguish sufficiently between performances. Specifically, consider the requirement for nine successive responses in any run of ten consecutive trials. An individual who has correctly responded on eight trials and who then goes on to make two errors, must be exposed to at least a further ten trials in order to complete the training block. They may, in fact, be required to complete a large number of runs of ten trials, on each occasion failing to satisfy the block completion criterion on the basis of a single error. In effect, the trial requirement metric for this participant is not reflective of near perfect fluency being demonstrated repeatedly across the

block and is a number that is being inflated exponentially with every repetition of the near perfect performance. In contrast, a participant who makes nine correct responses out of a run of ten consecutive trials on the first occasion after which they had produced eight consecutively correct responses, will achieve a considerably lower trial requirement score, despite having a marginally different fluency level in their performance compared to the first participant. Consequently, inter-individual trial requirements varied wildly, producing an unfavorable degree of noise in the data.

To address the crudeness of the SoR measure, Cartwright et al. (2016) developed an alternative scoring method, known as the Block Slope Score (BSS) method. To calculate the BSS, a cumulative response rate for each block is plotted on a graph to which a regression line is fit. A difference score representing the difference in the rates of learning across the blocks is then calculated by subtracting the slope of the learning curve for the inconsistent block from that of the consistent block. The BSS method effectively enabled in-vivo measurement of learning, and was more nuanced than its' predecessor. Moreover, calculation of the BSS necessitated a finite number of trials for each block, thus addressing the noise issue with the SoR index, and shortening the completion time of the FAST.

Despite offering more nuance and sensitivity than the SoR index, the BSS metric was not without its' flaws. Slope scoring did not control for sequences of rapid random responding that, by definition, produce a response accuracy rate approaching 50%, but at an increasing speed, based purely on the rapidity of the random responding, rather than the degree of stimulus control exerted over responding. Cummins et al. (2018) argued that this method could be improved upon by instead calculating a simple fluency score for each block based on the number of correct responses per minute. Importantly, however, this score would be penalized by the number of incorrect responses per minute produced by the participant, in order to control for the possibility of rapid random responding in the production of a fluency

metric. More specifically, this method involves subtracting the rate of incorrect responses per minute (IRPM) from the rate of correct responses per minute (CRPM) on each block to calculate a response fluency score for each block. A response fluency differential (RFD) score is then created by subtracting the inconsistent block fluency score from that for the consistent block.

The RFD metric is more mathematically transparent than the BSS measure. That is, while the slope measure has excellent face validity at a conceptual level from a behavioural perspective, it involves a degree of abstraction, in that it provides a metric that is not visible in the raw data. Given that one of the criticisms of the IAT levied by the developers of the FAST method concerned the mathematical sophistication and indeed obfuscation involved in calculating the D-score (c.f. O'Reilly et al. 2012, but see also Greenwald et al., 2003), the RFD method may be preferable in terms of its elegant simplicity.

Though an in-depth analysis of the advantages and disadvantages of each scoring approach are beyond the scope of this thesis, it would be a worthwhile consideration to compare the two methods systematically across studies and using large samples. While the reader may be wondering why this is not achieved within the current thesis, it is partly because the software coded for the current research did not include the recording of the relevant metrics and dynamics of responding to easily allow for the calculation of the Block Slope Score. But it is also because of a lack of conceptual commitment to the previous scoring method. That is, a central pillar of the current research agenda is that no one measure be treated as superior to others based on seniority alone, but should be selected based on proven utility in specific research contexts. It is far too early in the emergence of this method to make such a choice at this point but as noted above, a dedicated research agenda to examine this issue is warranted.

5.6 Conclusion

As modern behaviour analysis wades into research territory involving group designs and hypothesis testing, the field will inevitably inherit some of the necessary data analytic and procedural artifacts of such research designs. Specifically, behaviour analysts will be obliged to use methods of inferential statistics with which we are not always comfortable, and in order to facilitate this, engage in data cleaning methodologies involving, for instance, the removal of outliers with what ultimately will be arbitrary criteria (e.g., 1 SD, 2 SD, etc.). The larger sample sizes required to ensure sufficient statistical power within experimental group designs will also somewhat reduce the traditional behaviour-analytic focus on individual participant responses, and our ability to horn stimulus control on a participant-by-participant basis. In addition, the control of stimuli in terms of pre-existing stimulus functions will not be possible to assess in idiosyncratic methodologies with great ease, and so it may be more prudent in many cases to use standardized stimulus databases already assessed in terms of their stimulus functions, sometimes in a psychometric fashion. This possibility was employed in the current study in terms of the unconditioned stimuli, but not in terms of the conditioned stimuli, a feature which in hindsight, could have been achieved with ease and would have considerably improved the quality of the data obtained. While these extra burdens will be unfamiliar to many behaviour analysts, and will be approached with caution, they bear the promise of increased generalizability of our research findings. While the traditional behavioural approach attempts to apply core behavioural processes to all participants within a given cohort, this has been achieved at the cost of generalized principles or psychological "effects" with broad applicability, albeit reduced precision. It is for this reason, that one of the core aims of contextual behavioural science (the umbrella philosophical movement within which much of modern behaviour analysis occurs), is to achieve behavioural influence (rather than strict behavioural control) with *sufficient scope and precision*. The term "sufficient"

here is pointed, in that degrees of precision will be decided on a case-by-case basis, in terms of the needs of the study and the scientific goals of the scientist (see Hayes et al., 2021; Vilardaga et al., 2009). In effect, The nuanced task of the modern behaviouralist is to balance behavioural control with generating generalizable data. This will be no mean feat, and will involve a protracted research program, linking data gathered using idiographic methods and small sample research with effects observed at the group level, ensuring empirical and conceptual consistency between both levels at all times.

Similarly, as research is more frequently conducted online, it is worthwhile to consider the various strategies one may undertake to enhance data quality and exert as much behavioural control as is possible. Such strategies to enhance attention-specific tasks include employing in-line attention checks that are routinely employed in offline research (e.g., Gough et al. 2012). In addition, in order to protect the integrity of response time measures, researchers may need to understand the responsiveness of particular web browsers in the administration of particular procedures over particular types of networks.

Online research also raises the opportunity to gather large amounts of data in a short period of time by remunerating professional or semi-professional research participants, thus offering huge potential benefit to researchers. While this may raise ethical and procedural consternations for some researchers, the increasing use of these methods is not materializing concerns regarding either data quality or the exploitation of participants (see Palan & Schitter, 2017; Peer et al., 2017).

The foregoing suggests a convergence in methodologies between behaviour analysis and other fields who have already pioneered the use of online methodologies and large research samples, even within the realm of implicit testing. As long as we are mindful of the functional roots of our particular approach to behaviour (c.f., Hughes et al., 2012), we can

quite literally have our cake and eat it in this regard. The use of methodologies more recognizable to psychologists in other fields can only increase the opportunity for collaboration and the impact of our research publications.

In summary, we can conclude with some confidence that the FAST is sensitive to generalized conditioned evaluations that have been established for innocuous stimuli, notwithstanding methodological limitations. Data trends across the three experiments also suggest that the FAST is sensitive to the salience of an unconditioned stimulus during a brief evaluative conditioning procedure. In short, it would appear therefore, that while the current research was basic in nature, it does speak to the possibility that implicit tests of this kind can be used to index the intensity of an unconditioned stimulus as employed in a brief or casual real-world evaluative conditioning procedure. For example, an individual who has been bitten by a dog, and subsequently becomes fearful of them, will display an aversion for dogs in self-reports, or on implicit tests, that may be a function of the salience of the aversive experience caused by the dog (i.e., pain and physiological arousal). While exploring the clinical implications of such possibilities is beyond the scope of the current thesis, it raises an interesting and tantalizing possibility that such tools could be used as an adjunct in clinical assessment, alongside interview and self-report techniques, in assessing the impact of a traumatic stimulus on the behaviour of a client. More specifically, following a trauma, measures such as the FAST can give a relatively objective index of the degree of trauma produced by an unconditioned stimulus on the basis that test scores from such procedures should increase with increasing unconditioned stimulus intensity when used to assess the valence of such stimuli. Previous research has already established that implicit tests can be used to index the presence or absence of a phobic fear (Teachman, 2007; Teachman et al., 2001), and correlate with self-report phobic fear. Other studies have used the IAT as a measure of general anxiety and showed it is a good predictor of anxious behaviours and self-

reports of anxiety (Egloff & Schmukle, 2002). However, no study to date has demonstrated, under laboratory conditions, that the scores of an implicit test increase step wise with the laboratory-controlled increase in the intensity of the emotional experience which the implicit test is being used to index. In the current research, there was no effort to made to standardize the FAST scoring scale or to approach the measure as psychometric. Nevertheless, this research is highly progressive in providing a further dive into the fundamental behavioural processes underlying the FAST effect, and in doing so, conform to the signature characteristics of the modest FAST research agenda. To this extent, the questions outlined at the beginning of this thesis have been fairly roundly addressed and the path has been laid for a further investigation of these ideas along basic, as well as applied research lines.

References

Amd, M. & Passarelli, D. A. (2020). Dissociating preferences from evaluations following subliminal conditioning. *Acta Psychologica, 204* (103023). 1-10. https://doi.org/10.1016/j.actpsy.2020.103023

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N. & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behaviour Research Methods, S3,* 1407-1425. https://doi.org/10.3758/s13428-020-01501-5.

Arntzen, E., Eilertsen, J. M., & Fagerstrøm, A. (2016). Preferences in equivalence classes by low potency benign valenced stimuli. *European Journal of Behaviour Analysis, 17*(2), 142-153. https://doi.org/10.1080/15021149.2016.1247637

Arntzen, E., Norbom, A., & Fields, L. (2015). Sorting: An alternative measure of class formation? *The Psychological Record, 65*(4), 615-625. https://doi.org/10.1007/s40732-015-0132-5

Babchishin, K.M., Nunes, K.L. & Hermann, C.A. (2013). The Validity of Implicit Association Test (IAT) Measures of Sexual Attraction to Children: A Meta-Analysis. *Archives of Sexual Behaviour 42,* 487–499. https://doi.org/10.1007/s10508-012-0022-8

Baeyens, F., Eelen, P., Crombez, G. & van den Bergh, O. (1992). Human evaluative conditioning: Acquisition trials, presentation schedule, evaluative style, and contingency awareness. *Behaviour Research and Therapy, 30*(2). 133-142. https://doi.org/10.1016/0005-7967(92)90136-5

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie, 48*(2), 145-160. https://psycnet.apa.org/doi/10.1026/0949-3946.48.2.145

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., Waldron, D., Barnes-Holmes, Y., & Stewart, I. (2009). Testing the validity of the Implicit Relational Assessment Procedure and the Implicit Association Test: Measuring attitudes toward Dublin and country life in Ireland. *The Psychological Record, 59*(3), 389-406. https://doi.org/10.1007/BF03395671

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010a). 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., Murtagh, L., Barnes-Holmes, Y., & Stewart, I. (2010b). 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

Barnes-Holmes, Y., McEnteggart, C., & Barnes-Holmes, D. (2020a). Recent conceptual and empirical advances in RFT: Implications for developing process-based assessments and interventions for human psychological suffering. In M. E. Levin, M. P. Twohig, & J. Krafft (Eds.). *Innovations in acceptance and commitment therapy: Clinical advancements and applications in ACT.* Raincoast Books, 41-53.

Barnes-Holmes, D., Barnes-Holmes, Y. & McEnteggart, C. (2020b). Updating RFT (More field than frame) and its implications for process-based therapy. *The Psychological Record, 70*(4), 605-624. https://doi.org/10.1007/s40732-019-00372-3

Barnes-Holmes, D., & Harte, C. (2022). The IRAP as a Measure of Implicit Cognition: A Case of Frankenstein's Monster. *Perspectives on Behaviour Science, 45*, 559-578. https://doi.org/10.1007/s40614-022-00352-z

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical Foundations of Neuroscience.* Wiley-Blackwell.

Bentall, R. P., Jones, R. M., & Dickins, D. W. (1999). Errors and response latencies as a function of nodal distance in 5-member equivalence classes. *The Psychological Record, 49*(1), 93-115. https://doi.org/10.1007/BF03395309

Bevins, R. A., McPhee, J. E., Rauhut, A. S. & Ayres, J. J. B. (1997). Converging evidence for one-trial context fear conditioning with an immediate shock: Importance of shock potency. *Journal of Experimental Psychology: Animal Behaviour Processes, 23*(3). 312-324.

Binder, C. (1996). Behavioural fluency: Evolution of a new paradigm. *The Behaviour Analyst, 19*(2), 163-197. https://doi.org/10.1007/BF03393163

Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology, 69*(1), 62-79. https://doi.org/10.1111/bmsp.12059

Bortoloti, R., & De Rose, J. C. (2009). Assessment of the relatedness of equivalent stimuli through a semantic differential. *The Psychological Record, 59*(4), 563-590. https://doi.org/10.1007/BF03395682

Bortoloti, R., Rodrigues, N. C., Cortez, M. D., Pimentel, N., & de Rose, J. C. (2013). Overtraining increases the strength of equivalence relations. *Psychology & Neuroscience, 6*(3), 357-364. https://doi.org/10.3922/j.psns.2013.3.13

Boyle, S. Roche, B., Dymond, S. & Hermans, D. (2016). Generalisation of fear and avoidance along a semantic continuum. *Cognition and Emotion, 30*(2), 340-352. https://dx.doi.org/10.1080.02699931.2014.1000831

Brown, B. R., and Lathrop, R. L. (1971). "The Effects of Violations of Assumptions Upon Certain Tests of the Product Moment Correlation Coefficient."

Buades-Sitjar, F., Planchuelo, C. & Duñabeitia, J. A.(2021). Valence, Arousal and Concreteness Mediate Word Association. *Psicothema 33*(4), 602-609. https://doi.org/10.7334/psciothema2020.484.

Buchanan, E. M., Valentine, K. D. & Maxwell, N. P. (2019). English semantic feature production norms: an extended database of 4436 concepts. *Behaviour Research Methods, 51,* 1849-1863. https://doi.org/10.3758/s13428-019-01243-z

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

Castelli, L., & Tomelleri, S. (2008). Contextual effects on prejudiced attitudes: When the presence of others leads to more egalitarian responses. *Journal of Experimental Social Psychology, 44*(3), 679-686. https://doi.org/10.1016/j.jesp.2007.04.006

Catania, A. C., Horne, P., & Lowe, C. F. (1989). Transfer of function across members of an equivalence class. *The Analysis of Verbal Behaviour, 7*(1), 99-110. https://doi.org/10.1007/BF03392841

Church, R. M., & Raymond, G. A. (1967). Influence of the schedule of positive reinforcement on punished behaviour. *Journal of Comparative and Physiological Psychology, 63*(2), 329–332. https://doi.org/10.1037/h0024382

Ciarrochi, J., Brockman, R., Duguid, J., Parker, P., Sahdra, B., & Kashdan, T. (2016). Measures that make a difference: Optimizing psychological measurement to promote wellbeing and reduce suffering. In R. Zettle, S. Hayes, T. Biglan, & D. Barnes-

Holmes (Eds.), *Handbook of Contextual Behavioural Science* (pp. 320–347). John Wiley

Cohen J (1988). *Statistical Power Analysis for the Behavioural Sciences*, 2nd ed. Lawrence Erlbaum.

Cousineau, D. & Chartier, S. (2010). Outliers detection and treatment. *International Journal of Psychological Research, 3*(1), 58-67.

Cummins, J., & Roche, B. (2020). Measuring differential nodal distance using the function acquisition speed test. *Behavioural Processes, 178*, 104179. https://doi.org/10.1016/j.beproc.2020.104179

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 Behaviour, 110*(1), 24-38. https://doi.org/10.1002/jeab.437

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

De Houwer, J. (2002). The Implicit Association Test as a tool for studying dysfunctional associations in psychopathology: Strengths and limitations. *Journal of Behaviour Therapy and Experimental Psychiatry, 33*(2), 115-133. https://doi.org/10.1016/s0005-7916(02)00024-1

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation, 37*(2), 176–187. https://doi.org/10.1016/j.lmot.2005.12.002

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behaviour, 37*(1), 1-20. https://doi.org/10.3758/lb.37.1.1

De Houwer, J. (2019). Implicit bias is behaviour: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science, 14*(5), 835-840. https://doi.org/10.1177/1745691619855638

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology, 24*(1), 252-287. https://doi.org/10.1080/10463283.2014.892320

De Leeuw, J. R. & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual

search task. *Behavioural Research 48*. 1-12. https://doi.org/10.3758/s13428-015-0567-2

Dinsmoor, J. A. (1995). Stimulus Control: Part I. *The Behaviour Analyst, 18*(1), 51-68. https://doi.org/10.1007/bf03392691

Dougher, M. J., Auguston, E., Markham, M. R., Greenway, D. E., Wulfert, E. (1994). The transfer of respondent conditioning eliciting and extinction functions through stimulus equivalence classes. *Journal of the Experimental Analysis of Behaviour, 62*(3). 331-351. https://doi.org/10.1901/jeab.1994.62-331

Doughty, A. H., Brierley, K. P., Eways, K. R., & Kastner, R. M. (2014). Effects of stimulus discriminability on discrimination acquisition and stimulus-equivalence formation: Assessing the utility of a multiple schedule. *The Psychological Record, 64*(2), 287-300. https://doi.org/10.1007/s40732-014-0001-7

Doughty, A. H., & Soydan, J. A. (2019). Differential derived stimulus relations across probe-trial versus adduction testing are not a function of comparison-stimulus presentation. *Behavioural Processes, 166,* 103903. https://doi.org/10.1016/j.beproc.2019.103903

Egloff, B. & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of Personality and Social Psychology, 83*(6), 1441-1455. https://doi.org/10.1037//0022-3514.83.6.1441.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology, 17*(1), 74-147. https://psycnet.apa.org/doi/10.1080/10463280600681248

Field, A. P. & Moore, A. C. (2005). Dissociating the effects of attention and contingency awareness on evaluative conditioning effects in the visual paradigm. *Cognition & Emotion, 19*(2). 217-243. https://doi.org/10.1080/02699930441000292

Fields, L., (2015). Stimulus relatedness in equivalence classes, perceptual categories, and semantic memory networks. *European Journal of Behaviour Analysis, 17*(1), 2–18. https://doi.org/10.1080/15021149.2015.1084713

Fields, L., Adams, B. J., Verhave, T., & Newman, S. (1990). The effects of nodality on the formation of equivalence classes. *Journal of the Experimental Analysis of behaviour, 53*(3), 345-358. https://doi.org/10.1901/jeab.1990.53-345

Fields, L., Arntzen, E., & Moksness, M. (2014). Stimulus sorting: A quick and sensitive index of equivalence class formation. *The Psychological Record, 64*(3), 487- 498. http://dx.doi.org/10.1007/s40732-014-0034-y

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

Fields, L., Landon-Jimenez, D. V., Buffington, D. M., & Adams, B. J. (1995). Maintained nodal-distance effects in equivalence classes. *Journal of the Experimental Analysis of Behaviour, 64*(2), 129-145. https://doi.org/10.1901/jeab.1995.64-129

Finn, M. (2020). Exploring the dynamics of arbitrarily applicable relational responding with the implicit relational assessment procedure (Doctoral dissertation, Doctoral thesis). Ghent University, Belgium). https://hdl.handle.net/1854/LU-8654041

Gast, A., Langer, S. & Sengewald, M. (2016). Evaluative conditioning increases with temporal contiguity . The influence of stimulus order and stimulus interval on evaluative conditioning. *Acta Psychologia, 170.* 177-185. https://dx.doi.org/10.1016/j.actpsy.2016.07.002

Gavin, A., Roche, B., & Ruiz, M. R. (2008). Competing contingencies over derived relational responding: A behavioural 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 behaviour analytically modified implicit association test for measuring sexual categorization of children. *The Psychological Record, 62*(1), 55-68. https://doi.org/10.1007/BF03395786

Glaser, T. & Kuchenbrandt, D. (2017). Generalisation effects in evaluative conditioning: Evidence for attitude transfer effects from single exemplars to social categories. *Frontiers in Psychology, 8*(103). https://doi.org/10.3389/fpsyg.2017.00103

Goffman, E. (1959). *The Presentation of Self in Everyday Life.* Penguin

Gough, P. M., Riggio, L., Chersi, F., Sato, M., Fogassi, L., Buccino, G. (2012). Nouns referring to tools and natural objects differentially modulate the motor system. *Neuropsychologia, 50*(1), 19-25. https://doi.org/10.1016/j.neuropsychologia.2011.10.017

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4-27. https://doi.org/10.1037//0033-295x.102.1.4

Greenwald, A. G., & Breckler, S. J. (1985). To whom is the self presented: The self and social life, 126, 145.

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. https://doi.org/10.1037//0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*(2), 197. https://psycnet.apa.org/doi/10.1037/0022-3514.85.2.197

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., & Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 US presidential election. *Analyses of Social Issues and Public Policy, 9*(1), 241-253. https://psycnet.apa.org/doi/10.1111/j.1530-2415.2009.01195.x

Gregg, A. P., Seibt, B. & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90*(1), 1-20. https://doi.org/10.1037/0022-3514.90.1.1

Grey, I. M., & Barnes, D. (1996). Stimulus equivalence and attitudes. *The Psychological Record, 46*(2), 243-270.

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://psycnet.apa.org/doi/10.1037/0096-3445.132.2.266

Hansen, M., Schoonover, A., Skarica, B., Harrod, T., Bahr, N., & Guise, J. M. (2019). Implicit gender bias among US resident physicians. *BMC Medical Education, 19*(1), 19-396. https://doi.org/10.1186/s12909-019-1818-1

Havlicek, L. L., & Peterson, N. L. (1976). Robustness of the Pearson Correlation against Violations of Assumptions. *Perceptual and Motor Skills, 43*(3_suppl), 1319–1334. https://doi.org/10.2466/pms.1976.43.3f.1319

Hayes, S. C., Merwin, R. M., McHugh, L., Sandoz, E. K., A-Tjak, J. G. L., Ruiz, F. J., Barnes-Holmes, D., Bricker, J. B., Ciarrochi, J., Dixon, M. R., Po-Lun Fung, K., Gloster, A. T., Gobin R. L., Gould, E. R., Hofmann, S. G., Kasujja, R., Karekla, M.,

Luciano, C. & McCracken, L. M. (2021). Report of the ACBS Task Force on the strategies and tactics of contextual behavioural science research. *Journal of Contextual Behavioural Science, 20,* 172-183. https://doi.org/10.1016/j.jcbs.2021.03.007

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://psycnet.apa.org/doi/10.1037/0033-2909.109.2.340

Hoffman, W., De Houwer, J., Perugini, M., Baeyens, F. & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin, 136*(3), 390-421. https://doi.org/10.1037/a0018916.

Holth, P., & Arntzen, E. (1998). Stimulus familiarity and the delayed emergence of stimulus equivalence or consistent non-equivalence. *The Psychological Record, 48*(1), 81-110.

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioural alternatives. *The Psychological Record, 61*(3), 465-496. https://doi.org/10.1007/BF03395772

Hughes, S., Barnes-Holmes, D. & Vahey, N. (2012). Holding onto our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. *Journal of Contextual Behavioural Science, 1.* 17-38. http://dx.doi.org/10.1016/j.jcbs.2012.09.003.

Hughes, S, & Barnes-Holmes, D. (2013). A functional approach to the study of implicit cognition: The IRAP and the REC model. In B. Roche & S. Dymond. (Eds.). *Advances in Relational Frame Theory & Contextual Behavioural Science: Research & Applications* (pp. 97-126).

Hughes, S., Barnes-Holmes, D. & Smyth, S. (2017). Implicit cross-community biases revisited: Evidence for ingroup favoritism in the absence of outgroup derogation in Northern Ireland. *Psychological Record, 67,* 97-107. https://doi.org/10.1007/s40732-016-0210-3.

Hussey, I. (2020). The IRAP is not suitable for individual use due to very wide confidence intervals around D scores. https://doi.org/10.31234/osf.io/w2ygr

Hussey, I., Mhaoileoin, D. N., Barnes-Holmes, D., Ohtsuki, T., Kishita, N., Hughes, S., & Murphy, C. (2016). The IRAP is nonrelative but not acontextual: Changes to the contrast category influence men's dehumanization of women. *The Psychological Record, 6*6(2), 291-299 .https://psycnet.apa.org/doi/10.1007/s40732-016-0171-6

Hussey, I., Thompson, M., McEnteggart, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2015). Interpreting and inverting with less cursing: A guide to interpreting IRAP data. *Journal of Contextual Behavioural Science, 4*(3), 157-162. https://psycnet.apa.org/doi/10.1016/j.jcbs.2015.05.001

Ioannidis, J. P. A. (2005). Why most research findings are false. *PLOSMedicine, 19*(8), e1004085. https://doi.org/1-.1371/journal.pmed.1004085

Jost, J. T. (2019). The IAT is dead, long live the IAT: Context-sensitive measures of implicit attitudes are indispensable to social and political psychology. *Current Directions in Psychological Science, 28*(1), 10-19. https://doi.org/10.1177/0963721418797309

Kattner, F. (2014). Reconsidering the (in)sensitivity of evaluative conditioning to reinforcement density and CS-US contingency. *Learning and Motivation, 45.* 15-29. https://dx.doi.org/10.1016/j.lmot.2013.09.002

Karpinski, A., & Steinman, R. B. (2006). The single category implicit association test as a measure of implicit social cognition. *Journal of Personality and Social Psychology, 91*(1), 16-32. https://psycnet.apa.org/doi/10.1037/0022-3514.91.1.16

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. *Quarterly Journal of Experimental Psychology, 63*(3), 595-619. https://psycnet.apa.org/doi/10.1080/17470210903076826

Klein, C. (2020). Confidence intervals on implicit association test scores are really rather large. https://doi.org/10.31234/osf.io/5djkh

Kohlenberg, R. J., Hayes, S. C., & Tsai, M. (1993). Radical behavioural psychotherapy: Two contemporary examples. *Clinical Psychology Review, 13*(6), 579-592. https://psycnet.apa.org/doi/10.1016/0272-7358(93)90047-P

Laekens, D. (2022). Sample size justification. *Collabra: Psychology, 8*(1). https://doi.org/10.1525/collabra.33267

Bradley, M. M., & Lang, P. J. (2008). The International Affective Picture System (IAPS) in the study of emotion and attention. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 29–46). Oxford University Press.

Leslie, J. C., Tierney, K. J., Robinson, C. P., Keenan, M., et al. (1993). Differences between clinically anxious and non-anxious subjects in a stimulus equivalence training task involving threat words. *The Psychological Record, 43,* 153–161.

McGlinchey, A., Keenan, M. & Dillenburger, K. (2000). Outline for the development of a screening procedure for children who have been sexually abused. *Research on Social Work Practice, 10*(6), 721-747.

McLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological bulletin, 109*(2), 163-203. https://psycnet.apa.org/doi/10.1037/0033-2909.109.2.163

McLoughlin, S. & Roche, B. T. (2022). ACT: A Process-Based Therapy in search of a process, *Behaviour Therapy* https://doi.org/10.1016/j.beth.2022.07.010

McRae, K., Cree, G., Seidenberg, A. S. & McNorgan, C. (2005). Semantic feature production norms for a large set of living and non-living things. *Behaviour Research Methods, Instruments & Computers, 37*(4). 547-559. https://doi.org/10.3758/BF03192726

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

Mizael, T. M., de Almeida, J. H., Silveira, C. C., & de Rose, J. C. (2016). Changing racial bias by transfer of functions in equivalence classes. *The Psychological Record, 66*(3), 451-462. https://doi.org/10.1007/s40732-016-0185-0

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 Behaviour, 95*(3), 343-368. https://doi.org/10.1901%2Fjeab.2011.95-343

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation/

Nevin, J. A., & Grace, R. C. (2000). Behavioural momentum and the law of effect. *Behavioural and Brain Sciences, 23(*1), 73-90. https://doi.org/https://doi.org/10.1017/s0140525x00002405

Noel, J. G., Petzel, Z. W. & Mulderig, T. H. (2019). Of two minds about alcohol: Specific effects of evaluative conditioning on implicit, but not explicit , alcohol cognitions among heavy versus light drinkers. *Psychology of Addictive Behaviours, 33*(3), 285-296. http://dx.doi.org/10.1037/adb0000449

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math= male, me= female, therefore math≠ me. *Journal of Personality and Social Psychology, 83*(1), 44-59. https://psycnet.apa.org/doi/10.1037/0022-3514.83.1.44

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at Age 7: A Methodological and Conceptual Review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). Psychology Press.

Olson, M. A., & Fazio, R. H. (2001). Implicit Attitude Formation Through Classical Conditioning. Psychological Science, 12(5), 413-417. https://doi.org/10.1111/1467-9280.00376

Olson, M. A. & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32*(4), 421-433. https://doi.org/10.1177/0146167205284004

O'Reilly, A., Roche, B., & Cartwright, A. (2015). Function over form: A behavioural approach to implicit attitudes. In Z. Jin (Ed), *Exploring implicit cognition: learning, memory, and social cognitive processes* (pp. 162-182). IGI Global. https://doi.org/10.4018/978-1-4666-6599-6.ch008

O'Reilly, A., Roche, B., Gavin, A., Ruiz, M. R., Ryan, A., & Campion, G. (2013). A function acquisition speed test for equivalence relations (FASTER). *The Psychological Record, 63*(4), 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 behaviour analytic implicit test for assessing stimulus relations. *The Psychological Record, 62*(3), 507-528. https://doi.org/10.1007/bf03395817

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17,* 776-783.

Osborne, J. W. & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research, and Evaluation, 9*(6). https://doi.org/10.7275/qf69-7k43

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://psycnet.apa.org/doi/10.1037/a0032734

Palan, S. & Schitter, C. (2017). Prolific.ac- A Subject pool for online experiments. *Journal of Behavioural and Experimental Finance, 17,* 22-27. https://doi.org/10.1016/j.jbef.2017.12.004

Page, M. M., & Lumia, A. R. (1968). Cooperation with demand characteristics and the bimodal distribution of verbal conditioning data. *Psychonomic Science, 12*(6), 243-244. https://doi.org/10.3758/bf03331291

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioural research. *Journal of Experimental Social Psychology, 70,* 153-163. https://doi.org/10.1016/j.jesp.2017.01.006

Perkins, A., Forehand, M. R., & Greenwald, A. G. (2006). Decomposing IAT-measured self-associations: the relative influence of semantic meaning and valence. *Social Cognition, 24*(4), 387-408. https://doi.org/10.1521/soco.2006.24.4.387

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. & Martini, J. R. (1999). The respondent conditioning of male sexual arousal. *Behaviour Modification, 23*(2), 254-268. https://doi.org/10.1177.0145445599232004

Popper K. R. (1959). *The Logic of Scientific Discovery.* Basic Books.

Power, P. M., Harte, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2017). Exploring Racial Bias in a European Country with a Recent History of Immigration of Black Africans. *The Psychological Record, 67*(3), 365-375. https://doi.org/10.1007/s40732-017-0223-6

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://psycnet.apa.org/doi/10.1080/14640747708400645

Rachuri, K. K., Musolesi, M, Mascolo, C., Rentfrow, P. J., Longworth, C. & Aucinas, A. (2010). EmotionSense: A mobile phones based adaptive platform for experimental social psychology research. *Proceedings of the 12th ACM International Conference on Ubiquitous Computing.* 281-290. https://dx.doi.org/10.1145/1864349.1864393

Reimers, S., & Stewart, N. (2008). Using Adobe Flash Lite on mobile phones for psychological research: Reaction time measurement reliability and inter-device variability. *Behaviour Research Methods, 40*(4), 1170-1176. https://doi.org/10.3758/BRM/40.4.1170.

Reinecke, K. & Gajos, K., Z. (2015). LabintheWild: Conducting large-scale online experiments with uncompensated samples. *Community-Based Participatory Research,* 1364-1378. https://dx.doi.org/10.1145/2675133.2675246.

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

Robinson, M. D., Meier, B. P., Zetocha, K. J., & McCaul, K. D. (2005). Smoking and the Implicit Association Test: When the Contrast Category Determines the Theoretical Conclusions. *Basic and Applied Social Psychology, 27*(3), 201-212. https://doi.org/10.1207/s15324834basp2703_2

Roche, B., & Barnes, D. (1997). A transformation of respondently conditioned stimulus function in accordance with arbitrarily applicable relations. *Journal of the Experimental Analysis of Behaviour, 67*(3), 275-301. https://doi.org/10.1901%2Fjeab.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 behaviour. *The Behaviour Analyst, 25*(1), 75-91. https://doi.org/10.1007/BF03392046

Roche, B., O'Reilly, A., Gavin, A., Ruiz, M., & Arancibia, G. (2012). Using behaviour-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., O'Riordan, M., & Hand, K. (2005). A relational frame approach to the psychological assessment of sex offenders. In: Taylor, M. & McQuale, E. (Eds.) *Viewing child pornography on the Internet: Understanding the offence, managing the offender, and helping the victims* (109-125). Russel House Publishing

Roddy, S., Stewart, I. & Barnes-Holmes, D. (2011). Anti-fat, pro-slim, or both? Using two reaction-time based measures to assess implicit attitudes to the slim and overweight. *Journal of Health Psychology, 15*(3). 416-425. https://doi.org/10.1177/1359105309350232.

Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research, 21*(4), 37-59. https://doi.org/10.1300/J079v21n04_02

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

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology 13*(2), 90-100. https://doi.org/10.1037/a0015108

Shapin, S. & Schaffer, S. (1985). *Leviathan and the Air-Pump. Hobbes, Boyle, and the experimental life.* Princeton, NJ: Princeton University Press.

Sidman, M. (1960). *Tactics of Scientific Research: Evaluating experimental data in psychology*. Basic Books.

Sidman, M., Kirk, B., & Willson-Morris, M. (1985). Six-member stimulus classes generated by conditional-discrimination procedures. *Journal of the Experimental Analysis of Behaviour, 43*(1), 21-42. https://doi.org/10.1901/jeab.1985.43-21

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of behaviour, 37*(1), 5-22. https://doi.org/10.1901/jeab.1982.37-5

Skinner, B. F. (1976). Farewell, my LOVELY! *Journal of the Experimental analysis of Behaviour, 25*(2), 218. http://doi.org/10.1901/jeab.1976.25-218

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18*(6), 643-662. https://psycnet.apa.org/doi/10.1037/h0054651

Teachman, B. A. (2007). Evaluating implicit spider fear associations using the Go/No-go Association Task. *Journal of Behaviour Therapy and Experimental Psychiatry, 38*(2), 156-167. https://doi.org/10.1016/j.jbtep.2006.10.006

Teachman, B. A., Gregg, A. P. & Woody, S. R. (2001). Implicit associations for fear-relevant stimuli among individuals with snake and spider fears. *Journal of Abnormal Psychology, 110*(2), 226-235. https://doi.org/10.1037//0021-843x.110.2.226.

Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2010). *A practical guide to implicit association tests and related tasks*. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 117–139). The Guilford Press.

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

Tyndall, I. T., Roche, B., & James, J. E. (2009). The interfering effect of emotional stimulus functions on stimulus equivalence class formation: Implications for the understanding and treatment of anxiety. *European Journal of Behaviour Analysis, 10*(2), 215-234. https://doi.org/10.1080/15021149.2009.11434320

Van Dessel, P., de Houwer, J., Gast, A. & Tucker Smith, C. (2015). Instruction-based approach-avoidance effects: Changing stimulus evaluation via the mere instruction to approach or avoid stimuli. *Experimental Psychology, 62*(3). https://doi.org/10.1027/1618-3169/a000282

Varao-Sousa, T. L., Smilek, D. & Kingstone, A. (2018). In the lab and in the wild: How distraction and mind wandering affect attention and memory. *Cognitive Research: Principles and Implications, 3*(42). https://doi.org/10.1186/s41235-018-0137-0

Vilardaga, R., Hayes, S. C., Levin, M. E., & Muto, T. (2009). Creating a strategy for progress: A contextual behavioural science approach. *The Behaviour Analyst, 32*(1), 105-133. https://doi.org/10.1007/bf03392178

Vinson, D. P. & Vigilocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behaviour Research Methods, 40,* 183-190. https://doi.org/10.3758/BRM.40.1.183

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

Warriner, A.B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behaviour Research Methods, 45,* 1191-1207. https://doi.org/10.3758/s13428-012-0314-x

Weber, S. J. & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin, 77*(4), 273-295. https://doi.org.10.1037/h0032351

Whelan, R. (2008). Effective analysis of reaction time data. *The psychological record, 58*(3), 475-482. https://doi.org/10.1007/BF03395630

Appendix I- Rating Scales

Please look at each of the words below and rate it on the scale provided in terms of how much it reminds you of pleasant or unpleasant imagery. Click on the number that represents your choice.

Apple

Unpleasant Pleasant

1 2 3 4 5 6 7

Banana

Unpleasant Pleasant

1 2 3 4 5 6 7

Pear

Unpleasant Pleasant

1 2 3 4 5 6 7

Orange

Unpleasant Pleasant

1 2 3 4 5 6 7

Chair

Unpleasant Pleasant

1 2 3 4 5 6 7

Table

Unpleasant Pleasant

1 2 3 4 5 6 7

Sofa

Unpleasant Pleasant

1 2 3 4 5 6 7

Desk

Unpleasant Pleasant

1 2 3 4 5 6 7

Appendix 1I – Information Sheet

This research is being conducted by Aideen Watters (aideen.watters.2018@mumail.ie), a postgraduate student at the Department of Psychology, Maynooth University, under the supervision of Dr. Bryan Roche (contact: Bryan.T.Roche@mu.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 assessment procedure called the Function Acquisition Speed Test (FAST). This type of test is known as an "implicit test" insofar as it can function as a very indirect measure of your evaluations of any word or image and is relatively difficult to deceive.

The FAST works by measuring how you respond under instruction to pressing particular keys on a computer keyboard upon the presentation of a variety of items on screen (in this case words). Your response pattern can indicate a bias in favor of or against certain words or concepts.

The first phase will involve a brief learning experience in which images, some of which are potentially distressing (graphic images of bodily injuries) are briefly but repeatedly presented on a computer screen alongside various randomly chosen words. You will simply be required to notice which words tend to appear with which types of images. You are required to pay close attention at all times. If you stop attending to the task your data will be of no use to the researchers.

You will then be asked to rate your feelings (positive or negative) towards a set of English words, that are the same as or related to the ones you have been exposed to in the previous procedure. This will allow us to assess the effect of the learning experience on your feelings towards the words involved.

Finally, you will be asked to complete a brief task (the FAST) that involves you learning to

press specific computer keyboard buttons whenever particular words (such as those used in the learning experience) are presented on screen. This will also allow us to assess in a more indirect way your evaluations of the words involved in the initial learning experience.

If you have a history of anxiety-related issues that would make such a task unadvisable or upsetting, or if you would find viewing such images too distressing, then you should not participate in the study.

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 unique four-digit code. You should note this for proof of participation, if for any reason you wish to ask a question of the researchers.

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.

Please note that this research is best conducted on a desktop or a laptop computer in a quiet environment. This task requires concentration, and the data may be of no use to the researchers if it is conducted where there is any distraction or on a small device.

Participants must be over 18 years of age, and should not have a history of anxiety related issues that would make participation inadvisable (See above).

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@mu.ie or +353 (0)1 708 6019. Please be assured that your concerns will be dealt with in a sensitive manner.

Appendix III – Consent Form

By proceeding you are confirming that you have read and understood the information provided to you. You are also confirming that you are over the age of 18 years and do not suffer from any condition that would make exposure to aversive images harmful to you and that you are fully aware that some of the images that will be presented in this study are of the type that many people would find upsetting (i.e., images of bodily injuries). Finally, you are agreeing that you understand that it is not possible to withdraw your data following participation because all data is completely anonymous.

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 this page now. If at any point during experimentation you decide you no longer want to participate, you may leave, and your data will not be utilised.

At the end of testing, you will be provided with a unique four-digit code. You can use this code for proof of participation to earn course credit if you need it. A receipt can be provided by contacting the researcher.

Please note that this research is best conducted on a desktop or a laptop computer in a quiet environment. This task requires concentration, and the data may be of no use to the researchers if it is conducted where there is any distraction or on a small device.

Consent and Proceed ☐

Appendix IV Demographic Questionnaire

Please provide the following information to help us with this research

Q. What is your age in years?

Q. What is your sex?

- Male
- Female
- Non-Binary
- Prefer not to say

Q. Which of the following ethnicities describes you best?

- White
- Black
- Asian
- Arab
- Mixed
- Other

Appendix V– Debriefing

Thank you for taking the time to participate in this study. The purpose of this experiment was to test the hypothesis that a new implicit test called the FAST test is capable of measuring biases we form against words or other items after relatively brief emotional learning experiences. We were also interested in how the strength of these biases are related to the degree of aversiveness of the images used in these emotional learning experiences.

The researchers were particularly interested to see whether or not the results of your FAST test correlated with your ratings of the pleasantness of the words that were involved in the emotional learning experience, or of words that shared similar meanings.

The FAST works by measuring the rate at which an individual can learn to respond in the same way to words that do not share the same emotional meaning. That is, where an emotional learning experience is strong, individuals often learn more slowly to press the same computer keyboard button for aversive images and pleasant words. In simple terms, it is difficult for humans to learn to respond in the same way to things that are incompatible, and the more incompatible the items are the slower it is that humans learn to respond in the same way to these items. In effect, the FAST learning test gives us an index of how strongly biased an individual has become against thinking of certain types of images in a positive or negative way.

Should you have any questions or concerns about the study you can contact me at aideen.watters.2018@mumail.ie or my supervisor for this research Dr. Bryan Roche at Bryan.T.Roche@mu.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@mu.ie or +353 (0)1 708 6019. Please be assured that your concerns will be dealt with in a sensitive manner.

Appendix V

Appetitive Unconditioned Stimuli: *Condition 1*

Image No: 2345

Valence: 7.41, Arousal: 5.42

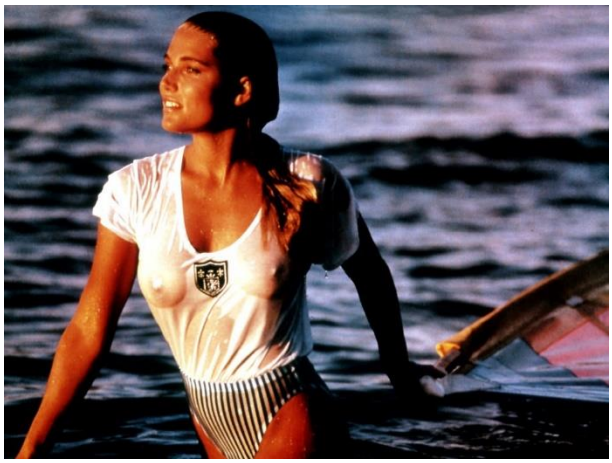


Image No: 4220

Valence: 6.60, Arousal: 5.18

Image No: 5260

Valence: 7.34, Arousal: 5.71

Image No: 7220

Valence: 6.91, Arousal: 5.30

Appendix VI

Aversive Unconditioned Stimuli- *Condition 1*

Image No: 7380

Valence: 2.46, Arousal: 5.88

Image No: 9400

Valence: 2.50, Arousal: 5.99

Image No: 9419

Valence: 2.82, Arousal: 5.10

Image No: 9500

Valence: 2.42, Arousal: 5.82

Appendix VII

Appetitive Unconditioned Stimuli- *Condition 2*

Image No: 2005

Valence: 6.00, Arousal: 4.07

Image No: 1947

Valence: 5.85, Arousal: 4.35

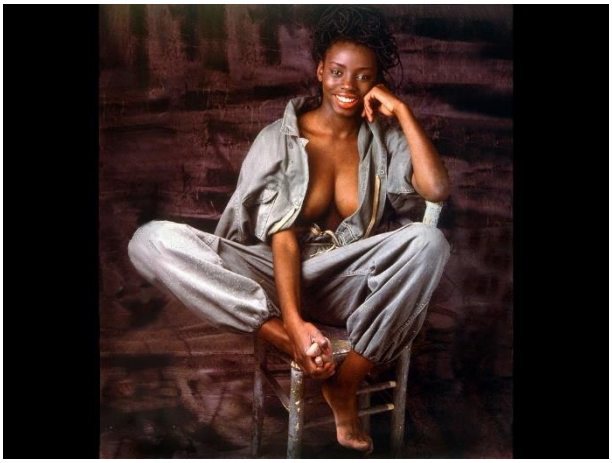


Image No: 4004

Valence: 5.14, Arousal: 4.44

Image No: 7402

Valence: 5.98, Arousal: 5.05

Appendix VIII

Aversive Unconditioned Stimuli- *Condition 2*

Image No: 1280

Valence: 3.66, Arousal: 4.93

Image No: 6561

Valence: 3.58, Arousal: 4.44

Image No: 7361

Valence: 3.10, Arousal: 5.09

Image No: 9404

Valence: 3.71, Arousal: 4.67

Appendix IX

Appetitive Unconditioned Stimuli- *Condition 3*

Image No: 2214

Valence: 5.01, Arousal: 3.46

Image No: 2396

Valence: 4.91, Arousal: 3.34

Image No: 2980

Valence: 5.61, Arousal: 3.09

Image No: 7484

Valence: 4.92, Arousal: 4.08

Appendix X

Aversive Unconditioned Stimuli- *Condition 3*

Image No: 1112

Valence: 4.71, Arousal: 4.60

Image No: 2752

Valence: 4.07, Arousal: 4.84

Image No: 6800

Valence: 4.02, Arousal: 4.87

Image No: 7484

Valence: 4.99, Arousal: 4.24