The Development of an Implicit Association Test for Measuring Forgiveness

Jeremy Goldring

Submitted January 2011 for the degree of Doctor of Philosophy,

in the School of Psychology, at the University of Adelaide

Contents

List of Tables

Abstract

The majority of psychological research on forgiveness has relied on self-report instruments as the primary mode of collecting data; there is a recognised need for alternative approaches to forgiveness measurement. This need is accentuated by the inherent/perceived 'prosocial' nature of forgiveness: there is a chance that people will self-report as more forgiving in order to present themselves as more socially desirable. Furthermore, a person may not always be consciously aware of their forgiving motivations, intentions or attitudes. Additionally, theorists typically frame forgiveness as a conscious, deliberate, controlled process, but it may also be comprised of more unconscious, spontaneous, and automatic components. Thus, self-report scales may be insufficient for exploring forgiveness. This thesis aimed to address these shortcomings by developing an Implicit Association Test (IAT: Greenwald, McGhee & Schwartz, 1998) suitable for the measurement of forgiveness. This forgiveness IAT was developed across 9 studies, with a total of 1304 participants. Studies 1 to 5 assessed the validity of the forgiveness IAT against several criteria: the choice of words and categories used to represent and compare with forgiveness; resistance to socially desirable responding; and convergence with self-reported forgiveness measures. Studies 6 to 9 assessed the forgiveness IAT's utility in predicting behavior, on the basis of recalled, hypothetical, and laboratory-based transgressions. Results suggest that the IAT is a valid, reliable, and useful measure of forgiveness attitudes, and may be able to predict some types of post-transgression behaviour that are not accounted for by existing self-report forgiveness measures. These findings will help psychologists to better understand the processes that drive forgiveness, particularly those operating at the automatic level.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Jeremy Goldring

Acknowledgements

This thesis would not have been possible without the help and guidance of the following people, to whom I owe many thanks:

Dr Peter Strelan, for always making the time for me (whether that was for a casual chat over coffee, giving detailed feedback on my drafts, or anything in between); for knowing when to challenge and when to encourage; and for generally being a top bloke. I couldn't have asked for more from my primary supervisor.

Dr Carolyn Semmler, for her sound advice, valuable (colour-coded) feedback, and for keeping me sensible.

Dr Ian McKee, for his wise and thoughtful words, and for his supreme generosity in providing last-minute feedback on a draft of the thesis.

Bob Willson and Mark Brown, for their regular assistance in setting up my studies online, and patience with me when it came to smoothing out the inevitable bugs.

Dr Matt Palmer, for really going out of his way to help me out, even after it was no longer part of his job description.

Emma Kemp, for providing swift feedback on my draft thesis.

The rest of the Forgiveness crew: Louise, Heather, and Letty, who have all been there for me in their own different ways.

Mum & Dad. For everything.

The 1304 participants who took part in my research.

And last, but certainly not least (not ever), Catherine. For encouraging and supporting me, and for helping me to keep it all in perspective.

Chapter 1:

Forgiveness and the Implicit Association Test

1.1 What is forgiveness?

"If one by one we counted people out
For the least sin, it wouldn't take us long
To get so we had no one left to live with.
For to be social is to be forgiving."
(Frost & Untermeyer, 1963, p. 160)

As so eloquently captured by poet Robert Frost, forgiveness is an integral part of being social creatures. Yet while forgiveness has been a matter of interest for philosophers, writers, poets and theologians for centuries, it has only been in the last two decades that researchers have started taking a serious interest in forgiveness as a construct within psychology (Fincham, 2000; Worthington, 2005a). Forgiveness research in psychology is now flourishing, being examined within and applied to a range of contexts including developmental (Denham, Neal, Wilson, Pickering, & Boyatzis, 2005; Hui & Chau, 2009), cultural (Kadima, Gauché, Vinsonneau, & Mullet, 2007; Sandage & Williamson, 2005; Suwartono, Prawasti, & Mullet, 2007), health (Friedberg, Suchday, & Srinivas, 2009; Toussaint & Webb, 2005b; Waltman et al., 2009; Webb, Toussaint, Kalpakjian, & Tate, 2010; Witvliet, Ludwig, & Vander Laan, 2001), education (Gassin, Enright, & Knutson, 2005) and policy development (Worthington, 2001).

The most significant attention has been paid to the domains of therapy (Enright & Fitzgibbons, 2000; Enright & Zell, 1989; McCullough & Worthington, 1994) and personality and social psychology (Emmons, 2000; McCullough & Hoyt, 2002; Thompson et al., 2005; Walker & Gorsuch, 2002). Psychologists now know much about the predictors of forgiveness, including offence-specific factors such as apology (Struthers, Eaton, Santelli, Uchiyama, & Shirvani, 2008), relational closeness (Bono, McCullough, & Root, 2008) and offence severity (Fincham, Jackson, & Beach, 2005), as well as person-

based factors such as empathy (McCullough et al., 1998), rumination (McCullough, Bono, & Root, 2007), attachment style (Burnette et al., 2007), and Big Five traits such as agreeableness, conscientiousness and neuroticism/emotionality (Balliet, 2010; Brose, Rye, & Lutz-Zois, 2005; McCullough & Hoyt, 2002; Walker & Gorsuch, 2002). We also know about some of the potential outcomes of forgiveness, especially those relating to psychological well-being (Ahadi & Ariapooran, 2009; Toussaint & Webb, 2005b), relationships (McCullough, 2008), and physical health (Friedberg et al., 2009; Webb et al., 2010; Whited, Wheat, & Larkin, 2010).

Despite this 'boom' in forgiveness research there is still some contention as to what forgiveness actually is. Psychologists generally agree that forgiveness is both an intrapersonal and interpersonal process[1] (Struthers, Eaton, Santelli et al., 2008). Despite this, the bulk of the research has tended to focus mostly on the intrapersonal experiences of only one party - the 'victim' (Baumeister et al., 1998) - either in relation to a specific event or relationship, or as a more general, stable disposition. Forgiveness represents a change in motivations (McCullough et al., 2007; McCullough et al., 1998), and/or cognition, affect and behaviour (Baumeister, Exline, & Sommer, 1998; Subkoviak et al., 1995) in response to a transgressor or transgression. However, there remains debate on the extent to which this change occurs, and the nature of the forgiving response. For some it is enough that negative feelings or motivations toward a transgressor are removed (Thompson et al., 2005). For others forgiveness must be more than just the

[1] There is a separate but related literature on self-forgiveness, which is purely an intrapersonal process, but this construct will not be addressed in this thesis.

absence of ill-will, and must also include feelings of benevolence toward the transgressor (Fincham, Beach, & Davila, 2004; McCullough, Root, & Cohen, 2006), and perhaps some intention to exhibit this benevolence through action (Worthington, 2005a). Finally, for some theorists, forgiveness requires more than just benevolence: it requires compassion, altruism or even love (Enright, Freedman, & Rique, 1998; Worthington, 1998).

In attempting to frame the parameters of what forgiveness is, theorists have also asserted what they believe forgiveness is not. We are told that forgiveness is not the same as pardoning, excusing, condoning, overlooking, forgetting or accepting because these terms imply (at most) a denial of the offence or (at least) discounting of the severity of it (Fincham, 2000; Enright et al., 1998). Forgiveness is seen to be *more* than these things (Enright & North, 1998b). We are told that forgiveness is not to be confused with reconciliation because the two are independent processes: a relationship can be reconciled without forgiveness having taken place (Enright & Zell, 1989). Conversely, forgiveness can occur without any need to restore or reconcile the relationship (Denham et al., 2005; Scobie & Scobie, 1998).

More recently, however, McCullough (2008) has challenged the idea that forgiveness and reconciliation are as different as is often claimed, suggesting instead that reconciliation may simply be the observable, behavioural expression of forgiveness, and that forgiveness only evolved because of the adaptive benefits that reconciliation afforded our ancestors. Similarly, Frise and McMinn (2010) argue that reconciliation is merely a form of 'relational forgiveness'.

It is worth noting, however, that much of the theorising about forgiveness has been prescriptive or 'top-down' in nature. In an attempt to explore lay

conceptualisations of forgiveness, Kearns and Fincham (2004) conducted a prototype analysis. In an exploratory study, participants generated a list of exemplars, which yielded 78 attributes that were associated with forgiveness. In a second study, these 78 attributes were then rated for their centrality on an 8 point Likert scale. Many of the abovementioned concepts such as reconciliation (centrality rating of 6.62), accepting (6.38), and forgetting (4.58), were relatively central to participants' understandings of forgiveness, which is in stark contrast to much of the theorising. Using a similar methodology, Friesen and Fletcher (2007) also found that lay people considered constructs such as forgetting (3.90 on a 7 point centrality scale), and generally letting go (4.92) to be more central to forgiveness than the idea that forgiveness is offered as a 'gift' (2.25).

Other studies also highlight some key differences in the way forgiveness is understood by lay people. Kanz (2000) administered a questionnaire consisting of 23 yes/no items representing commonly cited beliefs about forgiveness, with the majority (69%) of respondents believing that reconciliation is a fundamental component of forgiveness; that it is possible to forgive someone without them being aware of it (97%); and that it is possible to remain angry at someone even though you have forgiven (76%). Consistent with these understandings, Younger et al (2004) found that participants regarded forgiveness as being self-focused rather than altruistic, with the most commonly given reasons concerning the personal health and happiness of the victim. Participants were also asked to define forgiveness, with the most commonly reported elements being accepting, moving on, letting go, reconciliation, and even forgetting, many of which have no place in academic definitions of the construct. Thus, there appears to be some

disparity between the understandings of forgiveness put forward by theorists, and those put forth by lay people.

1.2 Measuring Forgiveness

Forgiveness can be measured at either the situation-focused (e.g. McCullough et al., 1998; Subkoviak et al., 1995), or person-focused (e.g. Brown, 2003; Thompson et al., 2005) levels.

Situation-focused ('state') measures attempt to capture a person's forgiveness thoughts, feelings, and motivations at a particular point in time and in relation to a specific transgression or transgressor. One of the most commonly used state measures of forgiveness is the Transgression-Related Interpersonal Motivations scale (TRIM; McCullough et al., 1998), which asks a person to remember a specific instance in which they were transgressed against, and then assesses that person's behavioural motivations toward that specific transgressor using a rating scale. The original version of the TRIM contained 12 items and assessed two kinds of motivations – the motivation to avoid the transgressor and the motivation to exact revenge on the transgressor (McCullough et al., 1998). The scale was later revised and expanded to 18 items, to include a third set of motivations: benevolence (McCullough & Hoyt, 2002; McCullough et al., 2006).

Another situation-based measure that has been used in several studies is the Enright Forgiveness Inventory (EFI; Coyle & Enright, 1997; Reed & Enright, 2006; Subkoviak et al., 1995). This is a 60 item measure that assesses both the presence and absence of cognitive, affective and behavioural dimensions of forgiveness, by using 6 subscales.

Person-focused forgiveness measures generally focus on either attitudes toward forgiveness, or forgiveness as an enduring personality trait (sometimes called "forgivingness": Berry et al, 2001). Trait measures of forgiveness assess the extent to which people are forgiving across all/most situations and in all/most relationships. This is typically done either by asking participants to report their general forgiveness tendencies, or by assessing projected intentions to forgive across a range of transgressions and then aggregating these responses. Measures of trait forgiveness include the Transgression Narrative Test of Forgivingness (TNTF; Berry, Worthington, Parrott III, O'Connor, & Wade, 2001), the Trait Forgivingness Scale (TFS; Berry et al., 2005), the Tendency To Forgive scale (TTF; Brown, 2003), as well as two independent Willingness To Forgive scales (WTF; DeShea, 2003; Hebl & Enright, 1993). A summary of some of the scenario-based measures can be found in DeShea (2003, p.203).

Attitudinal measures require individuals to indicate their general thoughts and feelings towards forgiveness, irrespective of whether this evaluation is related to forgiveness motivations, intentions, or behaviours. Examples of attitudinal measures of forgiveness include the Attitudes To Forgiveness scale (ATF: Brown, 2003), the Heartland Forgiveness Scale (HFS; Thompson et al., 2005), and the Forgiveness Attitudes Questionnaire (FAQ; Kanz, 2000).

Although the state, dispositional, and attitudinal approaches differ in their specific content, they invariably share a common structure: they are all self-report scales. This heavy reliance on a single mode of measurement poses a problem for forgiveness research.

1.3 The "Mono-Method Bias" & Limitations of self-report

In an early review, McCullough and Worthington (1994) noted that there was no consensus as to how to best measure forgiveness. Since then there have been numerous studies measuring both state and trait forgiveness, with the bulk of these relying on self-report questionnaires to collect their data (McCullough et al., 2000; McCullough, Root, Tabak, & Witvliet, 2009; Mullet, Neto, & Riviere, 2005). Such reliance on one mode of measurement has been termed a "mono-method bias" (Hoyt & McCullough, 2005), and has some worrying implications. A problem with using several measures of the same mode (i.e. self-report) is that there is increased possibility that they will also share the same kinds of 'bias' variance (Hoyt & McCullough, 2005). Specifically, self-report measures are based on the assumption that people are always willing to accurately report their 'true' attitudes, when there is evidence to suggest that this is not always the case. That is, people may respond in a manner that is deemed to be socially desirable, rather than indicating how they truly think or feel (Crowne & Marlowe, 1960; Jones & Sigall, 1971). It is also possible that people may not always be able to accurately report their 'true' attitudes, as they may be unaware that they hold particular attitudes (Wilson & Dunn, 2004).

1.3.1 Self-presentation concerns: *Unwillingness* to accurately report attitudes.

The inclination for people to complete scale items in a way that presents themselves favourably is now a well established phenomenon in social psychology (Crowne & Marlowe, 1960). Socially desirable responding may be magnified when the topic of interest has an encultured positive or negative bias, as cultural norms play an

important role in determining what an individual will perceive as socially desirable in the first place (Fisher & Katz, 2000). It could be argued that forgiveness is one such topic that would be particularly susceptible to socially desirable responding, particularly given that forgiveness is clearly a prosocial response (Enright & Fitzgibbons, 2000; Enright et al., 1998; McCullough & Worthington, 1994). McCullough and colleagues frequently describe forgiveness as 'prosocial' (Bono et al., 2008; McCullough et al., 2007), while Enright and colleagues go so far as to call it a "gift", focusing on its positive components – compassion, mercy and love (Enright et al., 1998). Forgiveness may be perceived as even more desirable in countries that have been founded on Christian ideals, owing to the revered place of forgiveness in Christian traditions.

Although forgiveness may be viewed as desirable at the societal level, individual experiences of forgiveness may be less positive. In their prototype analysis, Kearns and Fincham (2004) found that some participants reported perceiving forgiveness as a 'sign of weakness' or as 'giving the (transgressor) permission to (offend) again'. Similarly, Friesen and Fletcher (2007) found that anxiety about whether the transgressor might reoffend was relatively central to people's conceptualisations of forgiveness. Lamb (2006) suggests that forgiveness can often be particularly disempowering, especially in relationships where power balances are already unequal. People may be acutely aware of these costs of forgiving, and view forgiveness with suspicion as a result (Baumeister et al., 1998). There is also evidence to suggest that forgiveness may have genuinely negative consequences for a victim's self-respect if the perpetrator is unrepentant about what they have done (Luchies, Finkel, McNulty, & Kumashiro, 2010).

In short, forgiveness is particularly vulnerable to self-presentation concerns in *both* directions. Thus, self-report scales may not always be the most effective tools for assessing forgiveness attitudes.

1.3.2 The limits of introspection: *Inability* to accurately report attitudes

In addition to people often being unwilling to reveal their attitudes, it is also possible that people are sometimes unable to report these attitudes (Greenwald & Banaji, 1995; Greenwald, McGhee, & Schwartz, 1998). Self-report measures assume that attitudes are conscious and easily accessible, and can therefore be measured through explicit tests. This assumption is incongruent with some of the most famous work in psychology: the idea that a large proportion of the human mind is 'unconscious' (Frey-Rohn, 1974). Greenwald and Banaji (1995) argue that much of our reasoning about the world is performed at an unconscious level, a process they refer to as 'implicit social cognition'. Furthermore, these implicit cognitions and attitudes may differ from our explicit ones; i.e., those of which we are consciously aware (Greenwald, 1990).

In the literature, forgiveness is often framed as a deliberate, conscious, and controlled process – a choice that people make after some reflection (Worthington & Scherer, 2004). Recently, however, it has been suggested that forgiveness may at least partially operate at the unconscious level – that there can be a degree of automaticity in the forgiveness process – especially in close relationships (Karremans & Aarts, 2007; Karremans & Van Lange, 2008). Karremans and Aarts (2007) argue that forgiveness forms part of the mental representations that we have of our close relationships, which they refer to as *relational schemas* (Baldwin, 1992). Availability of these schemas means that

forgiveness takes place more automatically in close relationships than it does with transgressions involving those who are less close. Across four studies they found that forgiveness was more easily activated after being subliminally primed with a close (vs non-close) other. Further, forgiving a close (vs non-close) other required less cognitive resources (under time constraints). Although the idea that forgiveness may have a substantial automatic component is a relatively new one, and these processes remain largely unexplored, it does still suggest that self-report scales may not always be suitable for accurately measuring forgiveness.

1.4 Existing alternatives to self-report in forgiveness research

Examples of studies that examine forgiveness using measures other than self reports are scarce, but the few available studies take either a behavioural (Karremans, Van Lange, & Holland, 2005; Struthers, Eaton, Santelli et al., 2008; Wallace et al., 2008; Zechmeister, Garcia, Romero, & Vas, 2004) or physiological (Farrow & Woodruff, 2005; Witvliet et al., 2001) approach.

1.4.1 Behavioural approaches

Studies on forgiveness which experimentally manipulate actual transgressions and/or measure forgiveness through actual (rather than self-reported) behaviour are conspicuously scarce. Zechmeister et al. (2004) examined the effects of apologies in the forgiveness process by giving participants an unsolvable task, and then providing scathing feedback on their performance. Forgiveness was measured by the extent to which participants were willing to help the experimenter with another research project.

Struthers, Eaton, Santelli et al. (2008) had participants work cooperatively with a confederate on a reading comprehension task, for which they were informed that high task performance would increase their chances of winning a $50 prize. After the confederate sabotaged the task, participants were given the opportunity to assign ballots for the prize draw to themselves and their partner: assigning a greater number of ballots to the partner was interpreted as greater forgiveness. Karremans et al. (2005) asked participants to recall a transgression and then complete a task which required them to 'fill in the blanks' in a foreign (bogus) language paragraph by guessing the personal pronouns. More inclusive (e.g. 'we' versus 'I') pronouns indicated greater forgiveness, and correlated accordingly with TRIM-reported forgiveness. In a second study, forgiveness was measured using two indicators of support for a charity: nominating the number of hours they would be willing to volunteer, and the amount of money they placed in a donation box for this charity.

While these behavioural approaches to examining forgiveness help to better illuminate the ways in which forgiveness is actually carried out, they are not without their limitations. The most important of these is the way in which forgiveness is operationalised. Zechmeister et al. (2004) assessed the extent to which participants were willing to do a favour for the experimenter by volunteering their time (number of hours) to help with another study, and then treated this as a proxy for forgiveness. The problem with this approach is that it is difficult to determine participants' exact motivations for volunteering their time: this measure may be more indicative of prosocial behavioural tendencies than specific forgiveness behaviour.

The same can be said of the outcome measure used by Struthers, Eaton, Santelli et al. (2008). Allocating ballots to a partner captures both prosocial and vengeful motivations towards the offender, but it does not necessarily represent forgiveness: this behaviour may be indicative of generosity or general prosocial tendencies. It is also difficult to assess whether this behaviour is a result of desire to forgive the accident, punish the intentional act, or a combination of the two. Karremans et al. (2005) also took a behavioural measure of prosocial behaviour – the amount of money that participants chose to donate to a charity after they had spent time reflecting on a transgression - but the distinction here is that they did not claim that this equated to forgiveness.

1.4.2 Physiological approaches

Despite McCullough and Worthington's (1994) recommendation, physiological approaches to forgiveness measurement remain scarce. Witvliet et al. (2001a) measured a range of physiological responses while participants were mentally rehearsing either 'forgiving' or 'unforgiving' scripts. Participants were instructed to remember a real autobiographical transgression and then assigned to either a forgiveness (empathising with the offender, or granting forgiveness) or an unforgiveness script (mentally rehearsing the hurt, harbouring a grudge). Participants were asked to relate the scripts to the specific transgression they had remembered. While they were doing this, several physiological measures were taken, including facial (brow) EMG, skin conductivity, heart rate and blood pressure. Significantly higher levels on all of these measures were found for the 'unforgiving' condition, relative to both the 'forgiving' condition and baseline measures. More recently, forgiveness has also been assessed physiologically using fMRI

and PET techniques, assessing brain activity while participants make a series of forgiveness-relevant decisions (Farrow et al., 2001; Hayashi et al., 2010; Young & Saxe, 2009).

While these non-self-report measures do provide some much-needed alternatives for forgiveness measurement, and provide some further insight in to the intra-individual side of the forgiveness process, they still have their various limitations. Physiological approaches such as those explored by Witvliet et al. (2001a) and Farrow et al. (2001) are expensive and time consuming, and require that the researcher has access to specialised equipment (i.e. heart rate monitors or brain scanning equipment). In some of these cases, there is also the possibility that we are assessing correlates of forgiveness, rather than forgiveness itself. For example, Farrow et al. (2001) attempted to examine forgiveness by presenting participants with a series of decision-making tasks while they were under fMRI. The task presented participants with a short scenario ("You read in a newspaper that a well-known television presenter has appeared in court, charged with an offence"), and then asked them to make a series of forced-choice responses as to what crimes they would evaluate as "more forgivable" (e.g. "income tax evasion" versus "council tax evasion", or "speeding on a motorway" versus "speeding on a country road")[2].

It is debatable, however, as to whether these studies are actually tapping in to forgiveness or to a more general moral or ethical reasoning ability. It could be argued that these decisions merely assess a person's logical reasoning ability, rather than forgiveness per se. As already mentioned, forgiveness is frequently defined as an

[2] See Farrow and Woodruff (2005) for a review of other studies which employ a similar methodology

interpersonal process, with a significant focus on intrapersonal processes, yet this study examines forgiveness as being removed from the individual – assessing objective ideas about right and wrong. Finally, it is unclear exactly how to interpret the information gleaned from this line of research – even if we can definitively establish that forgiveness is related to specific areas of the brain, what application does this have for our understanding of how forgiveness works in the real world?

In summary, while these physiological approaches may help to develop our understanding of some aspects of the forgiveness process, the current cost (both economic and temporal) of administering them, coupled with uncertainty about how to interpret the results means that they are limited. Moreover, these approaches only measure forgiveness at the *state* level: alternative ways of measuring forgiveness *attitudes* and *dispositions* are still required. These limitations, considered together with the previously discussed constraints of self-report scales and behavioural approaches, illuminates a strong need to devise new and novel methods for forgiveness measurement. One non-self-report measure that appears to be particularly suitable for measurement of *attitudes* toward forgiveness is the Implicit Association Test (IAT; Greenwald et al., 1998).

1.5 The Implicit Association Test

The IAT is a computer-based sorting task that measures the time it takes for a person to sort one pair of target words or concepts relative to another pair of words or concepts (Greenwald et al., 1998; Lane, Banaji, Nosek, & Greenwald, 2007). To illustrate how the IAT works, consider a commonly used example IAT that seeks to explore how a

person evaluates, in terms of pleasant or unpleasant, flowers relative to insects. A standard IAT has seven trial blocks. In the first block, the target word 'flowers' appears on the left side of the computer screen and the target word 'insects' appears on the right (or vice versa). A series of words or labels that represent either target (e.g. rose, daffodil, cockroach, beetle) are then displayed in the centre of the screen and the participant is asked to sort them using a 'left' or 'right' response key on their keyboard. Block two utilises the same procedure but replaces target pair 'flowers' and 'insects' with the attributive dimension 'pleasant' and 'unpleasant', then presents a series of words that represent these categories (e.g. happy, heaven, evil, monster). The third and fourth block combine these two tasks, placing 'flower' and 'good' on the left side of the screen, and 'insect' and 'bad' on the right. Block five is identical to block one but reverses the positioning of the two targets – in this case 'insects' will now be on the left (with 'pleasant') and 'flowers' on the right (with 'unpleasant'). Block six and seven are identical to block three and four, with the target concept reversed and the attribute dimension remaining stable, in this case 'insects' and 'pleasant' on the left and 'flowers' and 'unpleasant' on the right.

The critical blocks are block three and four, and six and seven (the remainder are practice tasks) and the IAT score is computed by assessing response times on one of these relative to the other. That is, did someone respond faster when 'flowers' and 'pleasant' (and 'insects' and 'unpleasant') shared a response key, than when 'insects' and 'pleasant' (and 'flowers' and 'unpleasant') shared a response key? One might predict that flowers would indeed be evaluated more favourably than insects, as this appears to be the more

compatible pairing, and evidence from studies using the IAT regularly supports this prediction (Greenwald et al., 1998; Kim, 2003; *cf* Govan & Williams, 2004).

The IAT-measured preference for one category relative to its paired category is termed an 'implicit attitude' or 'implicit preference', and has been defined as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favourable or unfavourable feeling, thought, or action toward a social object or concept" (Greenwald & Banaji, 1995, p. 8). The term "implicit" is used because it is thought that these attitudes/preferences operate outside of conscious awareness, and evidence demonstrating that people find it difficult to consciously manipulate their own IAT scores appears to support this claim (Kim, 2003; Steffens, 2004). That is, the IAT uses reaction time and error rate data to measure an attitude/preference that is *implied*, rather than explicitly stated.

Examining attitudes at the implicit level is useful because it has the potential to uncover aspects of a person's attitudes that were previously unavailable through self-report, but may be equally predictive of behaviour. The IAT is also well suited as an alternative or complement to self-report measures as it is an indirect, rather than direct, measure of attitudes (Fazio & Olsen, 2003; Greenwald et al., 1998). While self-report instruments can directly assess what a person consciously believes about a given topic, the IAT does not require that a person be conscious of an attitude to be able to measure it.

1.5.1 Why the IAT?

The strengths of the IAT as a measure of implicit social cognition are now well-documented. Since its inception, the IAT has become arguably the most widely used and validated measure of implicit social cognition. A search of academic database PsycINFO using the keyword phrase "implicit association test" returns a list of almost 700 published papers on the topic, and it continues to be one of the most reliable of the implicit cognition measures (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005a; Lane et al., 2007). In a meta-analysis of 50 studies which made use of the IAT, Hofmann et al. (2005) reported a mean internal consistency reliability of r=.79 (an average using a linear model, across studies that reported either internal consistency or split-half estimates), which is much greater than reliabilities typically reported for other key measures of implicit cognition such as the Go No-go Association Task (GNAT; Nosek & Banaji, 2001; r=.20), and the Extrinsic Affective Simon Task (EAST; De Houwer, 2003; r=.25). The IAT also produces large effect sizes (e.g. Bosson, Swann, & Pennebaker, 2000; Houben, 2007), and can be easily adapted for use as an online measure (Nosek, Banaji, & Greenwald, 2002a).

Most importantly, the IAT appears to do what it claims to do: it is resistant to social presentation concerns. The IAT's resistance to socially desirable responding patterns is well documented (Lane et al., 2007; Poehlman, Uhlmann, Greenwald, & Banaji, 2005). The strongest evidence for this comes from a meta-analysis of 184 independent samples, in which high social sensitivity had a far greater impact on self-reported attitudes than on respective IAT scores (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). For each study in the analysis, social sensitivity was coded by three

independent raters, assigning a value between 1 and 7 in response to the item "likely to be affected by social desirability concerns". Overall, social sensitivity accounted for 24.4% of variance in the effect sizes for self-reported attitudes, while only accounting for 3.4% of the variance in IAT scores. This evidence forms a compelling case for the resistance of the IAT to self-presentation concerns.

The IAT's resistance to self-presentation probably owes to the fact that it is not easily 'fake-able' (Kim, 2003; Lane et al., 2007). Kim (2003) found that participants could only fake IAT scores if they were first *taught* strategies for doing so. Similarly, in two studies, Steffens (2004) found that participants were unable to fake conscientiousness but were able to fake extraversion on IATs, but this was only possible after participants had had prior experience completing an IAT, and were given explicit instructions on how to fake their scores. In contrast, a study on shyness found that participants were unable to fake IAT scores, even with instructions on how to do so (Asendorpf, Banse, & Mucke, 2002). Further, studies that provided no instructions on effective faking strategies found no significant effects of faking in the domains of implicit anxiety (Egloff & Schmukle, 2002) or implicit attitudes toward homosexuality (Banse, Seise, & Zerbes, 2001). Given the unlikelihood that the majority of research participants would have had prior experience with the IAT procedure, and that explicit faking instructions are not normally provided, it appears that under normal conditions the IAT remains robust to self-presentation concerns.

1.5.2 What does the IAT actually measure?

Despite the growing body of evidence for the validity, reliability and utility of the IAT, one key aspect of the IAT which still attracts some controversy is its construct validity: what does the IAT actually measure? Critics have suggested that because IAT measurement relies on *associations*, IAT-measured preferences may be as much a product of general cultural knowledge as they are of our specific, unique implicit attitudes. That is, the IAT may measure *familiarity* with the stimuli or categories rather than (or in addition to) *individual preference* (Kinoshita & Peek-O'Leary, 2006). These two competing explanations are difficult to separate, as the entire premise of the IAT is that it operates at a non-conscious level – and if we can't accurately introspect on this process then understanding the process is difficult.

Some evidence that the IAT does measure individual preferences comes from Siegel (2006). Across several studies, participants were given information about unfamiliar groups and then completed IATs. Participants who believed the information to be correct showed IAT scores that were consistent with the information provided, whereas those who doubted the accuracy of the information did not. This provides evidence that IAT scores are not just a function of a person's general exposure/knowledge, but also their individual endorsement of this information.

Perhaps the most compelling evidence that IATs measure individual preferences comes from research examining IAT convergence with self-report measures, and the predictive validity of IATs. Two large meta-analyses have revealed that IATs correlate with corresponding self-report measures at an average r=.23 (152 independent samples; Greenwald et al., 2009) to r=.24 (126 samples; Hofmann et al., 2005a), which suggests

that there must be at least some level of individual preference being captured by the IAT. It should be noted, of course, that these correlations are relatively small, meaning that there may still be *some* variance explained by factors beyond individual preference. However, the small magnitude of these correlations may not pose a significant problem in assessing the utility of the IAT, particularly as the IAT has repeatedly been shown to predict behaviour. In a recent meta-analyses of more than 100 studies, Greenwald et al. (2009) found that IATs were related to a range of behavioural, physiological and judgement measures at an average r = .27. Significantly, the IAT has been able to predict behaviour in a large number of applied contexts: implicit race stereotypes can predict courtroom judges' verdicts (Rachlinski, Johnson, Wistrich, & Guthrie, 2009) and job-hiring behaviour in an organisation (Rooth, 2010), and a suicide IAT can predict actual suicide (Nock et al., 2010).

In broad terms, the fact that the IAT has been shown to predict behaviour across a range of domains suggest that it may not be all that important whether it measures individual attitudes, broader cultural knowledge, or a combination of both – it still tells us something useful about a particular construct.

1.5.3 Summary of the IAT

Understanding the mechanisms and applications of the IAT is an ongoing process. Issues of IAT validity and utility have, and will, continue to receive much attention in the research literature. For additional discussion of this area, refer to reviews by Nosek, Greenwald and Banaji (2007), Lane et al. (2007), Nosek et al. (2007), Greenwald et al. (2009), Schnabel, Asendorpf and Greenwald (2008a). For now, we do know that the IAT is

able to address some of the issues surrounding self-report measures, because of its well-documented resistance to socially desirable responding. We also know that the IAT can be a useful predictor of real-world behaviour, especially in domains that are socially sensitive. Together, these two factors make the IAT a potentially promising resource for the future of forgiveness measurement.

1.6 Thesis aims and overview

The measurement of forgiveness has been criticised for (a) too often relying on a mono-method approach, and (b) that approach too often being self-report (Hoyt & McCullough, 2005; McCullough et al., 2000). As already discussed, reliance on one mode of measurement increases the probability of shared error variance in results, while self report methods have been criticised for being susceptible to socially desirable responding, as well as not being able to assess attitudes that operate at a more unconscious level. Developing an IAT for use in forgiveness measurement has the potential to address all three of these concerns. In combination with other (self-report) measures, an IAT could be used as part of a multi-method approach to forgiveness measurement and thus reduce shared error variance in results. Furthermore, relationships between the IAT and other measures may provide some insight into the way that forgiveness is conceptualised. The IAT, owing to its ability to measure attitudes at a non-conscious level and resistance to being 'faked', should also be more resistant to socially desirable responding than self-report scales. Finally, irrespective of social desirability concerns, the IAT should be able to examine forgiveness attitudes at a non-conscious level.

The primary aim of this thesis was to develop an Implicit Association Test that is suitable for the measurement of forgiveness. This measure was constructed and tested across a series of nine studies. Chapter 2 presents data from two studies, along with a pilot study, that address the construct validity for the forgiveness IAT at the *stimulus* level: primarily the selection of appropriate words to validly represent forgiveness in the IAT. In Chapter 3, construct validity is addressed at the *category* level, presenting a study which compares and contrasts potential "opposite" categories for forgiveness. Chapter 4 demonstrates that convergence between implicit and explicit forgiveness measures can be greatly improved by using a self-concept IAT rather than an attitudinal IAT. Data from two studies are presented in support of this. Chapters 5 and 6 examine the predictive validity of the Forgiveness IAT across four studies, utilising a range of research methodologies including retrospective prediction of past offenses, and measuring behaviour in response to both hypothetical scenarios and a "real-life" transgression using an iterated trust game. Finally, Chapter 7 summarises and discusses the work as a whole.

Chapter 2:

Developing an Implicit Association Test for Forgiveness and the potential effects of stimulus word valence

2.1 Chapter Overview

This chapter presents two studies, each of which shared two key aims. The first was to develop and validate an IAT that could be used to assess people's implicit forgiveness attitudes. The second aim was to address a potential methodological concern in designing this IAT; namely, to determine if the valence of stimuli words within IAT categories would be a significant confound in producing IAT scores for forgiveness.

2.2 Extra-attitudinal influences on IAT effects

Scores on *attitudinal* IAT measures (those which use "pleasant-unpleasant" as the evaluative dimension) are intended to assess implicit preference for one target category relative to the other paired target category. Thus IAT scores are meant to tap implicit *associations* and, as discussed in Chapter 1, there is now abundant evidence that IATs are often able to do this. However, it is also possible that scores on IATs may be produced by factors that are independent of actual associations: that is, performance on an IAT might depend on other features of the task (for reviews see Blair, 2002; Bluemke & Friese, 2006). Two of the more prominent (and somewhat inter-related) alternative explanations are the salience asymmetry account and the strategic recoding account.

2.2.1 Salience Asymmetries: the IAT as a figure-ground task

Rothermund and Wentura (2001, 2004; Rothermund, Wentura, & De Houwer, 2005) were among the first to challenge the *association account* of IAT effects, exploring extra-attitudinal influences on IAT scores across a number of studies. Using a figure-ground framework, they suggested that performance on an IAT is more dependent on the

salience of the category pairings than on actual implicit associations – that is, the greater the salience of a pair of categories, the faster people will respond. For the target dimension of the IAT (e.g. flowers vs insects, old vs young) they argued that the more novel of the two categories would be more salient and the more familiar category would be less salient. Similarly, they said, for the evaluative dimension of the IAT (e.g. pleasant vs unpleasant), 'unpleasant' words are, by default, more salient, owing to their negative valence. There is considerable evidence that attention is driven by negative emotion: people more readily attend to stimuli that are negatively valenced than stimuli that are positively valenced (e.g. Fox et al., 2000; A Öhman, Flykt, & Esteves, 2001; Pratto & John, 1991). Thus an IAT preference may not necessarily be the product of people responding faster when 'pleasant' words are paired with the target they favour. Rather, in the same component of the task 'unpleasant' words are paired with the more novel target, and in many cases this also happens to be the target that they do not favour. As such, the IAT task becomes more about sorting the salient *figure* against the less salient *background,* rather than about revealing implicit attitudes.

To test this *salience account* of IAT effects, Rothermund and Wentura (2004) conducted a series of experiments in which they manipulated the salience of the various IAT categories. In their first study they ran four experiments with young participants completing an old-young IAT, using typical names of old people (more novel, and therefore more salient) and young people (more familiar, less salient) as the stimulus words. In the first experiment they conducted a standard evaluative IAT, using categories of pleasant and unpleasant, and found typical IAT association effects: participants found the task easier when young/pleasant (less novel/salient) and old/unpleasant (more

novel/salient) were paired, a finding supportive of both the association and salience accounts. For each of the three subsequent experiments they manipulated the salience of the evaluative categories, replacing the pleasant-unpleasant dimension with words (familiar) vs non-words (salient); non-words vs negated non-words (a non-word preceded by the word "no" – this was deemed more salient); and single-coloured (less salient) vs multi-coloured (more salient) words. In all three experiments they found that participants could respond more quickly when the two 'salient' categories were paired than when the salient categories were placed on opposing sides of the IAT.

Note that these experiments only provided support for the salience account without discounting the association account, as there was no evaluative category (i.e. pleasant/unpleasant) present in any of them. Put simply, these experiments show that IAT effects can be driven by salience of categories when the usual mechanisms for evaluation based on implicit preference are removed. Rothermund and Wentura did attempt to address the relative contributions of the two accounts in one of their experiments (2004, study 3a). Using an old/young X pleasant/unpleasant IAT, they first primed participants with go/no-go tasks to make one of each of the two paired categories more salient than the other, following which participants completed a standard IAT. Results showed that priming salient pairings that were counter-intuitive to the association model (e.g. old-pleasant) actually produced IAT effects that were in the opposite direction to what would normally be expected if the IAT was measuring associations. This seems to provide evidence for the salience account of IAT effects.

Additional data which seem to at least partially support the salience account of IAT effects has been found by Kinoshita and Peek-O'Leary (2006), Mierke and Klauer

(2003), and Brendl, Markman and Messner (2001). However, all of the experiments reported operated under somewhat artificial conditions – they all sought to deliberately interfere with the IAT, either by removing the evaluative dimension (i.e. replacing "pleasant-unpleasant" with a non-evaluative pairing) or priming attention, to produce these salience effects. This kind of interference is rarely present in usual applications of the IAT, so this still does not discount the association account as a credible explanation for IAT effects. In a direct comparison of several IATs, Kinoshita and Peek-O'leary (2006) found some support for this idea that IAT effects may not merely be the result of a single mechanism, concluding that IAT effects can be the product of a variety of factors, including both salience and association. It could be that, as a default position, the IAT *does* measure implicit associations (i.e. in the absence of deliberate efforts to manipulate target salience, the IAT will capture actual implicit preferences), but that this can be over-ridden if the IAT task or process is sufficiently tampered with. In summary, there is some evidence to suggest that salience asymmetries *can* cause IAT effects, but this relationship has not been demonstrated in most normal uses of the IAT.

2.2.2 Strategic recoding of the IAT tasks: Beyond salience

In summarising their findings, Rothermund and Wentura (2004) suggested that their salience account of IAT effects had a significant implication for a great number of IATs that already existed in terms of what they called "strategic recoding". Put simply, they suggested that participants may mentally simplify an IAT task by grouping two categories according to their common/salient features, thus simplifying what should be a four-category double-discrimination task in to a two-category single-discrimination task.

For example, performance on an IAT with target categories of weapons/musical instruments and evaluative categories of pleasant/unpleasant (such as that used by Greenwald et al., 1998) might be easier when weapons-unpleasant and musical instruments-pleasant are paired purely because one can recognise that weapons are culturally seen as "bad" while musical instruments are more "good", independent of one's own implicit preferences for these two things. Thus, this part of the IAT is simplified to a single-discrimination task based purely on the *valence* of the stimuli – is this a "good" word or a "bad" word? During the opposite pairings of weapons-pleasant and musical instruments-unpleasant one has access to no such mental shortcut and thus performance on this "incompatible" block will be much slower, resulting in inflated IAT scores that are independent of the individual's own implicit preferences.

In effect, the strategic recoding account essentially states that categories can be paired according to any common features that they share, not just the salient ones. As such, this account has also been described as being about similarity (De Houwer, Geldof, & De Bruycker, 2005) or congruence (Bluemke & Friese, 2006) between categories, target familiarity (Kinoshita & Peek-O'leary, 2006) and the fluency with which categories can be processed (Chang & Mitchell, 2009). In a handful of studies, it has been shown that strategic recoding can occur as a result of a number of different shared features of stimuli, and that this is irrespective of whether the similarities are based on salience, or on other shared features.

2.2.3 Strategic recoding based on valence of category and/or stimuli

By far the most common form of IAT uses a pleasant-unpleasant (or good-bad) dichotomy as its evaluative dimension, and thus far there has been no published research showing that other shared features of categories can "over-ride" this particular pairing. However, as noted earlier, such a pairing may be problematic in its own right. On a flower-insect IAT one might show a preference for flowers relative to insects, but this could just be because one recognises that flowers are *generally* seen as being more positively valenced than insects, irrespective of one's own preferences. Specifically, a person may strategically recode the IAT task such that positively valenced (e.g. flowers/pleasant) words are sorted to one side, while negatively valenced (e.g. insects/unpleasant) words are sorted to the other.

Practically, strategic recoding based on valence may not be problematic. In a response to Rothermund and Wentura, Greenwald et al. (2005) argued that the former's suggestion that people may strategically recode IAT categories on the basis of category valence was "empirically equivalent" (p. 423) to their own association-strength account of IAT effects. At face value, this claim seems to hold some weight. If a person associates one target with pleasant words and/or the other with unpleasant words then this may be suggestive of their actual preference, i.e. if I am able to recognise that flowers are "good" and insects are "bad", then surely this reveals something about my own attitudes toward the two, at least at the category level. However, Greenwald et al.'s (2005) claim becomes problematic in light of evidence which examines valence of the specific stimuli used to represent the categories, rather than just the valence of the categories themselves. Specifically, several studies have shown that manipulating the valence of the stimuli

representing each category can reduce the magnitude – and in some cases even reverse the direction – of IAT effects (Govan & Williams, 2004; Mitchell, Nosek, & Banaji, 2003).

The most convincing example of the effects of stimuli valence comes from a study by Govan and Williams (2004) in which they reversed the typically found flower-insect IAT effect (study 1a). Participants were assigned to either a typical or an atypical flower-insect (pleasant-unpleasant) IAT, which differed only in the stimulus words used to represent the flower and insect categories. In the typical IAT, flower and insect exemplars were similar to those used by Greenwald et al. (1998); that is they were positive exemplars of flowers (e.g. rose, daffodil) and negative exemplars of insects (e.g. cockroach, wasp). For the atypical IAT the valence of exemplars was switched such that flowers were represented negatively (e.g. weed, poison ivy) and insects represented positively (e.g. ladybird, butterfly). While results from the typical IAT replicated previous findings, the atypical IAT produced atypical results – participants showed a small preference for insects over flowers, although the effect size was much smaller than that found in the standard condition.

It could be argued that attitudes toward flowers and insects are of relatively little consequence in the real world and that perhaps IATs measuring more meaningful associations might not be affected in the same way. However, similar (albeit weaker) results have been found using a race IAT (Govan & Williams, 2004; Mitchell et al., 2003). Mitchell et al. (2003, experiment 2) asked participants to complete one of two versions of a black-white (good-bad) IAT – one which represented black with three 'negative' and white with three 'positive' exemplars, and one which represented black with three

'positive' and white with three 'negative' exemplars[3]. While a significant and relatively strong preference was found for whites compared to blacks on the standard black-white IAT (typical finding), the atypical IAT revealed a non-significant preference that was still in the same direction (i.e. preference for white compared with black). Using a similar methodology, Govan and Williams (2004, study 1b) were also able to eliminate the IAT preference for white over black by using positive black (e.g. Eddie Murphy, Cathy Freeman) and negative white (e.g. Adolf Hitler, Pauline Hanson) exemplars but were unable to reverse it in the same way that they did with their flower-insect IAT.

2.2.4 Evaluating the impact of valence effects

The findings presented above suggest that valence of stimulus labels can have an impact on IAT effects over and above any actual implicit associations that may exist. However, it is worth noting that stimuli valence cannot *completely* account for any of the effects described above. If the IAT scores were exclusively a product of stimuli valence then switching the valence of the stimuli should have produced IAT preferences in the opposite directions i.e. there should have been a preference for black compared to white. Instead, in both of these studies switching valence still produced IAT scores which showed a preference for white over black, albeit significantly lower in magnitude.

Govan and Williams (2004) had more success in reversing the direction of IAT preference with their flower-insect IAT, but the effect size was much smaller (mean

[3] participants had previously completed a task where they classified images of famous white and black actors/politicians/athletes/etc in terms of how positive or negative they were, and based on ratings these images were then subsequently used in the participants' IAT tasks

difference in reaction times of 76.28 milliseconds) than the effect obtained for the IAT that was calibrated in the standard direction (mean reaction time difference of 307.98ms). This indicates that while stimuli valence can influence eventual IAT preferences, it does not completely override other associations that people make – either based on their own implicit attitudes, overall category valence, or perhaps another mechanism entirely. Thus the question becomes not *whether* valence effects determine IAT effects, but *to what extent* they influence them. This is actually a common feature of not only results specifically examining valence, but also salience asymmetry research generally: reversing the salience of an IAT task can diminish the magnitude of IAT effects but not completely reverse them (Houben, 2007; Rothermund & Wentura, 2004).

The argument presented thus far is that valence can have some effect on IAT scores. Why then do IAT scores also frequently show (at least some) convergent validity with self-reported attitude measures of the same construct (Hofmann et al., 2005a)? And why is it that IATs are often able to predict behaviour corresponding to the attitude they are attempting to measure (Greenwald et al., 2009)? The obvious answer is that the majority of IATs, for the majority of the time, actually do capture the implicit associations they are trying to measure, rather than other features of the categories or category labels. Alternatively, perhaps salience asymmetries do play some role in producing IAT effects, but this role is much less important than that played by implicit associations. In a response to Rothermund and Wentura, Greenwald et al. (2005) argued exactly that: IAT scores are produced mostly by what they call the "nominal features" of the IAT, which they define as the features that are actually indicated by the category labels and that the IAT is supposed to measure. However, even Greenwald et al. (2005) conceded that even

if IAT scores are really indicative of people's real implicit preferences, there may still be some effects of valence. The point that should be taken from the preceding discussion is not that salience or strategic recoding are the main sources of IAT effects, but that – given the right circumstances – they *can* contribute to them.

2.3 The present work

One might argue that stimuli valence is more of an issue for constructs which are already more valenced to begin with. As discussed in Chapter 1, forgiveness, by definition, might be such a construct that has an inherent positive valence, often being referred to positively as a "gift" (Enright et al., 1998) and a "prosocial response" (Bono, McCullough, & Root, 2008; McCullough et al., 2007). Based on this, it could be argued that people may find it easier to associate 'forgiveness' words with 'pleasant' words, purely based on the valence of these words, rather than as a function of their implicit attitudes. Thus one of the main aims of the first set of studies is to determine if strategic recoding based on stimuli valence will be an issue for using a Forgiveness IAT.

2.4 Study 1

2.4.1 Study Overview

This study aimed to develop a forgiveness IAT while, at the same time, attempting to determine if stimuli valence would be a significant factor in producing these IAT effects. The second aim was to assess two key components of the forgiveness IAT's validity: namely its resistance to socially desirable responding, and its convergence with other forgiveness attitude measures. These issues will be addressed shortly.

2.4.2 IAT design

2.4.2.1 Contrast category

As the IAT is a relative measure, the construct of interest must always be paired with an 'opposite' or contrast category. If no opposite exists then the contrast category should be a "sensible, mutually exclusive category that is ideally from the same domain" (Lane et al., 2007, p86). As noted earlier, an exact definition of forgiveness remains elusive, but its conceptual opposite appears even more so; there does not appear to be one single logical choice for its opposite category. There have been several suggestions in the forgiveness literature for potential candidates, with some of the most prominent being revenge and avoidance (McCullough et al., 1998) or holding a grudge (Baumeister et al., 1998). Retributive justice and punishment have also been framed as possible opposites to forgiveness (Exline, Worthington, Hill, & McCullough, 2003).

Revenge will be used in this study as a starting point. Revenge has traditionally been understood by experts as an opposite of forgiveness, with many definitions of forgiveness equating the presence of forgiveness with the absence of revenge (McCullough, 2008; McCullough et al., 1998; Worthington, 2001). Forgiveness has been defined as "a willingness to abandon one's right to resentment, condemnation, and subtle revenge" (Enright & The Human Developmental Study Group, 1991, p. 108), while both McCullough (McCullough, Worthington, & Rachal, 1997) and Worthington (1998) state that it necessarily involves reductions in revenge and retaliation. In reviewing the literature from an interdisciplinary perspective, McCullough (2008) dedicates an entire book to contrasting forgiveness and revenge as evolutionary adaptations, and concludes that they are essentially two sides of the same coin. That is, forgiveness and revenge are

two adaptive processes that arise when they are necessary for cooperation (and by extension, survival). Revenge, he claims, helps to maintain cooperation by discouraging social loafing – any individual who does not 'pull their weight' faces retaliation from the group, and thus revenge plays instrumental roles as both a punishment and a deterrent of future non-cooperative behaviour. Forgiveness is also necessary so that those who offend initially are given the chance to reintegrate in to the group. McCullough (2008) suggests that the two processes do not occur simultaneously, but both are equally important for ensuring our survival as a cooperative species. In this way, forgiveness and revenge fit Lane et al.'s (2007) criteria as an appropriate IAT pairing – they are mutually exclusive categories from the same domain.

One has to look no further than the way that forgiveness is frequently operationalised to see that forgiveness and revenge are often treated as dichotomous constructs. One of the most popular measures of forgiveness – the TRIM (McCullough et al., 1998) – includes a revenge scale as a proxy for forgiveness: the scale is interpreted such that if a person reports that they are less vengeful, then they are more forgiving. Several other forgiveness scales also include items measuring revenge, which are then reverse-scored and interpreted as being indicative of forgiveness attitudes or tendencies (although this is not done quite as comprehensively as in the TRIM). For example, Rye's (2001) Forgiveness Scale includes the item "I spend time thinking about ways to get back at the person who wronged me", while the Heartland Forgiveness Scale (Thompson et al., 2005) includes the item "I continue to punish a person who has done something that I think is wrong". The same is also true in reverse, with measures of revenge also including items that explicitly tap forgiveness such as "I find it easy to forgive those who have hurt

me" (Stuckless & Goranson, 1992). This common interpretation among researchers that less revenge equates to more forgiveness (and vice versa) suggests that revenge may be considered an appropriate contrast category to pair with forgiveness in the IAT.

2.4.3 Assessing the validity of the Forgiveness-Revenge IAT

2.4.3.1 Assessing the impact of stimuli valence on the forgiveness IAT

The present study aimed to examine the extent to which valence of target stimuli might contribute to forgiveness-revenge IAT effects, by experimentally manipulating target category labels. Fortunately, definitional debate surrounding forgiveness has established that there are words which are often used by people to describe forgiveness that have inherent positive (e.g. reconciliation) or negative (e.g. condone) connotations. Revenge, on the other hand, is more problematic, as there are very few – if any – positively valenced words that could be used to represent it. Consequently, this study will only manipulate the valence of stimuli for the forgiveness category. There will be three conditions: one in which forgiveness is represented by six positively valenced words, one where forgiveness is represented by six negatively valenced words, and a third "balanced" condition where there are three positively and three negatively valenced words representing the category. All words for revenge will be negatively valenced. Following standard IAT procedure outlined by Greenwald et al. (1998), the evaluative dimension of the IAT will consist of the categories 'pleasant' and 'unpleasant', and the words used to represent these categories will remain constant across the three IAT variants. If valence of forgiveness stimuli is an issue then a significant difference in mean scores between the three IATs would be expected, with the positive-valence condition

producing the highest and the negative-valence condition producing the lowest IAT scores.

2.4.3.2 Socially desirable responding and the IAT

As outlined in Chapter 1, one of the major appeals of developing an IAT for forgiveness is that it purports to be resistant to socially desirable responding (SDR). For detecting SDR on self-report measures, a common method is to use a SDR scale, such as that developed by Crowne and Marlowe (1960), or any of a large number of (shorter) derivations. These scales generally consist of lists of two kinds of behaviours; (a) behaviours that are rare but socially desirable, and (b) behaviours that are common but socially undesirable. A person who indicates that they engage in many of the former and few of the latter types of behaviour is deemed to be high in socially desirable responding.

In assessing whether scores on an *explicit* self-report measure have been influenced by SDR one can first examine bivariate correlations – the scale measuring the construct of interest should not significantly positively correlate with the SDR scale. If it does correlate then variance due to SDR can be partialled out before examining the relationships further (Paulhus, 1986; Stober, 2001). Theoretically, SDR scales should be able to be used in a similar manner with the IAT. As the IAT purports to be immune to SDR concerns, it should be expected that the IAT and SDR measures should not be significantly correlated. Furthermore, if the explicit forgiveness measures and SDR are correlated, then partialling out the effects of SDR should improve the correlations between the explicit and implicit forgiveness measures. To the author's knowledge, this has only been attempted once in the IAT literature, with an anxiety IAT (Egloff &

Schmukle, 2003). That study found that SDR did not significantly moderate the relationship between implicit and explicit anxiety, but suggested that this may be because anxiety may not provide the same motivations for people to hide their true attitudes that some of the more socially sensitive topics like race or aggression (and perhaps forgiveness) might. This study seeks to determine if the Forgiveness-Revenge IAT will be affected by socially desirable responding and, if so, examine the role of SDR in moderating the relationship between the IAT and self-reported forgiveness scales.

2.4.3.3 Convergent validity: Implicit-Explicit (IE) correspondence

Meta analyses of IAT studies have shown that the average degree of convergence between IAT and corresponding self-report measures is in the vicinity of r = .23 (152 independent samples; Greenwald et al., 2009) and r = .24 (126 independent samples; Hofmann et al., 2005a). However, within the IAT literature there are also large amounts of variation in implicit-explicit (IE) measure convergence. This is often a function of the nature of the construct under investigation (Hofmann et al., 2005a; Nosek, Greenwald, & Banaji, 2005). For example, studies which have used the IAT to measure consumer preferences (e.g. Maison, Greenwald, & Bruin, 2004; Scarabis, Florack, & Gosejohann, 2006) often find IE correspondence up to around .40, while domains like political preference (e.g. Nosek & Hansen, 2008b; Olsen & Fazio, 2004) have shown IE correspondence as high as .70.

However, such IE correspondence is the exception rather than the rule, with correlations usually falling below .30 (Hofmann & Schmitt, 2008). One argument for low correspondence is that the two measures capture different elements of the same

construct (Fazio & Olsen, 2003). Another argument is that the discrepancy can be explained by the error variance accounted for by each of the two measures. As already discussed, the IAT may be influenced by extra-attitudinal factors such as category salience or strategic recoding, while explicit measures may be influenced by self-presentation concerns, or a lack of introspective access to "true" thoughts or feelings (Greenwald et al., 1998). For the moment it is not necessary to explore these mechanisms further, but a more detailed discussion of IE correspondence will be provided in Chapter 4. For the present study, it is expected that the Forgiveness-Revenge IAT will be modestly correlated with explicit measures of forgiveness attitudes.

2.4.4 Summary of hypotheses

Valence

The effects of valence on the forgiveness IAT will be assessed using analysis of variance, with valence content of IAT (forgiveness-positive, forgiveness-negative, forgiveness-balanced) as the independent variable and the IAT score as the dependent variable. Two competing hypotheses are being tested, with the following possible outcomes:

a. If the IAT scores are purely a function of valence issues and not forgiveness attitudes it would be expected that IAT scores would be significantly different between the three conditions, with IAT scores being highest in the forgiveness-positive condition, and lowest in the forgiveness-negative condition.

b. If the IAT scores are purely a function of forgiveness attitudes and valence issues are of no concern it would be expected that there would be no significant differences between IAT scores across the 3 conditions.

Socially Desirable Responding

The relation of SDR to the Forgiveness-Revenge IAT will be assessed using correlational analyses. It is hypothesized that:

- The IAT will be resistant to socially desirable responding such that there will be no correlation between the IAT and a measure of socially desirable responding.

Convergent Validity

Correlational analyses will be used to explore the relationships between implicitly measured (IAT) and explicitly measured (self-report scale) forgiveness. It is hypothesised that the Forgiveness-Revenge IAT will significantly, but modestly, correlate with self-report measures of forgiveness attitudes.

2.4.5 Method

2.4.5.1 Design

This study utilised a between-groups experimental design.

2.4.5.2 Participants

Participants were 156 (103 female, 53 male) first year undergraduate psychology students at the University of Adelaide, Australia, who participated in exchange for course credit. Mean age for the sample was 20.2 years (*SD*=4.17).

2.4.5.3 Materials

2.4.5.3.1 IAT design

There were three IAT conditions, all with target categories of forgiveness-revenge and attribute categories of pleasant-unpleasant. The three conditions differed only in the words used to represent forgiveness. The first represented forgiveness with six 'positively' valenced words: absolve, compassion, mercy, empathy, reconcile, amnesty. The second used six 'negatively' valenced words: excuse, overlook, condone, justify, lenient, ignore. The third IAT condition was 'valence-balanced', representing forgiveness with three 'positive' and three 'negative' words: absolve, compassion, mercy, excuse, overlook, condone. All IATs used the same words for revenge (retaliate, vengeance, retribution, pay-back, vindictive, reprisal), pleasant (freedom, love, cheer, pleasure, gift, happy), and unpleasant (abuse, filth, hatred, poison, evil, tragedy)

2.4.5.3.2 IAT stimulus word selection

The stimulus words for forgiveness and revenge were selected from several thesauruses with consideration to the guidelines for stimulus word selection, presented in a comprehensive review of the IAT by Nosek, Greenwald, and Banaji (2007), as well as through examination of the findings of Kearns and Fincham's (2004) and Friesen and

Fletcher's (2007) prototype analyses. The stimulus words for pleasant and unpleasant were a smaller subset of those used by Greenwald et al. (1998).

The number of words per category was constrained by the nature of the forgiveness construct. In an effort to minimise the possibility that participants might strategically recode the IAT task it seems sensible to include as many stimulus items as possible to represent each category. At the same time, the functioning of an IAT also relies on the premise of using exemplars that are *unambiguously* related to the chosen category, to ensure that longer delays in reaction time are due to the associations with the construct under investigation, and not merely a result of people trying to decipher other features of the word (Lane et al., 2007). Forgiveness has very few (if any) direct synonyms, and relatively few words that can be unambiguously related to it, thus number of stimulus words was restricted to six per category.

This number of words should not be problematic as there is evidence to show that IAT effects should not differ greatly as a function of stimulus set size. In their original paper, Greenwald et al. (1998) demonstrated similar IAT effects, regardless of whether there were 5 or 25 items per category. More recently, Nosek et al. (2005) reviewed data from 11 studies, employing IATs of differing stimulus set sizes, and concluded that valid IAT scores could be attained with as few as 2 items per category (although using only one per category made it easy for participants to strategically recode the task).

2.4.5.3.3 Structure of the IAT

The Forgiveness-Revenge IAT followed standard procedures as outlined by Greenwald and colleagues (Greenwald et al., 1998), with the exception of the number of

trials per block. The present study used blocks of 24 and 48, rather than 20 and 40 respectively, as this more easily accommodated the number of word exemplars used in the IATs – in this case six per category, ensuring each exemplar was presented an equal number of times. The sequence and category pairings of the IAT trials are outlined in Table 2.1.

Table 2.1

Sequence of Trial Blocks in the IAT

Block	Classification Task	No. of trials	Function	Response Key Assignment	
				Left Key	Right Key
1	Initial Target Discrimination	24	practice	forgiveness	revenge
2	Initial Attribute Discrimination	24	practice	pleasant	unpleasant
3	Initial Combined Task	24	practice	forgiveness	revenge
				pleasant	unpleasant
4	Initial Combined Task	48	test	forgiveness	revenge
				pleasant	unpleasant
5	Reversed Target Discrimination	24	practice	revenge	forgiveness
6	Reversed Combined Task	24	practice	revenge	forgiveness
				pleasant	unpleasant
7	Reversed Combined Task	48	test	revenge	forgiveness
				pleasant	unpleasant

Before completing the IAT, participants were given the following instructions:

> *" In the next task, you will be presented with 4 sets of words and you will be asked to sort them into groups. You will be asked to sort them as quickly as you can, but try to keep your error-rate as low as possible. Going too slowly or making too many mistakes will mean that it will not be possible to interpret your results.*

Please read through the following category labels and the items that belong to each of the categories. You will be asked to sort the words according to these categories."

Participants were then provided with a table listing the four categories and the exemplars that belonged to each.

A correct response on any given trial meant that the participant proceeded immediately to the next trial. A trial was considered to be incorrect if (a) the participant sorted the stimulus word in to the wrong category, or (b) the participant took 10000 milliseconds or longer to respond. In both of these cases the trial was recorded as an error trial, and the participant was then given a chance to correct their response before the task proceeded to the next trial.

2.4.5.3.4 IAT reliability

Internal consistency reliability for IATs is usually calculated based on the log-transformed differences in responses latencies on corresponding trials between the compatible (blocks 3 and 4) and incompatible (blocks 6 and 7) blocks, e.g. the first trial of block 6 minus the first trial of block 3; the sixteenth trial on block 7 minus the sixteenth trial on block 4. These difference scores are then treated as individual scale items, and Cronbach's alpha calculated for them accordingly (for a more detailed explanation of these steps see Egloff & Schmukle, 2003, p.1701). For the Forgiveness IAT this resulted in 72 differences scores, and Cronbach's alpha for these scores was high ($\alpha = .93$).

2.4.5.3.5 Self-report questionnaire

Participants completed online versions of the Attitudes To Forgiveness (ATF) and Tendency To Forgive (TTF) scales (Brown, 2003), the 'other' subscale of the Heartland Forgiveness Scale (HFS, Thompson et al., 2005), Stober's (2001) Social Desirability Scale (SDS-17), and demographic items (age and gender). The first three of these measures were used to assess participants' forgiveness attitudes and dispositions.

Attitudes Toward Forgiveness (ATF)

The ATF (Brown, 2003) is a measure of general attitudes toward forgiveness. The scale consists of 6 items, rated on a 7 point Likert-type scale and includes statements such as "It is admirable to be a forgiving person" and "Forgiveness is a sign of weakness" (negatively-coded). Scale items were summed to produce total scores ranging from 6 to 42, with higher scores reflecting more positive attitudes towards forgiveness. Internal consistency was borderline ($\alpha = .63$).

Tendency To Forgive (TTF)

The TTF (Brown, 2003) is a 4 item scale, designed to measure forgiveness at a dispositional level. Unlike the ATF, the TTF aims to assess the degree to which people believe that they are actually able to forgive generally, their ability to 'let go' of hurts/grudges (e.g. "I have a tendency to harbour grudges", negatively-coded) and to move forward from a transgression (e.g. "I tend to get over it quickly when someone hurts my feelings"). It was designed to be distinct from attitudinal measures of forgiveness, and has been shown to be associated with outcomes such as life satisfaction,

depression, and state forgiveness, independently of actual forgiveness attitudes (Brown & Phillips, 2005). Also, quite deliberately, the TTF does not include items that address revenge, as Brown believed that current measures often confound absence of revenge with presence of forgiveness. Items are rated on a 7 point Likert-type scale, from (1) strongly disagree to (7) strongly agree, with item scores summed to produce a total range between 4 and 28, with higher scores indicative of higher dispositional forgiveness. The scale showed good internal consistency (α = .81).

Heartland Forgiveness Scale (HFS) – "other" subscale

The HFS (Thompson et al., 2005) aims to measure forgiveness at a general/dispositional level, and is comprised of three separate subscales; one focusing on self-forgiveness, one on forgiveness of others, and one on forgiveness of situations, with each containing 6 items. As the present thesis is only concerned with *interpersonal* forgiveness, only the forgiveness of others subscale was used. The 6 items, each rated on a 7 point Likert-type scale (from "almost always false of me" to "almost always true of me"), focus on the ways in which one generally thinks about a transgressor: how much they want to punish or see that person hurt (e.g. "I continue to punish a person who has done something that I think is wrong"), think badly of that person (e.g. "If others mistreat me, I continue to think badly of them") or show empathy and understanding for them (e.g. "With time I am understanding of others for the mistakes they've made"). Item scores were summed to produce a scale total ranging from 6 to 42, with higher scores reflecting more dispositional forgiveness of others. Internal consistency was acceptable (α = .77).

Social Desirability Scale (SDS-17)

Based on the social desirability scale developed by Crowne and Marlowe (1960), and following the same basic premises, the the SDS-17 (Stober, 2001) consists of 17 items that describe an action or behaviour and the respondent must nominate whether this action or behaviour is true or false of them. The items can be divided in to two categories – behaviours that are considered socially desirable but infrequent (e.g. "In conversations I always listen attentively and let others finish their sentences"), and behaviours that are considered socially undesirable but frequent (e.g. "I occasionally speak badly of others behind their back"). After reverse-coding items as appropriate, items are summed to produce a total that ranges from 0 to 17, with higher scores indicative of a greater tendency to respond in a socially desirable manner. Internal consistency reliability was borderline ($\alpha = .62$).

Demographic items

The final section of the questionnaire asked for demographic information regarding age and gender.

2.4.5.3.6 Online administration

The majority of IAT research is conducted online. Houben and Wiers (2008) found that IAT results did not differ depending on whether they were done in a controlled setting or at home. Furthermore, it seemed that IE correspondence was higher for those done at home, perhaps owing to lower self-presentation pressures in a non-face-to-face

setting. Following from this, and for ease of administration, the study was made available to participants online and could be completed at a time of their choosing.

2.4.5.4 Procedure

Student participants were recruited via a centrally-managed online research participation website, and through email advertisements. Once the participant had registered their interest on the website they were immediately provided with a web link[4] to the study which could be completed entirely online at a time of their choosing. Following the web link first took participants to an information page explaining the nature of the study, with a button for them to click to indicate their informed consent to participate in the study. Once they had clicked this button they were immediately taken to the self-report questionnaire. After completing the questionnaire, participants were then randomly allocated to one of the three IAT conditions. Upon completion of the IAT participants were taken to a final screen thanking them for their time.

2.4.6 Results

2.4.6.1 Data Preparation and IAT scores

IAT D scores (the current standard for IAT scoring) were computed following the revised scoring algorithm outlined by Greenwald et al. (2003). The algorithm includes

[4] All studies in this thesis relied on online completion of IAT and explicit attitude scales. At the time of thesis submission, the web links for studies 6 through 9 (for which the IAT and explicit attitude components were very similar to the present study) were still available for viewing online. Should the reader wish to view and/or complete these materials first-hand, they are referred to the web links for Study 6: http://www.psychology.adelaide.edu.au/expts/hurt1.html

steps to address extreme fast (<300ms) responses (which indicate that the participant may have not been taking the task seriously), as well as taking into account the number of errors made by participants. Participants for whom more than 10% of responses were faster than 300 milliseconds (N=20) were excluded from the calculations and subsequent analyses, leaving 136 participants. Before calculating D scores, response latencies for each error trial were replaced with the mean latency for that block plus a 600 millisecond penalty. D scores were then calculated by subtracting means latencies for blocks 3 and 4 from blocks 6 and 7, and then dividing by the pooled standard deviation of these four blocks.

D scores have a possible range between -2 and +2, with a score of zero indicating no/equal preference for the two target constructs. In this study the categories were coded such that positive D scores indicated an implicit preference for forgiveness relative to revenge, whereas negative D scores reflected an implicit preference for revenge relative to forgiveness. The greater a score deviates from zero, the stronger the implicit preference. Mean D scores for each IAT version, and overall, are presented in Table 2.2. Across all three conditions there was an implicit preference for forgiveness relative to revenge.

Table 2.2

Means, 95% Confidence Intervals and Standard Deviations for IAT D scores

	N	*M* (95% CIs)	*SD*
IAT (positive)	50	.90 (.73–1.08)	.62
IAT (balanced)	41	.74 (.57-.91)	.54
IAT (negative)	45	.72 (.58-.85)	.46
IAT (combined)	136	.79 (.70-.89)	.55

2.4.6.2 The effects of valence on IAT scores

A one-way ANOVA was conducted to determine if there were differences between the different IAT versions as a function of stimulus word valence. The ANOVA revealed that there were no significant differences between the three variations of the IAT[5], $F(2,133)=1.65$, $p=.20$, $\omega=.16$.

2.4.6.2.1 Confidence Intervals

Although the analysis of variance failed to detect significant differences between the IAT conditions, this does not mean that they do not exist. The nature of null-hypothesis significance testing (NHST) is such that a non-significant difference cannot be used as evidence for the absence of an effect. That is, a lack of evidence for a hypothesis does not provide evidence that the difference does not exist, it merely provides grounds for uncertainty (Frick, 1996; Greenwald, 1975). In fact there are numerous other factors which may contribute to a null effect, such as insufficient power, small sample size, or any

[5] The distributions of scores for both the IAT-positive and IAT-balanced conditions were negatively skewed, violating the assumptions of ANOVA. To address this, a negative square root transformation was applied, which improved the distribution of scores for the forgiveness-balanced condition, but not for the forgiveness-positive condition. The transformation also failed to significantly alter the outcomes of the ANOVA, with the difference between the three conditions still not reaching significance, $F(2,133)=2.08$, $p=.13$, $\omega=.18$. As this transformation did not alter the outcome of the ANOVA, the original (pre-transformation) values for the IAT scores were retained for the remainder of the analyses.

number of extraneous variables which may contribute to increased error variance in the model.

Accordingly, many statisticians recommend examining confidence intervals (Cumming, 2008; Gardner & Altman, 1986), and reporting of confidence intervals is also now recommended as a research standard in psychology (APA, 2001). Confidence intervals were calculated separately for the three pairwise comparisons. In order to maintain a constant alpha level of .05 across three independent comparisons, the confidence intervals were adjusted from 95% to 98.3% for each. The results are presented in Table 2.3 (below). As the confidence intervals for all three comparisons include zero, one cannot be confident that there are any meaningful differences between the three conditions.

Table 2.3

98.3% Confidence Intervals for Pairwise Comparisons of 3 IAT Variants

Comparison	Lower bound	Upper bound
IAT (positive) X IAT (balanced)	-.14	.47
IAT (balanced) X IAT (negative)	-.29	.24
IAT (positive) X IAT (negative)	-.09	.46

In summary, valence of stimuli used to represent forgiveness in the IAT appeared to play a negligible role in determining scores on a forgiveness-revenge IAT. As there were no significant differences between IAT versions the three conditions were collapsed for the remainder of the analyses.

2.4.6.3 Relationship between social desirability and implicit and explicit forgiveness

Pearson correlations (see Table 2.4) were computed to determine if SDR was significantly related to both the forgiveness-revenge IAT and the self-report forgiveness measures. Consistent with hypothesis, there was no significant correlation between the IAT and the SDS-17. Additionally, the SDS-17 was not significantly correlated with any of the self-report forgiveness scales.

Table 2.4

Intercorrelations between IAT D Scores, Self-Report Measures of Forgiveness Attitudes, and Socially Desirable Responding

	1	2	3	4	5
1. IAT	--				
2. Attitudes To Forgiveness (ATF)	.22*	--			
3. Tendency To Forgive (TTF)	-.11	.21*	--		
4. Heartland Forgiveness Scale (HFS)	.18*	.51**	.53**	--	
5. Social Desirability Scale (SDS-17)	-.10	-.00	.02	.15	--

* $p<.05$. ** $p<.01$.

2.4.6.4 Relationships between IAT D scores and self-report measures of forgiveness

As shown in Table 2.4, there were significant small positive correlations between IAT D scores and two of the self-report forgiveness measures; the ATF and HFS. There was no significant correlation between IAT D scores and TTF.

As expected, the three forgiveness scales all significantly correlated with each other. Finally, none of the explicit forgiveness measures were associated with SDR.

2.4.7 Discussion

The present study had two aims. The first aim was to assess whether the valence of IAT stimuli used to represent forgiveness would significantly influence the way in which people responded to the forgiveness-revenge IAT. Results indicate that IAT scores did not significantly differ as a function of IAT stimuli valence. The second aim was to assess the validity of the forgiveness-revenge IAT on two fronts: the extent to which the IAT was resistant to socially desirable responding and the degree to which the IAT converged with other (self-report) measures of forgiveness attitudes. As hypothesised, there was no significant relationship between the IAT and socially desirable responding tendencies, although SDR was also unrelated to other forgiveness measures. The forgiveness-revenge IAT showed modest convergent validity with self-reported forgiveness measures.

One of the key validity concerns with using the IAT as an attitudinal measure is that it should not be affected by incidental features of the IAT design. The valence of the stimulus words used for the forgiveness category was one of these features, and was of particular interest in this study, in light of IAT research showing that word valence can sometimes account for IAT effects (Govan & Williams, 2004; Rothermund & Wentura, 2004), and the argument that forgiveness may have an inherently positive valence. Despite these concerns, there were no significant differences in IAT scores as a function of word valence, with an examination of confidence intervals further supporting this claim.

Further evidence that the IAT effects were being determined by responses to nominal features of the stimuli, rather than being methodological artefacts, comes from exploring the relationships between implicit and explicit forgiveness attitudes. Modest correlations between the IAT and the two attitudinal self-report measures of forgiveness (ATF, *r*=.22; HFS, *r*=.18) are consistent with those found in much of the IAT literature (Greenwald et al., 2009; Hofmann et al., 2005a). This is what would have been expected for implicit and explicit measures that are measuring the same basic construct, while at the same time accessing different aspects of it. Although this does not discount the effects of valence, it does provide some evidence for the convergent validity of the Forgiveness-Revenge IAT, that it is to some extent measuring the forgiveness construct. Furthermore, IE convergence provides evidence that the Forgiveness-Revenge IAT effects are at least partially accounted for by the *association* model, or nominal features of the IAT.

Despite the IAT's convergence with two of the explicit forgiveness measures, it did not significantly converge with the TTF. The reasons for this may be illuminated by examining how the three self-report forgiveness scales correlated amongst themselves. The most noteworthy of these relationships was the small correlation between the ATF and TTF (*r*=.21). The small magnitude of this correlation is consistent with Brown's (2003) data, which showed correlations of between .33 (study 3) and .37 (study 4), suggesting that these scales measure two distinct (yet related) constructs. This makes sense, as the TTF aims to assess the extent to which a person *practices* forgiveness, whereas the ATF examines more the extent to which a person views forgiveness as *desirable*, irrespective of whether they believe themselves to practice it. To rephrase, the TTF aims to assess

forgiveness *behaviour*, while the ATF aims to assess forgiveness *attitudes* (Brown & Phillips, 2005). Given that the IAT is itself an *attitudinal* measure, it is perhaps unsurprising that it would correlate most strongly with explicit measures of forgiveness *attitudes*.

An alternative explanation for no relationship between the IAT and TTF may relate to the degree to which these measures assess *revenge*. The IAT used in this study included "revenge" as the paired category for "forgiveness", and as such, scores on the IAT should be indicative of participants' preferences for both forgiveness and revenge. The self-report scales, on the other hand, simply claimed to measure forgiveness, but there may have been subtle differences here between the three scales. In particular, Brown (2003) stated that the TTF was designed to measure forgiveness with a concerted effort to *not confound it with revenge,* and a brief scan of the scale's items reveals that it indeed includes no revenge-relevant content. In contrast, both the HFS and ATF each include at least one item that is more closely aligned with revenge. This is quite overt in the HFS, with the item "I continue to punish a person who has done something that I think is wrong" essentially providing a direct measure of revenge – the motivation to punish someone who has hurt you. It is therefore perhaps unsurprising that a self-report measure that taps in to notions of revenge would correlate more with a Forgiveness-Revenge IAT than a measure that does not.

In addition to convergent validity, another aspect of the forgiveness-revenge IAT validity that is of interest is its resistance to social desirability factors. One of the appealing advantages of the IAT over self-report is its apparent resistance to the effects of socially desirable responding (Greenwald et al., 2009; Lane et al., 2007). Results from

the present study provide further support for this position, with the Forgiveness-Revenge IAT showing no significant correlation with the SDS-17. However, data from the present study does not provide insight in to whether the IAT is more resistant to self-presentation than other measures of forgiveness, as none of the self-report measures correlated with the SDS-17 either.

One explanation for these null correlations is that the ATF, TTF and HFS may be generally robust to social desirability pressures. This is difficult to assess, however, due to the lack of available evidence in the forgiveness literature. Most of the available studies which use the TTF (e.g. Brown, 2003, 2004; Brown et al., 2007; Brown & Phillips, 2005; Eaton, Struthers, & Santelli, 2006; Hu, Zhang, & Ja, 2005), the ATF (e.g. Brown, 2003; Brown et al., 2007; Brown & Phillips, 2005; Eaton et al., 2006; Hu et al., 2005), or the HFS (e.g. Day & Maltby, 2005; Macaskill, 2007) did not control for socially desirable responding. In the one study using the HFS where social desirability was examined (Thompson et al., 2005), the "other" subscale of the HFS was found to correlate significantly (.34) with the Marlowe-Crowne Social Desirability scale (Crowne & Marlowe, 1960), which is in contrast to the present findings. The only known study to examine the impact of SDR on the ATF and TTF (Powers, Nam, Rowatt, & Hill, 2007) did find significant correlations between an impression management scale and both the ATF (r=.38, p<.001) and the TTF (r=.27, p<.01), although this has not been replicated. Of course the lack of correlation in the present study does not mean that SDR was not important – just that it was not detected using the SDS-17.

The results of the present study are encouraging for the further investigation and development of an IAT for measuring forgiveness. Importantly, the Forgiveness-Revenge

IAT has passed some initial tests of validity: it has displayed construct validity through being resistant to both strategic recoding by valence and social desirability concerns, and it has displayed a degree of convergent validity that is consistent with findings in the IAT literature.

However, it is also possible that the small correlations between the IAT and forgiveness scales could be interpreted as evidence against the validity of the IAT: that is, these correlations are only small because the IAT lacks construct validity. This argument has, in fact, sometimes been used to criticize the low levels of IE convergence that are typically found in IAT studies (see Hofmann & Schmitt, 2008, for a review of this argument). One particular factor which may have affected the construct validity of the Forgiveness-Revenge IAT in the present study is word selection. Words included in the IAT were selected subjectively by the researcher, and although the selection process was done with much thought and based on words found in the forgiveness literature, it is possible that these are not the best words for capturing the forgiveness construct. As noted earlier, there is now much evidence to suggest that the stimuli used to represent a construct within the IAT matter. As such, the outcomes of the present study may have been different if an alternative set of stimulus words had been used.

There is also a need to further develop our understanding of both convergent validity and social desirability factors in relation to the Forgiveness-Revenge IAT. The self-report attitude measures used in this study were limiting because they only assessed forgiveness, while the IAT measured attitudes to both forgiveness and revenge. One way this issue can be addressed initially is by the inclusion of a self-reported measure of revenge, in addition to forgiveness measures. Our understanding of the impact of self-

presentation concerns was also limited in this study by the relatively poor scale reliability (α = .62) of the SDS-17. As such it may be beneficial to consider an alternative measure of SDR.

In summary, there is a need to replicate the findings of the present study using an alternative set of stimulus words, and some alternative measures. The remainder of this chapter will address these concerns by (a) running a small study to pilot test some alternative words to represent forgiveness, and (b) based on the results of this pilot study, attempt to replicate the findings of the present study using an alternative set of stimulus words for the Forgiveness-Revenge IAT, and including additional measures of revenge and SDR.

2.8 Study 2a

2.8.1 Overview of Study

There is a possibility that the stimulus words used to represent forgiveness in Study 1 were not the most ideal choice, either in terms of their representatives of the forgiveness category and in their assigned valence. These words were chosen with reference to the forgiveness literature, along with several thesauruses, but it has already been established that the way in which theorists define forgiveness may in fact be different to how lay persons do. However, this was hopefully offset by the fact that many of these words came from two prototype analyses, which specifically asked lay people to nominate how central they believed words to be to the forgiveness construct (Friesen & Fletcher, 2007; Kearns & Fincham, 2004). Therefore, there is good reason to be confident

that *representativeness* of stimulus words should not have affected the Forgiveness-Revenge IAT's construct validity.

In contrast, assigning *valence* to stimulus words was much less of an exact process. Prototype analyses only address how *central* a word is to a category, without asking for any indication about whether that word is seen as positively, negatively or neutrally related to the category. Consequently, word valence was assigned subjectively by the researcher, and may have been done so inaccurately.

In order to address this issue it was deemed appropriate to attempt to replicate the findings of Study 1 using a different set of words to represent forgiveness: words that had been pilot tested for both centrality *and* valence.

2.8.2 IAT word selection

In reviewing the literature on the IAT it becomes apparent that there are no fixed conventions or expectations for justifying the words used to represent each target category. In their original paper, Greenwald et al. (1998) based their stimulus word selection on previously published norms (Bellezza, Greenwald, & Banaji, 1986) and lists of category memberships (Battig & Montague, 1969), but also applied some level of subjective judgment, with the final lists of words being "ones that the authors judged to be both familiar to and unambiguously classifiable by members of the subject population" (p.1466). While some other authors have similarly made an effort to justify their selection of stimuli, this practice appears to be the exception rather than the rule: the majority of papers on the IAT do not provide explanations or criteria for how the sets of words were selected. Perhaps ironically, some of the very people who challenge the

word choice of the IAT (e.g. Govan & Williams, 2004; Rothermund & Wentura, 2004) do not justify how they themselves came to choose their words.

This trend of not justifying IAT word selection seems odd considering there is now considerable evidence to suggest that the stimuli chosen can have an impact on subsequent IAT scores. In addition to the salience and valence concerns already outlined in this chapter, several published reviews have identified features of stimulus word selection that can impact IAT effects, and outline some basic principles for choosing appropriate words. Reiterating Greenwald et al.'s (1998) original assertions, both Lane et al. (2007) and Nosek et al. (2007) suggest that IAT stimuli should be (a) easily and quickly identifiable to the participant population, i.e. the participants need to actually understand what all of the words mean; (b) not negated, i.e. adding the prefix "un", e.g. "unfaithful", increases reaction time purely as a function of processing the negation, rather than indicating a more difficult attitudinal association; and (c) unambiguously related to only one of the four target categories. If these conditions are not met then there is an increased chance of reaction times being affected by factors other than a person's implicit preferences.

Although not overtly stated, it is quite possible that many IAT researchers do in fact carefully consider these principles when designing IATs. However, returning to the rationale offered by Greenwald and colleagues (1998), word selection relies on a subjective judgement by the researcher. In many situations this may be a sensible approach, and the researcher will have a similar understanding of the target concepts to that of their respective participant pool. However, in other cases, researchers may have a different perspective on these things to the average person who participates. This might

be particularly true for an area such as forgiveness, where there is ongoing debate about its precise definition, and there are known discrepancies in understanding between theorists and lay people (e.g. Kearns and Fincham, 2004).

In light of this, a pilot study was conducted to explore the extent to which people associate different words with forgiveness, with the aim of using these words to generate an alternative version of the Forgiveness-Revenge IAT. The study had two main aims. The first aim was to determine whether or not people associated various words as being representative of forgiveness (centrality). The second aim was to determine if these words were associated with forgiveness positively, negatively or neutrally (valence). As the study was exploratory, no specific hypotheses were set, however responses would be used to generate an alternative list of forgiveness words to those used in Study 1, suitable for use in constructing a forgiveness IAT.

2.8.3 Method

2.8.3.1 Participants

Participants were recruited using a variation on 'snowball' sampling methodology (Goodman, 1961). Participation was requested via an email invitation to members of the researcher's personal email distribution list, which consisted of 49 email addresses. The invitation asked the recipient to (a) participate in the study and (b) pass this invitation on to their email distribution lists. Thirty-three valid responses were received (23 female, 10 male).

2.8.3.2 Materials

A group of six researchers (all familiar with the literature on forgiveness) at the University of Adelaide generated a list of words deemed to be representative of forgiveness. Each member of this group was initially instructed to think as broadly as possible, in order to generate the largest possible list. This list was then culled by assessing each word in relation to the principles of IAT word selection outlined in a review of the IAT by Nosek et al. (2007), which includes specifications for category exclusivity, category representativeness, word difficulty and familiarity. This resulted in a final list of 29 words in total: Absolve, Acceptance, Altruism, Allow, Amnesty, Appease, Benevolence, Compassion, Condone, Empathy, Excuse, Forget, Generosity, Goodwill, Ignore, Indebtedness, Justify, Lenience, Letting go, Mercy, Moving On, Overlook, Pardon, Permit, Reconcile, Reprieve, Sacrifice, Tolerance, and Understanding.

2.8.3.3 Procedure

The list of words (presented in alphabetical order) was emailed to the researcher's email list following the snowball procedure outlined above. Participants were asked to assess the words and assign each of them to one of four nominal categories, using the labels 1 through 4:

1. The word positively relates to forgiveness.
2. The word negatively relates to forgiveness.
3. The word relates to forgiveness, but is neither positive nor negative
4. The word does not relate to forgiveness at all.

2.8.4 Results

The tables below arrange the stimulus words in to two subsets – words that received more 'positive' votes than of the other three options (Table 2.5), and words that received more 'negative' votes than any of the other three response options (Table 2.6). Each word could receive a maximum 33 votes. Five words did not fit either of these criteria: forget, sacrifice, justify, permit and allow. 'Forget' was judged to be neutrally (i.e. label 3) associated with forgiveness with 10 votes, 'sacrifice' had a completely even split (11 all) between positive and negative, while 'justify' was judged to be either negative or neutral (with 11 votes each). Permit and Allow were predominantly judged to be not related to forgiveness, with 12 and 11 votes respectively. As can be seen from the tables below, forgiveness-related words were largely seen as being more positive than negative.

For ease of interpretation, differential scores for valence and centrality were calculated for each word by subtracting the number of negative votes from the number of positive votes. As such, a positive score signals that a particular word was evaluated as being more positive than negative, with the opposite being true for negative differential scores. A record was also kept of the number of participants who believed that the word was unrelated to forgiveness. These data are presented in Table 2.7.

Only one word - "forget" - received more neutral responses than positive or negative ones. "Permit" and "allow" were the only words with more "not related" votes than anything else. These three words were excluded from consideration for the IAT.

Table 2.5

Positive Forgiveness Words (Response Frequencies)

Word	Votes (max 33)	% of total
Reconcile	31	93.9
Understanding	30	91.0
Letting Go	29	87.9
Compassion	29	87.9
Acceptance	29	87.9
Moving On	27	81.8
Empathy	26	78.8
Mercy	25	75.8
Amnesty	21	63.6
Absolve	20	60.6
Tolerance	19	57.6
Good Will	19	57.6
Pardon	18	54.5
Benevolence	16	48.5
Generosity	15	45.5
Reprieve	13	39.4
Altruism	11	33.3

Table 2.6

Negative Forgiveness Words (Response Frequencies)

Word	Votes (max 33)	% of total
Ignore	22	66.7
Overlook	16	48.5
Excuse	15	45.5
Indebtedness	13	39.4
Appease	12	36.4
Condone	10	31.3

Table 2.7

Forgiveness Words (as Differentials in Positive/Negative Responses)

Word	Differential	Not related
Reconcile	+31	1
Understanding	+30	2
Acceptance	+29	0
Compassion	+28	2
Letting Go	+26	0
Empathy	+26	3
Moving On	+25	0
Mercy	+21	1
Amnesty	+21	4
Goodwill	+19	5
Absolve	+16	2
Generosity	+15	11
Benevolence	+14	8
Tolerance	+12	4
Pardon	+12	3
Reprieve	+7	4
Altruism	+6	8
Lenience	+1	5
Sacrifice	0	6
Forget	-1	6
Condone	-3	8
Allow	-3	11
Appease	-5	6
Justify	-5	5
Permit	-5	12
Excuse	-9	4
Indebtedness	-9	9
Overlook	-11	7
Ignore	-22	8

2.8.5 Discussion

The results suggest that the stimulus words used for the IAT in Study 1 were appropriate choices, and representative of the forgiveness construct. Five of the six negatively-valenced forgiveness words from Study 1 feature in the top nine negative words from the pilot data, with the other one (lenience) sitting in "neutral" territory. The six positively-valenced forgiveness words from Study 1 all featured in the top eleven positive words from the pilot data.

As can be seen from the results, the words provided were evaluated as being more positive than they were negative, with 18 of the 33 words (55%) receiving positive differential scores, compared with only 10 of 33 (30%) receiving negative differentials. This abundance of positively-evaluated forgiveness words presents an opportunity to replicate the forgiveness-positive IAT variant from Study 1 using an entirely new set of stimulus words. After excluding the words used in Study 1 (absolve, compassion, mercy, empathy, reconcile, amnesty), 12 words remained on the positively-valenced list. Three of these words – generosity, benevolence, and altruism – were also seen by a substantial number of participants (24-36%) as being *not* related to forgiveness. Despite receiving significant positive responses, "letting go" and "moving on" can also be excluded as potential candidates for the IAT, owing to the fact that they are more aptly described as *phrases* – they are sets of two words, which may result in increased processing time, and therefore inflate IAT reactions time scores. After these exclusions, a final list of seven words remains, of which "lenience" only scored a positive differential of +1, and so could be excluded from the final list. This left six words to replace those used for the

Forgiveness-positive IAT variant used in Study 1: understanding, acceptance, tolerance, goodwill, pardon, and reprieve.

Unfortunately, owing to the relatively low frequency of negatively-evaluated words, there are too few to generate an entirely new condition for the forgiveness-negative IAT variant. After excluding the six words used for this condition in Study 1 (excuse, overlook, condone, justify, lenient, ignore), there are only 5 remaining words: forget, allow, appease, permit, and indebtedness. However, these words all received substantial responses (18-36%) indicating that they are *not* related to forgiveness at all, casting doubt over their usefulness as potential IAT candidate words. Guidelines for IAT word selection presented by both Lane et al. (2007) and Nosek et al. (2007) suggest that stimulus items must be unambiguously related to the target category, ruling out words that have significant disagreement regarding centrality and or/valence to forgiveness.

The IAT variants used in Study 2b will need to be a hybrid of both the old and the new lists. The negatively valenced words from Study 1 will need to be retained, while the positively valenced words can be replaced with a fresh list.

2.9 Study 2b

2.9.1 Method

2.9.1.1 Participants

Participants were 114 (87 female, 27 male) first year undergraduate psychology students at the University of Adelaide, who participated in exchange for course credit. Mean age of participants was 19.0 years (*SD*=3.72).

2.9.1.2 Materials

2.9.1.2.1 IAT Design

The IAT was identical to that used in Study 1 except for the stimulus words used in the "forgiveness-positive" category, and half of the words (the positive set) used in the "forgiveness-balanced" category. The "forgiveness-balanced" condition was comprised of three of the new positively-valenced forgiveness words, and three of the original negatively-valenced forgiveness words. The stimulus word sets for all other categories were the same as those used in Study 1. The stimulus word sets for forgiveness for the three IAT versions are presented in Table 2.8.

Table 2.8

Forgiveness-Revenge IAT Stimulus Word Sets for Forgiveness-positive, Forgiveness-negative and Forgiveness-balanced Categories

Forgiveness-positive	Forgiveness-negative	Forgiveness-balanced
Understanding	Excuse	Understanding
Acceptance	Overlook	Acceptance
Tolerance	Condone	Tolerance
Good-will	Justify	Excuse
Pardon	Lenient	Overlook
Reprieve	Ignore	Condone

All other aspects of the IAT procedure, including sequence and number of trials per block, were identical to those used in the previous study.

2.9.1.2.2 Self-report questionnaire

The questionnaire used for this study was identical to that administered in Study 1 but with the inclusion of an additional measure of socially desirable responding (Ballard, 1992), and a dispositional measure of vengeance-seeking (Stuckless & Goranson, 1992).

Vengeance scale

The Vengeance scale (Stuckless & Goranson, 1992) consists of 20 items which measures a person's endorsement of vengeful beliefs/attitudes (e.g. "revenge is morally wrong") or behaviours (e.g. "honour requires that you get back at someone who hurt you"), as well as how strongly they identify as a vengeful person (e.g. "if I am wronged, I can't live with myself unless I get revenge"). Each item was scored on a 7 point Likert type scale, resulting in total scale scores which had a possible range between 20 and 140. Internal consistency for the scale was very good (α = .92).

The Vengeance scale was included to assess implicit-explicit correspondence from an alternative perspective. The IATs being used in these studies include *revenge* as the paired category of *forgiveness*, which means that IAT scores should reflect a person's preferences for both forgiveness *and* revenge. Study 1 only assessed convergent validity by using self-reported measures of forgiveness, when a self-reported measured of revenge should be equally applicable. It is anticipated the Vengeance scale should correlate negatively with Forgiveness-Revenge IAT scores.

Marlowe-Crowne Short Form (MCSF)

This scale, devised by Ballard (1992), is one of many available short forms of the original Marlowe Crowne Social Desirability Scale (Crowne & Marlowe, 1960). The scale consists of a subset of 13 items from the original 33 item scale, originally derived through factor analysis and comparison with several other short forms of the same scale. An example item is "I take out my bad moods on others now and then". Participants rate each item as either true or false of them. After reverse scoring as appropriate, items were summed to create a scale score ranging from 0 to 13. Internal consistency for this scale was poor (α = .57), and could not be improved by excluding items.

2.9.1.2.3 *Internal consistency reliabilities*

For most of the remaining scales, internal consistency reliabilities were borderline to acceptable (TTF, α=.71; HFS, α=.67; SDS-17, α=.68). Reliability for the ATF was poor (α=.57) but was improved to a borderline level (α=.62) with the exclusion of one item: "People should work harder than they do to let go of the wrongs they have suffered". The IAT demonstrated good reliability, α=.87.

2.9.1.3 Procedure

The procedure for this study was identical to Study 1.

2.9.2 Results

2.9.2.1 Data Preparation

IAT D scores were computed following procedures previously outlined in section 2.4.7.1, which resulted in six participants being excluded from subsequent analyses. Mean D scores for each IAT version are presented in Table 2.9.

Table 2.9

Means, 95% Confidence Intervals and Standard Deviations for IAT D scores

	N	*M* (95% CIs)	*SD*
IAT (positive)	44	.98 (.81-1.15)	.56
IAT (negative)	35	.75 (.60-.90)	.45
IAT (balanced)	29	.87 (.69-1.06)	.49
IAT (combined)	108	.88 (.78-.97)	.51

2.9.2.2 The effects of valence on IAT scores

Consistent with the hypotheses and findings from Study 1, a One-way ANOVA[6] and examination of confidence intervals[7] revealed no significant differences between the

[6] Again, the distribution of D scores for the forgiveness (positive) condition was negatively skewed. A negative square root transformation failed to improve the distribution to within acceptable bounds. An ANOVA was re-run using the transformed variables, but this also yielded no significant differences between the three IAT versions.

[7] 98.3% confidence intervals for the three pairwise comparisons were as follows: IAT(positiveXbalanced) CI range -.21 to .41; IAT(balancedXnegative) CI range = -.41 to .16; IAT(positiveXnegative) CI range = -.06 to

three variations of the IAT, $F(2,105)=1.94$, $p=.15$, $\omega=.19$. Valence of stimuli used to represent forgiveness in the IAT appeared to play a negligible role in determining scores on a forgiveness-revenge IAT. As there were no significant differences between IAT versions the three conditions were collapsed for the remainder of the analyses.

2.9.2.3 Relationships between IAT D scores and self-reported forgiveness and revenge

The IAT showed only a low level of convergence with self-reported forgiveness and revenge attitudes and dispositions. As shown in Table 2.10, there was a significant (albeit small) positive correlation between IAT D scores and the HFS, and a significant (albeit small) negative relationship between the IAT and vengeance. Neither the ATF nor TTF was significantly correlated with IAT scores.

2.9.2.4 IAT's sensitivity to socially desirable responding

In accordance with hypothesis and consistent with findings from Study 1, there was no correlation between the IAT and either measure of SDR. In contrast, of the three self-reported forgiveness measures, two of these significantly correlated with both measures of SDR. Surprisingly, the vengeance scale also correlated with both SDR scales, but in the opposite direction to expected.

.51. As the confidence intervals for all three comparisons include zero, we cannot be confident that there are any meaningful differences between the three conditions

Table 2.10

Intercorrelations Between IAT D Scores, Self-Report Measures of Forgiveness Attitudes, and Socially Desirable Responding Scales

	1	2	3	4	5	6	7
1. IAT	--						
2. ATF	.07	--					
3. TTF	.17	.28**	--				
4. HFS	.19*	.41**	.56**	--			
5. Vengeance	-.19*	-.57**	-.47**	-.64**	--		
6. SDS-17	-.05	.13	.28**	.21*	-.22*	--	
7. MCFS	.05	.09	.33**	.27**	-.34**	.64**	--

* $p<.05$. ** $p<.01$.

2.9.3 Discussion

The primary aim of this study was to determine if the results of Study 1 would replicate with an IAT that used an alternative set of stimulus words. The results were at least partially replicated.

Similar to Study 1, valence of IAT stimuli words had a negligible impact on IAT scores. This adds further support that the IAT is doing what it claims to do, rather than assessing a methodological artefact.

Findings were also consistent with those of Study 1 in respect to convergent validity. There were correlations between the IAT and *some* of the self-reported attitude measures, but these correlations were still only small. Consistent with prediction, the vengeance scale *did* correlate with the IAT, providing some preliminary evidence that it

may be important to consider the 'opposite' category when constructing and interpreting an IAT for forgiveness.

The IAT was again unrelated to SDR, despite an additional measure being used. However, this time both SDR measures were significantly related to three out of the four explicit forgiveness scales. On face value this seems encouraging, as a key rationale for developing a forgiveness IAT is that explicit measures of forgiveness are limited by their susceptibility to self-presentation biases, and this data provides support for this argument. However, closer inspection of the direction of relationships provides grounds for concern. If SDR has been an issue on any scales, then these SDR measures should correlate *positively* with the scales that have been affected - which is what occurred with the TTF and HFS. However, the correlations between the vengeance scale and the two SDR scales are significantly *negative* – which suggests that these scales may be assessing something beyond just SDR.

Recall that SDR scales include two kinds of items: behaviours that are socially desirable yet infrequent (e.g. "No matter who I'm talking to, I'm always a good listener"), and behaviours that are socially undesirable yet frequent (e.g. "There has been an occasion when I took advantage of someone else"). Given the nature of these items, it is possible that these types of scales also measure *prosocial* behaviour – or at least self-perceived prosocial behaviour. The positive correlations with self-reported forgiveness and negative correlation with vengeance may be more indicative of the relationship between forgiveness and prosocial behavior. That is, perhaps those high in forgiveness actually just *do* engage in more prosocial behaviours and less antisocial behaviours. This is plausible, as it has already been established in that there is an empirical link between

forgiveness and prosocial behaviour (e.g. Karremans et al., 2005; Zechmeister et al., 2004). Thus this study may not actually provide evidence that the Forgiveness-Revenge IAT is resistant to SDR.

2.10 General Discussion

The studies presented in this chapter provide some preliminary evidence that measuring forgiveness using an IAT may be a promising angle for forgiveness research. Across two studies, the validity of a Forgiveness-Revenge IAT was examined in three key ways, with encouraging findings in at least two of these.

First, IAT scores were not significantly influenced by the valence of stimulus words used to represent the forgiveness category. This is an important finding in light of some general criticisms of the IAT – namely, that IAT scores might reflect strategic recoding of category/stimuli valence, rather than actual implicit associations (Govan & Williams, 2004; Mitchell et al., 2003; Rothermund & Wentura, 2004; Rothermund et al., 2005). These effects had no significant bearing on Forgiveness-Revenge IAT scores, meaning that the Forgiveness-Revenge IAT passes an important test of validity. This was irrespective of the specific words chosen to represent the positive forgiveness category, as the two studies used an entirely different set of words for this category, which provides further evidence for the construct validity of the Forgiveness-Revenge IAT.

Second, across the two studies it was encouraging that the IAT was unrelated to measures of SDR, while in Study 2b the majority of the explicit forgiveness measures were. These findings inspire confidence that a Forgiveness-Revenge IAT may have one advantage over self-report forgiveness scales. However, these results must be viewed

with some caution. The nature of null hypothesis significance testing means that just because no significant relationship was found between the IAT and SDR, this does not mean that such a relationship does not exist. To the extent that the SDR scales may have also been assessing more general prosocial motives, the effects of socially desirable responding on the forgiveness IAT remain unclear.

Third, these studies demonstrate that the Forgiveness-Revenge IAT possesses some degree of convergent validity, although these findings are not entirely clear-cut. On the one hand, the fact that some of the explicit forgiveness measures correlated with the IAT should be encouraging, particularly considering that IE correspondence reported in the literature is usually quite low. In Study 1, two of the three explicit forgiveness scales correlated significantly with the IAT with an average of r = .19. In Study 2, two of the four forgiveness/revenge scales significantly correlated with the IAT, also at an average of r = .19. Considering that the average degree of IE correspondence found in the IAT literature is in the vicinity of r = .23 (Greenwald et al., 2009) to r = .24 (Hofmann et al., 2005a), the Forgiveness-Revenge IAT appears to have fared reasonably well in this regard.

One factor which may have affected IE correspondence is construct clarity. In both of these studies, three variants of the IAT were used, and then scores for these three versions were combined after both ANOVA and examination of confidence intervals showed no significant differences in scores between the three. However, the three variants may have measured implicit preferences in slightly different ways, which may have impacted the manner in which they would co-vary with the self-report attitude scales – nuances which may have been lost once the scores were aggregated.

Unfortunately, low numbers of participants in each cell means that there is insufficient statistical power to conduct separate correlation analysis using each IAT variant.

Perhaps the most significant reason for low convergence, and one that has already been highlighted, is the relative structure of the IAT. The IAT, by its very nature, requires that forgiveness is evaluated against a paired 'opposite' category, which for these studies was *revenge*. Consequently, the IAT was measuring participants' preference for forgiveness relative to revenge, as opposed to the self-report scales which had no such anchor point. Convergence between a vengeance scale and the IAT (Study 2) highlights this point. At least some of the variance in the Forgiveness-Revenge IAT may be explained by attitudes toward revenge, rather than forgiveness. Thus it is clear that, theoretically, the two types of measures are assessing slightly different constructs, which should in turn result in lower levels of convergence between the two.

The argument presented above suggests that the way in which the IAT can assess people's forgiveness attitudes is largely dependent on what they are comparing forgiveness to. In these two studies, forgiveness was contrasted with revenge, as this seemed both a logical and theoretically appropriate choice. However, the Forgiveness IAT may behave quite differently depending on the category that is paired with forgiveness, and this in turn may influence the way in which IAT scores covary with other, more explicit measures of forgiveness. This is an important direction for research on the forgiveness IAT, and one which will be addressed in Chapter 3.

Chapter 3:

Selecting appropriate categories for the Forgiveness IAT

3.1 Chapter Overview

This chapter aimed to explore the ways in which the non-forgiveness IAT categories might impact the way that forgiveness is implicitly measured. This represents a shift in focus from the previous chapter from the *stimulus* level to the *category* level of the IAT. Chapter 2 addressed how the specific stimuli used to represent the IAT 'forgiveness' category might affect scores on the IAT – this was an important step in validating an IAT suitable for the measurement of forgiveness. Up until now, the focus has been on only one of the four IAT categories ('forgiveness') as the present work is concerned with forgiveness as its central theme. However, the IAT involves a double dissociation task. As such, one must recognise that IAT scores are a product of not one, but four, categories, organised into two pairs (Lane et al., 2007; Nosek et al., 2007; Schnabel et al., 2008a). Consequently, information about forgiveness that can be captured by an IAT is constrained by the other three categories – which up to this point have been 'revenge', 'pleasant' and 'unpleasant'. This chapter presents data from a study which aimed to investigate whether 'revenge' is the most suitable category to contrast forgiveness with, or if there may be a more useful alternative.

3.2 IAT category selection

An IAT effect is the product of how a person responds to two sets of contrasting category pairings. These pairings are often arranged as *target* and *attribute* categories (Greenwald et al., 1998). The target categories often (although not always) comprise a pair of constructs that could be semantically seen as 'opposites', such as "black-white", "gay-straight", "fat-thin", or "old-young". The second pairing may be another set of

"target" categories (such as "male-female" being used as a second target pairing on an "arts-science" IAT [e.g. Nosek et al., 2002a]) but it is more often a set of evaluative 'attribute' categories, usually indicating *preference* (e.g. pleasant-unpleasant, good-bad) but sometimes reflecting other kinds of evaluations like *self-concept* (e.g. self-other, me-not me). The second category in each pairing is often referred to as the *contrast* (Nosek et al., 2007) or *comparison* (Lane et al., 2007) category.

3.2.1 Choosing contrast/comparison categories

Although Greenwald et al. (1998) did not originally prescribe any specific guidelines or criteria for selecting appropriate IAT categories, there are a handful of subsequent papers which have since attempted to do so. For example, Nosek et al. (2005) stress the importance of selecting categories which are unambiguous and cannot be confounded with any of the other categories. In regards to selecting the contrast/comparison category, Lane et al. (2007) present several options, the first of which is to use a target pairing that is naturally dichotomous (such as male-female), in which case the contrast category is obvious. This recommendation is often echoed in other reviews of the IAT, with the suggestion that the IAT is not suitable for measuring non-dichotomous target pairings (e.g. Nosek et al., 2005; Schnabel et al., 2008a). However, this suggestion – if followed strictly – would render the IAT almost useless, as for the majority of constructs explored in IAT research purely dichotomous categories are not readily available. In this case, Lane et al. (2007) recommend that the comparison category should be "a sensible, mutually exclusive category that is ideally from the same domain" (p.86).

The most obvious example of this approach to category selection is present in one of the most applied contexts of IAT research – prejudice. For example, IATs designed to measure racial stereotypes/preferences often use the categories white-black, generally referring to a contrast between Caucasian Americans and African Americans. However, this is clearly not a pure dichotomy; a Caucasian-Hispanic IAT also measures implicit race preference as does a Caucasian-Asian IAT. In this case, the category "black" is not simply an *opposite* of "white", but one of many possible "mutually exclusive" comparison categories coming from the "same domain" (i.e. race).

In the same way, the most classic example of the IAT – flowers versus insects – does not consist of dichotomous constructs, and in this case they are not even from the same domain. In fact, target category pairings consisting of 'opposites' are actually quite rare in IAT research, with the comparison category often being just one of several possible 'opposites' of its corresponding paired target. Despite this, attempting to select a contrast category that appropriately complements its paired counterpart is an important step in IAT design, and has significant implications for the IAT's validity. Greenwald et al. (2009) found that IATs which had higher "complementarity" also had higher convergence with explicit measures of corresponding constructs, highlighting the need for careful consideration in choosing a comparison category that is complementary to its paired target.

3.2.2 Choosing an appropriate category to complement forgiveness

3.2.2.1 Revenge

In Studies 1 and 2, the target categories were 'forgiveness' and 'revenge', which appear on face value to be complementary constructs. However, revenge may not be the only – or the best – contrast category for forgiveness. Certainly forgiveness and revenge have often been found to be inversely related. For example, significant negative correlations between the two constructs were found in Study 2, as well as in the broader literature (Brown, 2003, 2004; Johnson, Kim, Giovannelli, & Cagle, 2010; Thompson et al., 2005). However, these correlations are usually small to moderate[8], suggesting that the two constructs cannot be considered as strict opposites. Others have also argued that forgiveness and revenge are not necessarily mutually exclusive constructs. For instance, North (1998) suggests that forgiveness does not necessarily mean that a victim must forego the desire for revenge, nor the right to punishment. Rather, revenge and/or punishment may be important catalysts for the forgiveness process to take place. In this way, forgiveness and revenge may not necessarily always be opposing forces, but may be part of the same process.

An alternative perspective on the relationship between forgiveness and revenge is that forgiveness may be a logical opposite of revenge, even if the reverse is not entirely true. If one searches for "revenge" in any number of thesauruses, "forgiveness" most often heads the list of suitable antonyms, indicating that forgiveness is indeed probably a

[8] One exception is McCullough et al. (1998), who found correlations between a single item forgiveness measure and state-specific revenge motivations to be as high as $r = -.67$

suitable opposite of revenge. In contrast, the same does not work in reverse – looking up "forgiveness" will sometime turn up "revenge" as an antonym, but this is far less consistent, and many other alternatives are presented. This is perhaps suggestive of the idea that forgiveness actually has several 'opposites', rather than just the one. The forgiveness literature in psychology would appear to support this claim, with several alternatives mentioned from a theoretical standpoint. Brown (2004) argued that one cannot be simultaneously high in forgiveness and high in revenge, but that being low in forgiveness does not necessarily imply that a person is vengeful. He also found that the extent to which unforgiving participants were vengeful or not was moderated by individual differences in narcissism. Thus, whether or not revenge can be seen as an appropriate opposite of forgiveness may be influenced to at least some extent by personality and individual difference factors. For some people in some situations revenge may be an appropriate opposite of forgiveness, whereas for others there may be more relevant alternatives.

3.2.2.2 Justice

One alternative category, which is somewhat related to revenge, is *justice*. Whether justice can be considered a suitable complement to forgiveness largely depends on how it is defined or understood. Both theoretical and lay understandings of justice have tended to focus on punitive qualities and the idea of 'just deserts', referred to more broadly as *retributive* justice (for a review see Darley, 2002). The inexorable links between retribution and justice are summarised by Vidmar (2000, p. 31): "retribution and revenge….are arguably the oldest, most basic, and most pervasive justice reactions

associated with human social life". This understanding of justice as retributive can also be readily observed in the discourse that is frequently associated with justice in everyday language – sayings such as "do the crime, do the time", and "justice must be done" being almost universally understood as "this person must be *punished* for what they did".

This emphasis on justice as *retributive* has, for the most part, been no different in the forgiveness literature (Exline, Worthington, Hill, & McCullough, 2003)(although there have been some recent attempts to place forgiveness within alternative *restorative* justice frameworks [e.g. Karremans & Van Lange, 2005; Strelan, Feather, & McKee, 2008; Wenzel & Okimoto, 2010])[9]. As a consequence, forgiveness and justice have often been framed as opposing and incompatible. It has been suggested that forgiving means (at least) a "loosening of justice standards" (Exline & Baumeister, 2000, p. 147) or perhaps even the sacrificing or foregoing of justice altogether (Reed & Aquino, 2003). Exline et al. (2003) argue that forgiveness and retributive justice can be seen as complementary constructs, with the former being an alternative to the latter. If defined retributively, justice appears to be an appropriate candidate as a contrast for forgiveness in the IAT. Fortunately, operationalising justice in this way is relatively easy to do in an IAT, by

[9] Of course, justice may be less suitable as a contrast/comparison category for forgiveness if it is conceptualized in more prosocial ways. For example, Strelan (2007) and Strelan and Sutton (2010) have found just world beliefs to be positively related to both state and trait forgiveness; Karremans and van Lange (2005) found that priming participants with social justice promoted forgiveness; and (inversely) Wenzel and Okimoto (2010) found that priming participants to forgive increased subsequent perceptions of justice. The present work will only focus on retributive notions of justice.

selecting stimulus words that represent ideas of punishment or just deserts, and excluding words that represent the more prosocial elements of the justice construct.

3.2.2.3 Unforgiveness and holding a grudge

Another logical alternative as a complement to forgiveness is "unforgiveness". This concept - championed by Worthington (1998, 2001) and referred to by several other scholars (e.g. Exline et al., 2004; Harris & Thoreson, 2005; Witvliet et al., 2008) – initially appears to be a sensible complement for forgiveness, as at face value the former is merely a negated form of the latter. However, Worthington concedes that unforgiveness should not be considered a "polar opposite" of forgiveness (Witvliet et al., 2008, p. 11). According to Worthington, forgiveness is just one of many alternatives to unforgiveness, with others including acceptance, cognitive reframing, seeking legal justice, seeking revenge, forgetting, excusing or condoning the offense, all of which decrease the 'injustice gap' by either minimising the importance of the offence, or seeking to narrow this gap (Wade Brown & Worthington, 2003; Worthington, 2001). Thus, theoretically forgiveness and unforgiveness cannot be seen as two sides of a coin. There is some empirical data to support this. Wade Brown and Worthington (2003) found that those high in forgiveness were generally low in unforgiveness, but those low in forgiveness were not necessarily high in unforgiveness.

Further doubt about the suitability of forgiveness and unforgiveness as complementary constructs comes from examination of what precisely is meant by the latter of these terms. Unforgiveness is defined as the 'cold' emotions that follow a transgression, such as "resentment, bitterness, hatred, hostility, [residual] anger, and

[residual] fear" (Worthington, 2001, p. 172), as distinct from the 'hot' emotions that form people's initial reactions to a transgression (Worthington, 2000). Thus unforgiveness corresponds to only one component of forgiveness – emotions – yet there is consensus that forgiveness also involves cognitions (Enright et al., 1998), motivations (McCullough et al., 1998) and perhaps behaviours (Zechmeister et al., 2004).

Finally, and perhaps most importantly, unforgiveness may not be ideal as an IAT category because of its relative ambiguity. To operate effectively, the IAT requires categories that are "precisely defined", as it is the "construal of the category that determines how it is evaluated" (Lane et al., 2007, p. 85). Categories that do not fit these criteria introduce additional room for error in IAT performance, as items in this category require additional processing time, which may subsequently impact IAT scores. For this reason, unforgiveness may not be appropriate for use in the IAT. Although Worthington's definition of unforgiveness is relatively precise, it is questionable as to whether this word has the same meaning amongst a general population. The word unforgiveness has the potential to be broad and highly ambiguous, and may mean different things to different people, encompassing ideas of revenge, justice, grudge-holding, as well as combinations of emotions, cognitions, motivations and behaviours. This ambiguity and subjectivity creates problems for selecting a subset of stimuli that would universally represent the construct within an IAT.

Perhaps this problem of ambiguity can be overcome by using a word for which there should be less definitional disagreement between academics and lay people: grudge. A grudge can be defined in very similar terms to unforgiveness, but has the added benefit of being more universally understood: the average person does not

typically use the word unforgiveness, yet they do use words like revenge, justice and grudge. Forgiveness is often defined by both theorists (Yandell, 1998) and lay people (Kearns & Fincham, 2004; Younger et al., 2004) as a form of "letting go". The question then becomes, what are people letting go of? Worthington would argue that they are letting go of unforgiveness: sustained emotions such as bitterness, anger, hostility, and resentment (Worthington, 2001). Other authors (e.g. Baumeister et al., 1998) would call these same things a *grudge*. Baumeister et al. (1998, p.80) go so far as to call holding a grudge the "opposite" or "mirror image" of forgiveness. In fact, many authors frame forgiveness in opposition to holding a grudge as part of their forgiveness definitions (McCullough, 2008; Struthers, Eaton, Shirvani, Georghiou, & Edell, 2008). McCullough (2008, p116) is quite explicit in this, defining forgiveness as "getting over your grudge and starting to feel positively again toward someone who harmed you".

Additionally, like revenge, grudge is used in measuring forgiveness. The Heartland Forgiveness scale (self-forgiveness subscale) includes the item "I hold grudges against myself for negative things I've done" (Thompson et al., 2005), while the TTF (Brown, 2003) includes the item "I have a tendency to harbor grudges", both of which are negatively scored as a proxy for greater forgiveness. Thus grudge appears to be a suitable contrast category for forgiveness, while also circumventing some of the potential pitfalls of using "unforgiveness".

3.2.3 Comparing contrast categories for the Forgiveness IAT

As already mentioned, it is possible that forgiveness does not have one single opposite, but several. Therefore, there may be several possible constructs suitable for

use as a comparison category for forgiveness in the IAT, with revenge, retributive justice, and grudge all appearing to be sensible choices. Note that the IAT-imposed constraint of needing to find a suitable complement for forgiveness is not necessarily a limitation. The forgiveness literature struggles to agree on a definition of forgiveness; perhaps an IAT-based comparison of these constructs will provide more information about how people implicitly understand/define forgiveness.

3.3. Study 3

3.3.1 Study Overview

The present study aimed to explore the significance of the contrast category used to complement forgiveness in the IAT. This was done by comparing three IAT versions, each one identical except for the contrast category used: revenge, justice or grudge. All three versions retained the pleasant-unpleasant attribute dimension. The suitability of the contrast category was assessed in two ways: a comparison of mean scores across conditions, and the extent to which each IAT versions displayed convergent validity with self-report measures of forgiveness.

3.3.2 Research questions

The present study was of an exploratory nature, and as such no specific hypotheses were set.

Research question 1: Differences between IAT variants

The three IAT variants were compared using a Oneway ANOVA, with IAT variant as the independent variable and IAT D score as the dependent variable. Of particular interest was whether or not the choice of contrast category would produce IAT effects that were significantly different from each other and, if so, which variants would produce the larger effects. Specifically, two possible outcomes were being investigated:

(a) If the IAT contrast category plays a significant role in producing IAT effects, it would be expected that D scores would be significantly different between the three IAT conditions.

(b) If the IAT contrast category does not play a significant role in producing IAT effects, it would be expected that D scores would not be significantly different between the three IAT conditions.

Research question 2: Convergent validity

It was anticipated that using different contrast categories in the IAT might impact the degree of convergence between the IAT variant and self-report forgiveness measures. Specifically, this study aimed to identify which contrast category would produce IAT scores with the highest levels of convergent validity. That is, the IAT variant that showed the greatest convergence with self-reported forgiveness scales should be the most effective for measuring implicit forgiveness.

3.3.3 Method

3.3.3.1 Design

This study utilised a single factor between-groups experimental design.

3.3.3.2 Participants

Participants were 215 (145 female, 70 male; age *M*=20.81, *SD*=5.76) first year undergraduate psychology students at the University of Adelaide, who participated in exchange for course credit.

3.3.3.3 Materials

3.3.3.3.1 IAT Design and Structure

There were three IAT conditions, differing only in the category that was paired with forgiveness. Forgiveness stimulus words were selected based on those deemed most central to the construct from the pilot data reported in Study 2a, and as such were a combination of those used in Studies 1 and 2b. Forgiveness words remained constant across the three IAT conditions. Stimulus words for grudge and justice were selected through consultation of several thesauruses, with an initial list being rated and ranked for centrality by a group of forgiveness researchers at the University of Adelaide. Due to limited availability of synonyms for the grudge and (retributive) justice categories, stimulus word set sizes were reduced to five per category (compared to six per category used in studies one and two).

It should be noted that the aim of the "justice" category was specifically to examine attitudes regarding *punitive/retributive* justice, rather than broad ideas of justice found in dictionary and thesaurus definitions. As such, word lists for this category were also drawn from thesaurus searches of "punishment". Words representing prosocial understandings of justice were excluded. Thus, the category itself should be thought of specifically as "retributive justice", rather than simply "justice" in general.

Revenge stimulus words were identical to those used in Studies 1 and 2, with the omission of one word (retribution), in order to keep word set sizes consistent between categories. "Retribution" was selected as the word to omit due to its conceptual overlap with the justice category. Stimulus words for each of these categories are shown in Table 3.1.

Table 3.1

IAT Stimulus Word Sets for Target Categories Forgiveness, Revenge, Grudge and Justice.

Forgiveness	Revenge	Grudge	Justice
Acceptance	Retaliate	Resentment	Punish
Mercy	Vengeance	Bitterness	Discipline
Reconcile	Pay-back	Grievance	Reprimand
Understanding	Vindictive	Animosity	Penalty
Compassion	Reprisal	Malice	Compensation

Attribute categories of pleasant and unpleasant were retained, but each of these word sets were reduced from six to five words for consistency with the other categories,

resulting in the omission of “freedom” and “abuse” from the respective lists. It should be noted that this reduction in category set sizes should have no bearing on IAT results, as IAT scores generally remain unaffected by number of words per category (Lane et al., 2007), with Nosek et al. (2005) showing specifically that reliable IAT effects can be produced with as few as 2 stimulus items per category.

The reduction of category set sizes from six to five words also meant that IAT blocks were reduced to 20 and 40, rather than 24 and 48.

All other aspects of the IAT design, instructions to participants, and administration, were identical to those described for studies 1 and 2b.

3.3.3.3.2 *Self-report questionnaire*

The questionnaire was identical to that used in Study 1, with the addition of a scenario-based Willingness To Forgive (WTF: DeShea, 2003) scale.

Willingness To Forgive scale

The WTF scale (DeShea, 2003) comprises 12 items, designed to assess how willing participants are to forgive when confronted with a range of conflict situations. Each item presents a brief hypothetical context in which forgiveness might be appropriate, and then asks participants to rate their willingness to forgive in this particular scenario on a 7 point Likert type scale, ranging from “not at all willing” to “completely willing”. The original scale range was numbered from 0-6 but in the present study this was changed to 1-7, in order to be consistent with the scales used for other included measures. An example of a scenario presented is: “You come from work and catch your roommate looking at your

private journal. Your roommate claims to have been looking for a dictionary and really hadn't read much of your journal". Items vary in both the severity of the offense, and the relationship to the offender. The scale is intended to be used as a measure of dispositional forgiveness, whilst recognising that situation/context may play an important role.

3.3.3.3.3 *Internal consistency reliabilities*

Cronbach's alpha for the scales ranged from borderline for the ATF (α=.61), TTF (α=.67), SDS-17 (α=.68) and MCSF (α=.63), to good for the WTF (α=.87) and acceptable for the HFS (α=.74). Borderline reliabilities were not substantially improved by deleting scale items.

3.3.3.4 Procedure

The procedure for this study was identical to Study 2b, with the exception of the previously described variations in the materials used.

3.3.4 Results

3.3.4.1 Data preparation

IAT D scores were computed, with twenty participants excluded due to extreme fast responses (<300ms, remaining N=195). Mean D scores for each IAT version are presented in Table 3.2.

Table 3.2

Means, 95% Confidence Intervals and Standard Deviations for IAT D scores

	N	*M* (95% CIs)	*SD*
IAT (forgiveness-revenge)	63	.91 (.77-1.06)	.57
IAT (forgiveness-grudge)	63	.97 (.81-1.12)	.60
IAT (forgiveness-justice)	69	.92 (.82-1.01)	.40
IAT (combined)	195	.93 (.86-1.01)	.53

3.3.4.2 The effect of the IAT contrast category on magnitude of IAT D scores

A Oneway ANOVA[10] revealed that there were no significant differences between the three variations of the IAT, $F(2,192)=.21$, $p=.81$, $\omega=.05$. An examination of 98.3% confidence intervals for the three pairwise comparisons further supports this, with all three comparisons including zero[11]. In summary, the contrast category used to complement forgiveness in the IAT appeared to play a negligible role in determining the magnitude of IAT scores.

[10] Distributions of scores for both the forgiveness-revenge IAT and forgiveness-grudge IAT were negatively skewed. A negative square root transformation was able to correct this skew for the forgiveness-grudge IAT only. However, the outcome of the ANOVA was still non-significant, and so original (non-transformed) scores were retained.

[11] 98.3% CIs for the three comparisons: IAT(revengeXgrudge) CI range = -.31 to .20; IAT(revengeXjustice) CI range = -.21 to .21; IAT(grudgeXjustice) = -.17 to .27.

3.3.4.3 The impact of contrast category on IAT convergence with self-report scales

There were only two significant correlations between IAT versions and self-report forgiveness scales (see Table 3.3). The Forgiveness-Revenge IAT correlated with the HFS and the Forgiveness-Grudge IAT correlated with ATF. The Forgiveness-Justice IAT did not significantly correlate with any of the self-report forgiveness measures. Intercorrelations among the explicit forgiveness measures were all significant, ranging from small to moderate.

Table 3.3

Intercorrelations Between IAT D Scores, Self-Report Measures of Forgiveness Attitudes, and Socially Desirable Responding scales

	1	2	3	4	5	6	7	8	9
1. IAT									
2. IAT:F-R									
3. IAT:F-G									
4. IAT:F-J									
5. ATF	.18*	.14	.30*	.12					
6. TTF	-.03	.04	-.03	-.12	.40**				
7. HFS	.17*	.26*	.13	.09	.52**	.58**			
8. WTF	-.08	-.05	.00	-.22	.34**	.48**	.46**		
9. SDS-17	-.17*	-.21	-.15	-.17	.06	.18*	.18	.16*	
10. MCSF	-.24**	-.24	-.28*	-.20	.10	.27*	.23**	.27**	.67**

* $p<.05$. ** $p<.01$.

IAT:F-R = Forgiveness-Revenge IAT; IAT:F-G = Forgiveness-Grudge IAT; IAT:F-J = Forgiveness-Justice IAT; ATF = Attitudes To Forgiveness (Brown, 2003); TTF = Tendency to Forgive (Brown, 2003); HFS = Heartland Forgiveness Scale ("other" subscale; Thompson et al., 2005); WTF = Willingness To Forgive (DeShea, 2003); SDS-17 = Social Desirability Scale (Stober, 2001); MCSF = Marlowe-Crowne Short Form (Ballard, 1992)

3.4 Discussion

The present study aimed to explore whether or not the choice of IAT comparison category for forgiveness (revenge, grudge, or justice) would impact (a) the magnitude of scores on the forgiveness IAT, and (b) the degree to which the forgiveness IAT converged with self-reported forgiveness attitudes or tendencies. Although the magnitude of IAT scores did not differ across IAT conditions, the degree of correspondence between implicitly and explicitly measured forgiveness did.

Negligible differences between the three IAT variants initially suggests that the choice of contrast category for forgiveness in the IAT may not actually be important. Mean IAT effects were remarkably similar across the three IAT variants. However, this conclusion becomes less certain when considering the relations between the IAT variants and self-report forgiveness measures. The Forgiveness-Revenge IAT only significantly correlated with the HFS, the Forgiveness-Grudge IAT only significantly correlated with the ATF, while the Forgiveness-Justice IAT did not show significant correlations with any of the self-reported forgiveness measures.

These disparate relations suggest that each IAT variant, just like each self-report forgiveness measure, is capturing a slightly different aspect of a person's forgiveness attitudes. On the surface, this may make sense. For example, Brown (2003) argues that many measures confound forgiveness with revenge, and therefore he deliberately kept ideas of revenge out of the TTF and ATF. This may explain why the Forgiveness-Revenge IAT variant does not correlate with either of these measures, but correlates with the HFS, which *does* include items about how much an individual wishes to inflict harm on the person who hurt them. However, this does not explain the patterns of convergence seen

for the Forgiveness-Grudge IAT. Of all the included self-report measures, the TTF is the only one to contain an explicit item about holding a grudge, while its other three items also pertain to ideas of grudge-holding. If the Forgiveness-Grudge IAT was to correlate with any of the self-report forgiveness measures it should have been expected to have been the TTF, rather than the ATF.

Attempting to explain why these two particular correlations did occur is fraught with difficulty. One possibility is that they occurred by chance: computing large numbers of correlations increases the chance of making a Type II error. Perhaps a more useful question regarding the IAT's convergence in this study is why most of the correlations between implicitly and explicitly measured forgiveness were *not* significant. The broader IAT literature suggests that IAT and self-report measures should be reasonably expected to correlate at an average of r=.23 (Greenwald et al., 2009). In the present study, generally the IAT and self-report measures did not correlate at all.

What is less clear from the current literature is the mechanism by which low explicit-implicit correlations arise. One explanation is that participants may be unwilling to report their true attitudes, possessing both the ability and motivation to control their responses on the self-report measures (Fazio & Olsen, 2003). If this had been the case in the present study, self-reported forgiveness measures should have been associated with the two social desirability scales, whereas the IAT should not have been, which is indeed what occurred. With the exception of the ATF, the explicit forgiveness measures all positively correlated with either one or both of the social desirability scales. However, interpreting the effects of SDR is complicated by the fact that the IAT (combined) significantly *negatively* correlated with both measures of SDR – again suggesting that

these scales are tapping into more than just socially desirable responding tendencies. Furthermore, controlling for the effects of these SDR measures only marginally improves convergence between the IAT and explicit measures, with a strengthening of one of the two significant relationships already found (the relationship between the Forgiveness-Revenge IAT and the HFS increases from *r*=.26 to *r*=.34, but the relationship between Forgiveness-Grudge IAT and the ATF only changes from *r*=.30 to *r*=.32) but no overall gain in the number of significant correlations between the IAT variants and self-report scales. Thus, self-presentation appears to be relatively unrelated to the low convergence found between implicit and explicit forgiveness measures.

Aside from the influence of social desirability, there are a number of other explanations for why implicit and explicit measures might diverge. One possibility is that the two types of measures assess distinct and independent attitudes, which may sometimes correspond with each other and sometimes not (Banaji, 2001; Greenwald & Nosek, 2008; Wilson, Lindsey, & Schooler, 2000)[12]. A second possibility is that specific sources of error are impacting either the IAT (e.g. salience, strategic recoding, etc), self-report measures (e.g. limited introspective ability), or both.

A third explanation is that scores on the two types of measures diverge for structural reasons. An Implicit Association Test and a standard self-report attitude scale differ in many ways beyond merely the implicit-explicit distinction. That is, they are fundamentally different in structure, with recent evidence demonstrating that addressing

[12] This dual attitude account will be further elaborated in Chapter 4

these structural differences can also improve convergence between implicit and explicit measures (Payne, Burkley, & Stokes, 2008).

A structural account may be particularly useful in explaining the low convergent validity found thus far for the forgiveness IAT. It is possible that the IATs and self-report scales used to measure forgiveness thus far have been too structurally dissimilar. The most obvious structural differences have been that the IAT is a *relative* measure of attitudes – it always requires a category to be evaluated in contrast to an alternative – whereas the forgiveness attitude scales used in this study are absolute, requiring no such comparison. As such, a positive score on the Forgiveness-Revenge IAT could be interpreted in several ways. It could be that an individual likes both revenge and forgiveness, but that they like forgiveness slightly more. It could also be that they do not like either of the two constructs, but they dislike forgiveness a little less. If the former is the case, then it should be expected that a positive forgiveness IAT score would be relatively unrelated (or only slightly related) to self-reported forgiveness attitudes, whereas the latter scenario could potentially result in a negative correlation between the two measures. Given this, the low implicit-explicit convergence found in the present study may not be completely unexpected, and it may be possible to increase the correspondence between the two by making some minor modifications to their respective structures. Chapter 4 will address this issue.

Chapter 4:

Addressing the low correspondence between IAT-derived and self-reported forgiveness

4.1 Chapter Overview

Studies 1, 2b, and 3 found that relations between Forgiveness IATs and self-reported measures of forgiveness were either small or non-significant. The current chapter addressed whether convergence between the measures could be improved by making structural modifications to both the IAT and the explicit forgiveness measures. Across two studies, the attitude IAT (pleasant-unpleasant) was replaced by a forgiveness self-concept IAT (self-other), as it was theorised that this should be more conceptually similar to forgiveness attitude scales. For the same reason, the existing attitude scales were complemented by a series of more 'relative' explicit measures (comparing forgiveness with other constructs), to more closely mirror the relative structure of the IAT. Use of a self-concept IAT also allowed for replication of the findings from Study 3, in assessing the suitability of three contrast categories for forgiveness.

4.2 Explanations for high or low convergent validity in IAT research

Convergence between IAT and self-report measures of the same construct generally tends to be low (for reviews see Greenwald et al., 2009; Hofmann et al., 2005a; Hofmann & Schmitt, 2008). However, there is still some debate about how these findings should be interpreted, and the extent to which implicit and explicit measures should *ideally* converge. Some theorists have argued that the low Implicit-Explicit Correspondence (IEC) that has generally been found provides evidence *for* the validity of the IAT. That is, the two kinds of measures should only minimally converge as the IAT is not constrained by a person's introspective limits or self-presentation concerns (Nier, 2005), or that implicit and explicit attitudes are independent constructs (Banaji, 2001;

Greenwald & Nosek, 2008). Others argue that IEC should be much higher than what is typically found, and that it is other features of the measures which prevent them from converging at a higher level (Payne et al., 2008). These are obviously contradictory positions, thereby posing significant problems for the interpretation of IAT data. These conflicting assumptions mean that there is not yet a consensus as to the extent to which the IAT should converge with explicit measures of the same constructs. Thus, it is currently possible to use both IE convergence and IE divergence as evidence that the IAT is an effective measure.

4.2.1 Low IE convergence means the IAT is doing its job

The low correlations found between IAT and self-report measures are often rationalised as being expected, on one of two main grounds. The first rationale is that explicit measures are affected by non-construct-related extraneous factors such as socially desirable responding, whereas implicit measures are not, creating divergence between the two (e.g. Fazio & Olsen, 2003). The second rationale is that implicit and explicit attitudes are independent constructs, and therefore they should be expected to diverge (e.g. Strack & Deutsch, 2004; Wilson et al., 2000).

4.2.1.1 Problems with explicit measures

The IAT was originally designed to address two key limitations of self-report scales: that people may be either unwilling or unable to accurately report their attitudes (Greenwald et al., 1998). Each of these factors imply that explicit measures provide inaccurate or incomplete information about a person's actual attitude, which could

potentially explain their low convergence with IAT measures of the same construct. That is, low IEC is caused by specific error variance in the explicit measures, error variance that does not affect the IAT.

The self-presentation account of IEC is succinctly captured by the MODE (Motivation and Opportunity as DEterminants) model (Fazio & Olsen, 2003). As the model title suggests, MODE argues that the greatest differences between implicit and explicit measures occur when a person is both *motivated* and has the *opportunity* to engage in deliberative, controlled processing. In socially sensitive domains, this allows an individual to moderate their responses on self-report measures, and increases the likelihood of socially desirable responding. One of the core premises used to justify using the IAT is that implicit attitudes are immune to this kind of conscious manipulation. Thus, if an individual attempts to deliberately misrepresent their attitude on a self-report scale, and the IAT is immune to such efforts, lower levels of IEC should be indicative of greater IAT validity.

The MODE explanation for low IEC is empirically well-supported. In their meta-analyses, both Nosek (2005) and Greenwald et al. (2009) found that social sensitivity of the construct was negatively related to the level of convergence between implicit and explicit measures. Specifically, Greenwald et al. (2009) found a significant negative correlation of r=-.35 between the magnitude of IEC and ratings of social sensitivity. The role of self-presentation in moderating the IE relationship has also been demonstrated experimentally, using a bogus pipeline procedure (Nier, 2005). Participants were assigned to one of three experimental conditions, in which they were either (1) primed to believe that the IAT was essentially a lie-detector, (2) primed to believe that the IAT was

extremely unreliable at determining a person's true attitudes, or (3) no priming information. Participants all completed the same race IAT and explicit racism measure. There was no significant correlation between the IAT and the explicit measure in either the 'no' or 'unreliable' prime conditions. However, for those participants who were led to believe that the IAT was able to detect their true attitudes (bogus pipeline), a correlation of r=.51 (p<.001) was found, providing strong evidence that low IEC may sometimes be the result of self-presentation factors.

In addition to self-presentation factors, low IEC may be due to the limitations of conscious introspection. Unlike self-report scales, the IAT is thought to operate at the non-conscious level. Thus, theoretically, IEC should be greater for individuals who are more adept at introspection, and/or in situations where there are greater resources for introspecting. The influence of introspective limits as a moderator of the IE relationship is difficult to assess, as if a person does become consciously aware of an implicit attitude surely it then ceases to be 'implicit'. Despite this difficulty, there have been a few attempts to investigate introspective limits as an explanation for discrepancies between implicit and explicit preferences, at both the trait and state levels. At the level of individual differences, Brown and Ryan (2003) found that those higher in 'mindfulness' – a trait which at least partially refers to a person's ability to consciously engage with their own perceptions – showed a significantly stronger level of implicit-explicit attitude convergence.

A similar moderating effect on IEC has been found for individual differences in another construct that more explicitly taps introspective capabilities: attitude awareness. Across two studies Hofmann, Gschwendner and Schmitt (2005b) investigated attitudes

towards Turks/Germans among a sample of German participants, and the potential moderating effects of attitude awareness and adjustment. Attitude awareness was operationalised as private self-consciousness, attitudinal self-knowledge, and attitude importance, whereas attitude adjustment was operationalised using measures of public self-consciousness, social desirability, prejudice control, and self-monitoring. A main effect was found for only one of the three measures of attitude awareness: attitude importance moderated the convergence between implicit and explicit measures. More interestingly, several three way interactions were observed, with attitude awareness and attitude adjustment interacting to predict differences in IEC. Specifically, the relationship was such that the highest IEC was found when attitude awareness was high and motivation to adjust that attitude was low. These findings suggest that individual differences in introspective capabilities may play some role in moderating IE convergence.

Other attempts to investigate the importance of introspection in moderating IE convergence have taken a more situation-specific approach, by examining/imposing situational constraints that limit a person's ability to introspect. One such constraint is the degree of spontaneity or restrictions on the amount of time taken to report an attitude. That is, an explicit attitude that is reported more quickly/spontaneously (more driven by a 'gut feeling') should be more comparable to an implicitly-measured attitude. In their meta-analysis, Hofmann et al. (2005a) found that spontaneity of explicit measures significantly moderated the IE relationship, such that self-report measures that were deemed to be based more on 'gut feeling' (i.e. low opportunity for introspection) correlated better with their respective IATs.

Some evidence for this relationship has also been found experimentally. Ranganath, Smith and Nosek (2008) sought to limit participants' opportunity to introspect when answering explicit measures by either explicitly asking them to do so (i.e. report your 'gut reaction') or by placing time restrictions on their responses (700ms). Using confirmatory factor analysis to compare two-factor models of the IE relationship (comparing process versus content), they found that both the gut reaction and time-pressured explicit attitudes were a better fit to the model when sharing a factor with implicit measures (IAT and GNAT) than with standard explicit measures.

Together, self-presentation and introspective limits appear to exert at least some moderating influence on the convergence between implicit and explicit measures.

4.2.1.2 Dual attitude accounts

Another prominent account of why IEC *should* be low is that the IAT and self-report scales do not measure the same construct, but rather they measure two distinct (but related) constructs (Greenwald & Nosek, 2008; Wilson, Lindsey, & Schooler, 2000). These theories are often called dual-system (Strack & Deutsch, 2004), dual-process (Smith & DeCoster, 2000) or dual attitude (Greenwald & Nosek, 2008; Wilson et al., 2000) theories, and distinguish implicit and explicit attitudes by the processes that underlie them: associative versus propositional (Gawronski & Bodenhausen, 2006), associative versus rule-based (Smith & DeCoster, 2000), impulsive versus reflective (Strack & Deutsch, 2004).

Common to these theories are two key assumptions: (a) that an individual simultaneously holds implicit and explicit attitudes towards the same construct, and (b)

newly formed attitudes do not replace previously held ones, they merely add another attitudinal layer. Explicit measures may capture more recent (and therefore more easily accessible to conscious introspection) attitudes, while implicit measures capture the relics – those older, more ingrained attitudes that have since been partially 'over-written' by newer attitudinal layers. There has in fact been some empirical support for dual attitude models, with studies using latent variable structural modelling demonstrating that implicit and explicit attitudes can be best represented using two-factor, rather than single factor models (Nosek & Smyth, 2007; Rudolph, Schröder-Abé, Riketta, & Schütz, 2010).

A key tenet of all of these dual attitude models is that neither the explicit nor implicit attitude is considered to be more "true" or more important than the other – the two work in tandem to influence behaviour (Smith & DeCoster, 2000; Strack & Deutsch, 2004). This view is supported by evidence showing that even when the two do not converge they can still predict independent variance in behaviour (Greenwald et al., 2009). Which of the two attitudes is the better behavioural predictor may rely on several factors, such as the relative strength of each attitude (Wilson et al., 2000), or motivation to engage in cognitive effort: newly acquired explicit attitudes require conscious attention which – if absent – may see individuals reverting to more automatic 'habitual' attitudes (Strack & Deutsch, 2004; Wilson et al., 2000). Still another possibility is that the two types of attitudes may predict different *types* of behaviour: implicit attitudes predict spontaneous/automatic behaviour whereas explicit attitudes predict more controlled/deliberated behaviour (Asendorpf et al., 2002). Whatever the explanation, the dual attitude accounts suggest that a high level of implicit-explicit convergence is by no means a prerequisite for the usefulness of either type of measure.

In summary, dual process models propose that implicit and explicit attitudes are two distinct constructs, and therefore they should not be expected to converge at a particularly high level. However, low convergence is not an obstacle to them each having unique utility, with both types of measures being potentially useful in understanding the complexities of an individual's attitudes.

4.2.2 High IE convergence means the IAT is doing its job

4.2.2.1 Problems with the IAT

Perhaps ironically, low IE convergence has also been used as evidence *against* the construct validity of the IAT. Whereas proponents of the idea that IE convergence should be low query the validity of self-report measures, those who believe that IEC should be greater are critical of the validity of the IAT. As summarised by Hofmann and Schmitt (2008, p. 208), the assumption is that "given that the construct validity of established personality questionnaires has been proven, this interpretation implies that indirect measures lack construct validity". There has been much literature devoted to testing the construct validity and limitations of the IAT. Specifically, it has been argued that the IAT can be influenced by a range of method-specific sources of error, although these factors may be different to those which affect self-report scales. Many of these potential sources of error such as valence, salience of (or familiarity with) IAT categories, and heuristic processing, have already been discussed in the present work (Chapter 2).

Of particular interest to understanding the IE convergence is the possibility that IATs may capture at least some element of a person's cultural knowledge or "extrapersonal associations", in addition to their own attitudinal preferences (Karpinski &

Hilton, 2001; Olsen & Fazio, 2004). Specifically, a person may be aware of cultural stereotypes towards specific groups, and this knowledge may serve to increase performance on the 'compatible' IAT block, irrespective of whether or not the person actually endorses that particular stereotype. The more an implicit measure taps extrapersonal associations, the less it should be expected to correlate with explicit attitude measures. Or course, there has been some debate about how to determine exactly what is classified as an 'extrapersonal' association, whether or not these associations should be considered as separate from a person's attitude, and whether or not this distinction is even relevant (Gawronski, Peters, & LeBel, 2008; Nosek & Hansen, 2008a).

Nosek and Hansen (2008a) make two important points in questioning the relevance of the personal/extrapersonal debate. First, they point out that a person's explicit attitudes will often correlate with their predictions of culturally held attitudes (i.e. what do other people think about this issue?). That is, our understanding of social norms is related to our attitudes, even at the explicit level, so why should we expect this process to operate differently at the implicit level? Second, if implicit measures purely tapped cultural knowledge, then we would expect there to be low variability among scores from people from the same culture or group, as they should all share that cultural knowledge. However, there is substantial variability in IAT scores among members of the same cultural group, and these scores can predict individual differences in behaviour (Fazio, Jackson, Dunton, & Williams, 1995; Greenwald et al., 2009).

This argument is further augmented by studies that have examined the known-groups validity of IAT measures: smokers show more positive implicit attitudes toward

and stronger implicit identity with smoking than non-smokers (Swanson, Rudman, & Greenwald, 2001); heavy drinkers strongly associate alcohol and arousal at the implicit level whereas light drinkers do not (Wiers, van Woerden, Smulders, & de Jong, 2002); psychopathic murderers show less implicit dislike for violence than do non-psychopathic murderers (Gray, MacCulloch, Smith, Morris, & Snowden, 2003). Thus, although implicit measures may capture some elements of broader cultural knowledge, this does not necessarily impede the utility of such measures, or their convergence with explicit measures.

It is possible that a number of factors interact to cause low IE convergence. It may be that self-report scales are impacted by both social desirability pressures and limited cognitive resources, while IAT scores are simultaneously affected by several sources of error, which may include cultural influences or any number of features of the IAT design. Irrespective of the source of the low IEC, there is a compelling argument to try and improve it: predictive validity. Greenwald et al. (2009) found that higher degree of IE convergence also increased the predictive validity of each type of measure.

4.2.3 Both can't be true?

Questions regarding the degree to which implicit and explicit measures should ideally converge remain ongoing and unresolved (Hofmann & Schmitt, 2008). A major reason is the lack of understanding and agreement about the theory underpinning what an IAT actually measures. As such, it remains possible to draw the same conclusion about the IATs validity irrespective of whether there is high or low IE convergence. Attempts to reconcile these two positions are further complicated by the fact there "is actually no

possibility for using behavioural evidence to choose decisively among the…..[various] interpretations of dissociation data patterns" (Greenwald & Nosek, 2008, p. 74). As such, the claim that the IAT shows an *appropriate* level of convergent validity with explicit measures is scientifically unfalsifiable, which poses significant problems for establishing the validity of IAT measures.

In light of this difficulty, Hofmann and Schmitt (2008, p. 208) suggest that "we need to abandon the idea that there is a simple solution to the question of convergence and divergence", and instead focus on the conditions under which we expect greater or less convergence between measures, including potential moderators of the process. That is, the focus should not be on *by how much* implicit and explicit measures should converge, but rather on *when* they should do so. As stated by Fazio and Olsen (2003, p.304), the best answer to whether (or when) implicit and explicit measures should converge is "it depends".

Several potential moderators of IEC have been proposed, including self presentation concerns (Nosek, 2005), trait introspective tendencies (Hofmann et al., 2005b), degree of cognitive elaboration (Fazio & Olsen, 2003; Gawronski & Bodenhausen, 2006), expertise with the content domain (Czellar & Luna, 2010), emotional processing ability (Dentale, San Martini, De Coro, & Di Pomponio, 2010), among numerous others. In their meta-analysis of IAT/self-report convergence, Hofmann et al. (2005a) found that moderator variables accounted for more than half of the variability in IEC across studies. Significantly, they found that one of the most important of these moderators was methodological in nature: the degree of IE convergence was moderated by method-

specific factors such as the structure of either the self-report or IAT measures – factors which had little connection to the construct under investigation.

4.3 A moderator of convergence: Structural fit

One of the most novel – and perhaps also one of the most compelling – methodological explanations for low IEC is what has been termed *structural fit* (Payne et al., 2008). Attempts to understand the convergence/divergence of implicit and explicit measures have typically focused exclusively on the fact that one is *implicit* and one is *explicit*, when it may be equally important to focus on the many other ways in which the two types of measures differ. Payne et al. (2008) suggest that one of the most obvious explanations for low IE convergence is that the two types of measures are fundamentally different in *structure*. Across four studies, they systematically showed that modifying the features of the measures to increase their structural similarity had the effect of significantly improving convergence between the two, to a mean *r* of .53 across the studies. This level of correspondence is much greater than that typically seen in implicit measurement research (especially in the investigated domain of racial prejudice). That is, improving structural "fit" also improved convergence, independent of the attitude content.

Specifically, the authors measured implicit attitudes using their own tool – the Affect Misattribution Procedure (AMP: Payne, Cheng, Govorun, & Stewart, 2005). This procedure briefly (100ms) shows participants a picture of a black or white face, then a Chinese character, then a neutral/noise picture (grey square), before asking them to explicitly rate the Chinese character in terms of its 'pleasantness'. Participants are

instructed to ignore all other stimuli (faces/grey squares) except the characters. The 'implicit' part of the procedure occurs through priming: rating a Chinese character as more favourable when it is preceded by images of white (versus black) faces is interpreted as an implicit preference for whites over blacks. The nature of this task allowed for relatively easy modification to create an explicit measure – instead of evaluating the Chinese characters, participants were asked to evaluate the faces, while ignoring the Chinese characters. By using the same type of task as both an implicit and explicit measure, Payne et al. (2008) found IE convergence to be relatively high, and approximately equivalent to the convergence found between explicit measures of the same construct. In light of these findings, they recommended that effort should be made to ensure that implicit and explicit measures be designed to be as structurally similar as possible.

Although Payne et al. (2008) did not specifically address structural fit as it relates to the IAT, there is good reason to believe that this idea may also have merit in explaining the (usually low) relationships found between the IAT and self-report scales. In support of the structural fit account is mounting evidence that the IAT frequently shows low to null correlations not only with explicit measures, but also with other implicit measures of the same construct. Recent work, particularly on implicit self-esteem, suggests that the IAT generally correlates poorly with other *implicit* measures (Cunningham, Preacher, & Banaji, 2001; Krause, Back, Egloff, & Schmukle, 2010; Olsen & Fazio, 2003; Rudolph, Schröder-Abé, Schütz, Gregg, & Sedikides, 2008). If the IAT is correlating equally poorly with explicit measures as it is with other types of implicit measures, this seems to provide

at least some support for the idea that low IE measure convergence is not just due to the implicit/explicit distinction, but could be at least partially due to structural factors.

On the surface, the IAT differs structurally from self-report scales in a multitude of ways. Whereas self-report scales rely on conscious, cognitively elaborated ratings of complex statements, the IAT purports to rely on unconscious, automatic responses to single words (or pictures). As such, altering the structure of either (or both) of these measures to increase structural similarity may be useful in improving convergence between the measures. Fortunately, there are some relatively simple ways in which both the Forgiveness IAT and self-report forgiveness scales can be modified to become more similar in structure.

4.3.1 Improving structural fit #1: Change the structure of the *explicit* measures

The very nature of the IAT dictates that a construct of interest is never evaluated in its own right; it is always evaluated through a comparison/contrast with another target category. Conversely, explicit attitude scales do not typically require one construct to be compared against another. According to the structural fit account, modifying the measures so that they are either both relative or both absolute should result in improved correspondence between the two.

There has been some effort among IAT researchers to use self-report measures that at least partially mirror the relative structure of the IAT. Many of the studies conducted through the Project Implicit website (http://implicit.harvard.edu) by the original developers of the IAT have assessed explicit preferences using 'feeling thermometers', which provide a relative evaluation of affect towards target constructs.

For example, Nosek, Banaji and Greenwald (2002b) asked participants to complete a Maths-Arts/Pleasant-Unpleasant IAT and coupled this with two feeling thermometers on which participants had to independently rate their feelings of warmth toward the concepts "Math" and "Arts" (as academic disciplines), on a scale from 0 (cold/unfavourable) to 100 (warm/favourable). Scores on the "Arts" item were subsequently subtracted from scores on the "Maths" item to give a measure of the participant's relative preference for maths over arts. As such, the resulting measure accurately mirrored the *s*tructure of the IAT in terms of *relativity*, representing a double-dissociation.

Using feeling thermometers to make explicit measures more relative appears to have the desired effect on improving IEC. Nosek (2005) conducted a meta-analysis on data from a very large internet sample gathered via the Project Implicit website, comprising 57 different tasks (across numerous content domains; N=6836[13]). All 57 tasks consisted of an IAT, along with an explicit feeling thermometer for each target category, on which participants indicated levels of warmth towards that category on a 9 point scale. Difference scores were subsequently computed by subtracting one feeling thermometer from the other, and this formed the explicit measure for each task. Meta analysis across the 57 tasks revealed mean IEC of r=.37 (reaching as high as r=.70), with 52 of the 57 tasks showing significant positive correlations between the IAT and parallel explicit measures.

Importantly, the level of correspondence found by Nosek (2005) is substantially greater than that found in other meta-analyses of IEC. The most likely explanation for

[13] 12563 tasks were completed in total, with each participant averaging 1.8 tasks each.

this discrepancy is methodological: the explicit measures used in the 57 tasks analysed by Nosek (2005) were all *relative* measures of preference and were all structurally identical (difference between two feeling thermometers), whereas there was considerably more variability in the structure of explicit measures used in the 126 studies examined by Hofmann et al. (2005a) or the 184 samples analysed by Greenwald et al. (2009). Hofmann et al. (2005a) investigated this structural question directly by comparing IEC as a function of the type of explicit measure that was used. Standard attitude scales, which comprised the largest number of explicit measures (N=74), displayed the lowest IEC of any of the explicit measures (*r*=.18), performing worse than some of the more relative measures like semantic differentials (*r*=.28) or feeling thermometers (*r*=.24). The authors concluded that IEC should reasonably be expected to be greater when self-report measures reflect relative rather than absolute judgements.

4.3.2 Improving structural fit #2: Change the structure of the *implicit* measures

4.3.2.1 Single Target Implicit Association Tests

Alternatively, one could make the IAT *less* relative. This has been attempted through the development of IAT measures that use only one target category, such as the Single Category Implicit Association Test (SC-IAT: Karpinski & Steinman, 2006) or Single Target Implicit Association Test (ST-IAT: Wigboldus, Holland, & van Knippenberg, 2004; as cited in Bluemke & Friese, 2008). These measures largely follow the same procedure as a standard IAT, with the exception that they only have three categories, as opposed to four: a single target category (no pairing) and two paired attribute/evaluative categories (e.g. pleasant-unpleasant). For example, in one of the key blocks both the single target

category and ‘pleasant’ would be sorted to one side of the screen, and ‘unpleasant’ (only) would be sorted to the other. In the other key block this would be reversed such that the single target category and ‘unpleasant’ are sorted with the same response key, and ‘pleasant’ (only) is sorted with the other. Theoretically, this method provides an implicit evaluation of a construct that is not grounded in a relative comparison with any other category.

At face value, using a single category IAT appears to be a promising solution for improving structural fit between the IAT and self-report measures. However, SC-IATs pose an alternative critical methodological problem – they create far greater opportunity for heuristic processing than do standard IATs. Specifically, the task of *strategically recoding* the IAT (Rothermund & Wentura, 2004) is grossly simplified by the removal of one of the target categories, because participants are then able to sort stimuli using a simple sorting rule. For example, imagine a SC-IAT with the target category ‘Animals’ and the attribute pair ‘pleasant’ and ‘unpleasant’. For the block in which ‘Animals’ and ‘pleasant’ are sorted with the same key and ‘unpleasant’ is sorted with the other, the only discrimination that a participant has to engage in is whether the stimuli belong to the ‘unpleasant’ category. Essentially, the rule becomes: “if I see one of the unpleasant words I press this key, and for *anything else* I press the other key”. This is problematic as it strongly invites responses that are not representative of actual attitudes or associations. As noted by Schnabel et al. (2008a) “a lot of questions need to be answered” (p. 213) about the single category IATs before they can be considered superior alternatives to standard IATs.

4.3.2.2 The Self-Concept IAT

To this point, the present chapter has focused on ways that the IAT and explicit measures could each be modified so that they are either both relative (or both absolute) measures. However, this discussion has thus far neglected the idea that they may both already be relative measures, albeit relative in different ways. Attitude scales do not typically instruct participants to directly compare/contrast the construct of interest (in this case 'forgiveness') with another construct: instead they usually require a participant to rate how strongly they endorse a single target. However, just because participants are not instructed to make relative judgements *does not* mean that they do not do so. For example, a forgiveness scale might ask participants to rate (on a scale from 1 to 7) the extent to which they 'like' forgiveness, on which they nominate a '6'. At face value this appears to be a non-relative measure of forgiveness - but how do individuals reach their decision? Nosek (2005) suggests that two comparisons are made: one is intrapersonal and the other interpersonal. At the intrapersonal level, an individual might decide that they like forgiveness, but that there are some constructs that they like *more*, and that a rating of 7 should only be reserved for these such things. At the interpersonal level, the individual may decide that they like forgiveness, but may recognise that there are some *other* people who are more forgiving than themselves.

Of course, the idea that people reach conclusions about their own attitudes through a comparison with others is not a new phenomenon in social psychology. More than fifty years ago Festinger (1954) theorised that attitudes do not operate in a cultural vacuum, and that social comparisons are necessary for understanding our own attitudes and abilities, especially when objective checks of these attitudes or abilities are

unavailable. For instance, there is no single objective test of whether one is a forgiving person or not: conclusions about a person's own forgivingness come from a comparison of knowledge of the many kinds of situations in which forgiveness might be relevant, and then an appraisal of how likely they are - compared to others - to grant forgiveness in those situations. Thus, forgiveness attitude scales may in fact assess relative attitudes, even if their structures do not always make this obvious.

Given the argument that both implicit and explicit attitudes may be seen as relative, attempting to address this structural difference may now seem quite redundant. However, even though both types of attitudes can be perceived as relative evaluations, they are relative in different ways. The 'attitude' IAT used thus far in the present work is a double-dissociation task, and therefore contains information about two different relative evaluations: Forgiveness relative to Revenge (or Grudge, or Justice), as well as Pleasant relative to Unpleasant. In contrast, self-report attitudes scales may be relative appraisals of an individual's own attitudes compared to those of other people's: an evaluation of the self relative to others. Thus, one way in which the two types of measures could become more structurally or conceptually similar is to include a *self-other* comparison in the IAT.

Such an IAT has been implemented quite successfully in the IAT literature, using what has been termed a 'self-concept' IAT (Asendorpf et al., 2002; Greenwald & Farnham, 2000). The self-concept IAT is identical to a standard attitude IAT except that the attribute dimension (e.g. pleasant-unpleasant, good-bad) is replaced with paired categories that contrast the self with others, either generically or idiographically. Generically, this is done using category labels "self" (represented by stimuli words like "I",

"me", "mine", "myself", etc) and "other" (represented by words such as "they", "them", "theirs", etc) (see Asendorpf et al., 2002; Greenwald & Farnham, 2000, Experiment 2; Nosek et al., 2002b; Pinter & Greenwald, 2005). Alternatively, self-concept IATs can be constructed ideographically, following procedures such as those employed (on a self-esteem IAT) by Greenwald and Farnham (2000, Experiment 1). Before completing the IAT, participants were asked to list 18 'self-descriptive' words (such as their own name, hometown) and 18 'non-self-descriptive' words, which were subsequently transformed into the IAT categories of "me" and "not me", for use in a personalised IAT. Regardless of its particular design, the self-concept IAT provides an implicit measure akin to the social comparison used when making explicit attitudinal judgements, and may prove to be useful in improving the convergence between the forgiveness IAT and explicit measures of forgiveness attitudes.

Implementing a self-concept IAT should contribute to improved IEC. Although there have been no direct comparisons between attitude and self-concept IATs of the same construct, there is some evidence to suggest that modifying the evaluative/attribute categories of an IAT can have implications for its convergence with explicit measures (Houben, Nosek, & Wiers, 2010).

4.4 Study 4

4.4.1 Aims and Hypotheses

The main aim of the present study was to determine if modifying both IAT and self-report measures of forgiveness would improve convergence between the two.

Hypothesis 1: Improved structural fit will improve IEC

This hypothesis has both a between-study and within-study component. Data will be analysed using correlations, and a comparison of these across studies. Between studies:

(a) The forgiveness self-concept IAT will correlate with more self-report forgiveness scales than has been found using the attitude IAT

(b) The forgiveness self-concept IAT will show *stronger* correlations with the self-reported forgiveness scales than has been found using the attitude IAT

Within this study:

(c) The forgiveness self-concept IAT will show stronger correlations with an explicit measure of forgiveness *self-concept* (Forgiveness Identification Scale) than with explicit measures of forgiveness *attitudes or dispositions*.

(d) The forgiveness IAT will show stronger correlations with *relative* explicit measures than with *standard* (non-relative) forgiveness scales.

A secondary aim was to replicate the findings of Study 3, using a self-concept IAT. Low IEC found in Study 3 meant that there was little opportunity to explore the way in which different implicit conceptualisations of forgiveness related to explicitly reported forgiveness, or to assess which of the contrast categories was most valid. The anticipated improvement in IE convergence would make this analysis possible.

4.4.2 Method

4.4.2.1 Design

This study utilised a between-groups experimental design.

4.4.2.2 Participants

A total of 217 (144 female, 73 male) participants completed the study, comprising 42 first year undergraduate psychology students, and 175 participants from the broader Australian community. Overall, participants had a mean age of 25.35 years (*SD*=11.10). The undergraduate students were recruited at the University of Adelaide, and participated in exchange for course credit. The community sample was recruited using paid advertisements on the internet social networking site Facebook, with participants self-selecting to participate. The advertisement was entitled "Are you forgiving?" and contained the following description: "We're doing research to find out what people really think about forgiveness. Complete our online survey and tell us what you think". Clicking on the advertisement redirected potential participants to an information page and online consent form. The advertisement was only shown to Facebook users who were aged 18 or older and currently residing in Australia.

4.4.2.3 Materials

4.4.2.3.1 IAT Design and Structure

The same three IAT variants were used as per Study 3 (contrasting forgiveness with revenge, grudge and justice). The attribute categories of 'pleasant' and 'unpleasant'

were replaced with ‘self’ and ‘other’. The ‘self’ category consisted of 5 personal pronouns focusing on the self: I, me, myself, my, mine. The ‘other’ category was comprised of 5 other-focused personal pronouns: they, them, themselves, their, theirs. All other aspects of the IAT design, instructions to participants, and administration, were identical to those described for Study 3.

4.4.2.3.2 *Self-report questionnaire*

The questionnaire consisted of several measures used in the previous studies: the ATF and TTF (Brown, 2003), the “other” subscale of the HFS (Thompson et al., 2005), the Willingness To Forgive scale (WTF: DeShea, 2003), the SDS-17 (Stober, 2001), the Marlowe-Crowne Short Form (MCSF: Ballard, 1992), and demographic items relating to age and gender. There were also two new additions to the questionnaire battery. The first was a five item scale designed by the author to assess the extent to which forgiveness was part of an individual’s self-concept and to provide a forgiveness scale measure that was as conceptually consistent as possible with the forgiveness self-concept IAT. (“Forgiveness Identification Scale” or FIS). The second inclusion was a set of three items that assessed forgiveness as a *relative* construct, designed to mirror the target pairings in the three IAT variants.

Forgiveness Identification Scale (FIS)

This scale consisted of five items: “Generally, I am a forgiving person", “Compared to others, I am a particularly forgiving person", “I would be upset if someone thought that I was unforgiving", “Being forgiving is part of my identity", and “Forgiveness is a part of

who I am". All items were positively-keyed but appeared in random order, shuffled amongst items from the ATF and TTF (some of which were negatively-keyed). Items were rated on a 7 point Likert type scale, ranging from “strongly disagree” to “strongly agree”, and subsequently summed to form a scale score ranging from 5 to 35. The scale displayed good internal consistency reliability (α=.86).

Relative forgiveness measures

Three items were included to assess forgiveness in a way that was structurally similar to the self-concept IAT. This was done in two ways. First, each item contrasted forgiveness with one of the same three IAT contrast categories: revenge, grudge, or punitive justice. Second, each item asked specifically about *what kind of person you are,* as this should be more equivalent to a self-concept IAT than merely asking more generally about whether or not the individual prefers or has favourable attitudes towards forgiveness. For each item, participants clicked on a line to indicate where they believed they belonged, with each line representing a dichotomous choice. Specifically:

1. "I am a forgiving person" to "I am a vengeful person” (reverse-scored)
2. "I hold grudges" to "I forgive”
3. "I think people should be punished" to "I think people should be forgiven”

The selected position on each line was transformed to a score between 0 and 100, such that higher scores indicated a greater preference for forgiveness when compared to that particular alternative.

4.4.2.3.3 *Internal consistency reliabilities*

Cronbach's alpha for the forgiveness scales was acceptable to good: ATF (.73), TTF (.70), HFS (.83), and WTF (.88). The social desirability scales were less reliable: SDS-17 (.65) and MCSF (.58), with no improvement by deleting particular scale items.

4.4.2.4 Procedure

The procedure for this study was identical to Study 3, with the exception of the previously described variations in the materials used.

4.4.3 Results

4.4.3.1 Scoring of IAT

IAT D scores were computed, with 22 participants excluded based on extreme fast responses (remaining N=195). Mean D scores for each IAT version are presented in Table 4.1. Internal consistency reliability for the IAT overall was very good, $\alpha = .85$.

Table 4.1

Means, 95% Confidence Intervals and Standard Deviations for IAT D scores

	N	*M* (95% CIs)	*SD*
IAT (forgiveness-revenge)	63	.60 (.53-.67)	.28
IAT (forgiveness-grudge)	66	.52 (.45-.60)	.32
IAT (forgiveness-justice)	66	.58 (.50-.66)	.32
IAT (combined)	195	.57 (.52-.61)	.31

4.4.3.3 Improving structural fit will improve IEC

Hypotheses 1a to 1d were concerned with the levels of convergence between the IAT and explicit measures, and were assessed using correlations. The results of correlational analyses are reported in Table 4.2.

Table 4.2

Intercorrelations Between IAT D Scores and Self-Report Measures of Forgiveness Attitudes

	1	2	3	4	5	6	7	8	9
1. IAT									
2. IAT:F-R									
3. IAT:F-G									
4. IAT:F-J									
5. ATF	.32***	.45***	.19	.38**					
6. TTF	.15*	.15	.05	.22	.39***				
7. FIS	.20**	.28*	.04	.28*	.53***	.62***			
8. HFS	.27***	.22	.18	.40**	.48***	.62***	.65***		
9. WTF	.07	.10	-.12	.22	.43***	.56***	.48***	.60***	
10. FRrel	.10	.11	.08	.15	.39***	.45***	.53***	.56***	.39***
11. FGrel	-.02	-.13	-.12	.18	.15*	.40***	.40***	.44***	.45***
12. FJrel	.03	-.10	-.07	.26*	.26***	.25***	.31***	.36***	.47***

* p<.05. ** p<.01. *** p<.001.

IAT:F-R = Forgiveness-Revenge IAT, IAT:F-G = Forgiveness-Grudge IAT, IAT:F-J = Forgiveness-Justice IAT, FRrel = single item measure of forgiveness relative to revenge, FGrel = single item measure of forgiveness relative to grudge, FJrel = single item measure of forgiveness relative to justice.

Correlation coefficients were compared across studies 3 and 4 using procedures outlined by Cohen and Cohen (1983) for comparing correlations from independent samples.

Coefficients were first transformed to *z* scores using Fisher's *r* to *z* transformation, and then differences between these *z* scores were computed. The differences in these correlation *z* scores between studies are reported in Table 4.3 (below).

Table 4.3

Correlation z score differences between studies 3 and 4

	IAT	IAT:F-R	IAT:F-G	IAT:F-J
ATF	1.47	3.37***	-1.15	2.74**
TTF	1.77*	1.09	.78	3.37***
HFS	1.03	-.42	.50	3.27***
WTF	1.47	1.47	-1.18	4.38***

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

Note: differences were computed by subtracting Study 3 values from Study 4 values. Thus, a positive difference in *z* scores reflects a stronger correlation in Study 4.

4.4.3.3.1 Using the forgiveness self-concept IAT to improve IEC

As hypothesized, using a self-concept IAT resulted in greater IEC than that observed with the attitude IAT used in the first three studies. The IAT (overall) significantly correlated with four of the five forgiveness scales: the ATF, TTF, HFS and FIS. This is in contrast to Study 3, in which the IAT only correlated with the ATF and HFS, and is a substantial improvement on the limited IEC found in Studies 1 and 2b. Moreover, for the forgiveness-revenge and forgiveness-justice IAT (but not forgiveness-grudge) versions, correlations with several of the explicit measures were significantly greater than those reported in Study 3.

4.4.3.3.2 Using a forgiveness identification scale to improve IEC

Although the Forgiveness Identification Scale *did* significantly correlate with the IAT, the magnitude of this correlation was comparable to other explicit forgiveness measures. Using this scale did not improve IEC.

4.4.3.3.3 Using relative explicit measures to improve IEC

None of the three relative measures of forgiveness significantly correlated with the (combined) IAT. A separate examination of the individual IAT versions revealed that of nine possible relationships with the relative forgiveness measures, only one reached statistical significance: the relative measure contrasting forgiveness with justice correlated with the forgiveness-justice IAT. However, this effect was weaker than that found between the forgiveness-justice IAT with the ATF, FIS and HFS, and not significantly different to the smaller relationships found with the TTF and WTF. Thus, the relative forgiveness items provided no advantage over standard attitude scales in enhancing IEC.

4.4.3.4 Replicating Study 3: Differences between IAT variants

A Oneway ANOVA found no significant differences between the IAT versions, $F(2,192)=1.08$, $p=.34$, $\omega=.11$. An examination of confidence intervals provided further evidence that there were no meaningful differences in the IAT D scores produced by the three IAT variants[14].

[14] 98.3% Confidence intervals for the three comparisons: IAT(revengeXgrudge) CI range = -.08 to .23; IAT(revengeXjustice) CI range = -.17 to .14; IAT(grudgeXjustice) CI range = -.24 to .06.

4.4.4 Discussion

The present study sought to improve convergence between the forgiveness IAT(s) and explicit forgiveness measures, by modifying the two types of measures to be more structurally and conceptually similar. Consistent with hypotheses, modifying the structure of the IAT to include a self-concept rather than a preference evaluation led to an improvement in correspondence with explicit forgiveness measures. This is evidenced in both the increased number of significant correlations, as well as an improvement in the magnitude of several of these correlations. Conversely, modifying the structure of the explicit forgiveness measures to more closely resemble the IAT did not result in any significant improvement in IEC.

In terms of correspondence with explicit forgiveness measures, the forgiveness self-concept IAT used in the present study performed at least equally as well as, and in some cases better than, the forgiveness attitude IAT used in the previous study. This was particularly true for the IAT variants that contrasted forgiveness with either revenge or justice, with the latter showing improvement from Study 3 in convergence with all four explicit measures. Importantly, using a self-concept IAT did not significantly reduce the magnitude of the implicit-explicit correlations that had previously been found using an attitude IAT. Thus implementing a self-concept IAT for forgiveness appears to have been useful for improving IEC.

However, modifying self-report measures to better structurally resemble the IAT did not produce the same benefits. To the contrary, the explicit measures that were included to match the structure of the IAT – the FIS and the three relative measures – converged with the IAT(s) at either a similar level to standard attitude scales, or

performed worse. The FIS was designed as a measure of forgiveness *self-concept.* Thus, it was expected to more strongly correlate with a self-concept IAT than self-report forgiveness measures that did not explicitly assess self-concept (e.g. ATF, HFS). Although the FIS *did* significantly correlate with the IAT, this correlation was only modest. Contrary to prediction, this effect was not significantly different to the IAT's correlation with the ATF, TTF, or HFS. Thus, despite the hypothesised greater conceptual similarity between an explicit and implicit measure of forgiveness self-concept, using an explicit forgiveness self-concept scale was not successful in improving convergence with the forgiveness IAT.

The other way in which explicit measures were modified to more closely reflect IAT structure was to include measures that were explicitly *relative* in nature. Three items required participants to report their attitudes towards forgiveness relative to another construct (either revenge, grudge or punishment). It was hypothesised that these measures would show greater convergence with the IAT(s) than would standard forgiveness attitude scales. Contrary to prediction, the three relative measures performed worse than the other explicit measures in regards to IEC, with only one of nine possible relationships reaching significance. Modifying the self-report measures was not useful for improving convergence between implicitly and explicitly measured forgiveness.

It is possible that measures were not modified in an appropriate way. Specifically, the three relative measures were essentially dichotomous. That is, indicating that one is more forgiving also means indicating that one is (proportionately) less vengeful (for example). In contrast, the IAT does not assess preferences as a dichotomy, but as a *difference* score: an IAT score is computed by calculating the *difference* between evaluations of forgiveness and revenge/grudge/justice associations. As noted in Chapter

3, forgiveness arguably does not have one single indisputable 'opposite', which means that it is unlikely that it is evaluated as dichotomous within an IAT context. This subtle distinction suggests that perhaps the two measures should not be expected to correlate all that well, as they are essentially tapping two distinct types of relative comparisons.

In addition to being limited by their dichotomous nature, the three relative explicit measures may have also been limited by the fact that they only explicitly assessed a single, rather than double-dissociation. Each item asked for participants to indicate the extent to which they thought that they were forgiving relative to revenge (etc), which represents just a single dissociation equivalent to the forgiveness-revenge pairing on the IAT. In contrast, an IAT score represents a double-dissociation. The self-concept IAT includes this same association between forgiveness and revenge but also includes a *social comparison*, measuring a relative comparison of the self relative to others. The explicit relative measures did not collect any information about the relative position of 'others' in understanding one's own forgiveness attitude. Again, this may mean that these explicit items may have been inadequate for capturing the same structural and conceptual nuances that the IAT does.

Together these findings provide mixed support for the role of structural fit in improving convergence between the forgiveness IAT and self-reported forgiveness measures. On the one hand, modifying the IAT to reflect forgiveness self-concept rather than evaluative preference appears to be a promising step forward in developing an IAT forgiveness measure. On the other hand, the modifications made to the explicit forgiveness measures appeared to have been inadequate for assessing forgiveness in a way that is structurally similar to the IAT. Either way, more work needs to be done. The

greater IEC that resulted from using a self-concept IAT – while encouraging – was not overwhelming. Only three of the correlations significantly improved on those from Study 3, and several IE relationships still remained non-significant. In particular, the forgiveness-grudge IAT did not converge with any of the self-report forgiveness measures. Replication of this finding is needed before the self-concept IAT can be adjudged as having superior convergent validity to the forgiveness attitude IAT.

Similarly, further work needs to determine if explicit measures that reflect *difference* scores (not dichotomous scores) can further enhance convergence with the forgiveness IAT. This is not a new idea in IAT research, with the explicit measures derived from the Project Implicit website generally representing a difference score between two relative feeling thermometers (Nosek, 2005; Nosek et al., 2002a). What *is* new is that instead of using feeling thermometers, these difference scores must reflect a double dissociation which includes a social comparison of an individual relative to others. These issues will be addressed in Study 5.

4.5 Study 5

4.5.1 Study Overview

It is possible that the three explicit relative measures developed for Study 4 did not accurately mirror the social comparisons inherent in a self-concept IAT. On the basis of this, the present study aimed to replicate Study 4, using a broader range of relative self-report measures that more accurately mirrored the structure of the forgiveness self-concept IAT.

4.5.2 Method

4.5.2.1 Design

This study utilised a between-groups experimental design.

4.5.2.2 Participants

A total of 224 (193 female, 31 male) participants completed the study, recruited from the broader Australian community via paid advertisements on Facebook. Details of the advertisement were identical to those described for study 4. Overall, participants had a mean age of 32.65 years (*SD*=12.54).

4.5.2.3 Materials

4.5.2.3.1 IAT Design and Structure

The IAT design and structure was identical to that described for study 4.

4.5.2.3.2 Self-report questionnaire

The questionnaire contained the same measures used in Study 4, with the addition of eight new items assessing forgiveness as a relative construct.

Relative forgiveness measures

The three dichotomous items used in Study 4 were retained for the present study, and complemented with eight additional items. These new items were used to generate explicit difference scores of forgiveness attitudes, with four of these also used to obtain a

double-dissociation measure that encouraged a social comparison. Four of these instructed participants to “On the following lines, please indicate where you think you fit ***in general***”, and comprised the following dichotomies:

1. “I am a forgiving person" to "I am not a forgiving person”
2. "I hold grudges" to "I don't hold grudges”
3. "I think people should be punished" to "I don't think people should be punished”
4. "I am a vengeful person" to "I am not a vengeful person"

The position clicked on the line for each item was transformed in to a score ranging from 0 to 100, and reversed scored so that higher values reflected greater endorsement of those specific behaviours. Following this, three separate difference scores were calculated by subtracting scores on item 2, 3, or 4 from the score on item 1. For example, to obtain a measure of a participant’s preference for forgiveness relative to revenge, their score on item 2 was subtracted from their score on item 1. Each of these difference scores had a theoretical range of -100 to +100, with scores greater than 0 representing an explicit preference for forgiveness relative to each of the three alternatives.

The remaining four items were designed to obtain a difference score of forgiveness attitude that also encouraged a social comparison of the self relative to others, and as such were a double-dissociation measure. Participants received the instructions “On the following lines, please indicate where you think you fit ***compared to other people***”, and consisted of the following dichotomies:

1. “Other people are much more forgiving than me" to "I am much more forgiving than other people”

2. "Other people are much more likely to seek revenge than I am" to "I am much more likely to seek revenge than other people are”
3. "Other people are much more likely to support punishment than I am" to "I am much more likely to support punishment than other people are”
4. "Other people hold grudges for much longer than I do" to "I hold grudges for much longer than other people do"

Again, values on each item ranged from 0 to 100, and three difference scores were calculated in the same manner as described above. Each difference score thus mirrored the structure of the IAT insofar as it represented a double dissociation, providing information about forgiveness relative to revenge/grudge/justice as well as how much these constructs were associated with the self relative to other people.

Two additional difference scores were also computed. For each of the two sets of (four) items, a *combined* difference score was calculated. That is, the mean score for the revenge, grudge and justice items was subtracted from the corresponding forgiveness item, to produce a measure that represented evaluations of forgiveness relative to a combination of alternatives.

4.5.2.3.3 Internal consistency reliabilities

Cronbach’s alpha for the forgiveness scales ranged from acceptable to very good: ATF (.74), TTF (.72), FIS (.88), HFS (.82), and WTF (.92). The social desirability scales were less reliable: SDS (.61) and MCSF (.59), with no improvement by deleting particular scale items.

4.5.3 Results

4.5.3.1 Scoring of IAT

IAT D scores were computed, with 27 participants excluded based on extreme fast responses (remaining N=197). Mean D scores for each IAT version are presented in Table 4.4.

Table 4.4

Means, 95% Confidence Intervals and Standard Deviations for IAT D scores

	N	*M* (95% CIs)	*SD*
IAT (forgiveness-revenge)	69	.56 (.47-.65)	.36
IAT (forgiveness-grudge)	66	.50 (.40-.61)	.43
IAT (forgiveness-justice)	62	.61 (.51-.71)	.39
IAT (combined)	197	.56 (.50-.61)	.39

4.5.3.3 Improved structural fit will improve IEC

4.5.3.3.1 Using the forgiveness self-concept IAT to improve IEC

Correlational analyses are reported in Table 4.5, with comparisons between *z* scores for these correlation and those from Study 3 reported in Table 4.6. Consistent with the findings of Study 4, using a self-concept IAT appears to have improved IEC for both the forgiveness-revenge IAT and forgiveness-justice IAT variants, when compared with the attitude IAT used in Study 3.

Table 4.5

Intercorrelations Between IAT D Scores and Self-Report Measures of Forgiveness Attitudes

	IAT	IAT:F-R	IAT:F-G	IAT:F-J
ATF	.08	.29*	-.18	.16
TTF	.02	.29*	-.24	.02
FIS	.06	.22	-.15	.16
HFS	.12	.30*	-.17	.31*
WTF	-.03	.23^	-.30*	.02
FRrelative	.04	.29*	-.20	.03
FGrelative	.12	-.00	.13	.21
FJrelative	.17*	.34**	.09	.06
FRdifference1	.10	.16	.02	.12
FGdifference1	.11	.18	.01	.16
FJdifference1	.09	.22	-.05	.13
FCdifference1	.11	.20	-.01	.15
FRdifference2	.18*	.36**	-.07	.22
FGdifference2	.14*	.39**	-.15	.16
FJdifference2	.15*	.37**	-.09	.13
FCdifference2	.17*	.39**	-.11	.18

* $p<.05$. ** $p<.01$. *** $p<.001$. ^ $p=.055$. IAT:F-R = Forgiveness-Revenge IAT, IAT:F-G = Forgiveness-Grudge IAT, IAT:F-J = Forgiveness-Justice IAT

Note: intercorrelations among the explicit measures were all significant, and ranged from .16 to .94. However, as these relations were of no relevance to the present study, they have been omitted from the table.

Table 4.6

Correlation z score differences between studies 3 and 5

	IAT	IAT:F-R	IAT:F-G	IAT:F-J
ATF	-1.00	1.55*	-4.83***	.40
TTF	.49	2.54**	-2.11*	1.38
HFS	-.50	.43	-2.97**	2.26*
WTF	.49	2.79**	-.30**	2.39**

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

Note: differences were computed by subtracting Study 3 values from Study 5 values. Thus, a positive difference in *z* scores reflects a stronger correlation in Study 5.

The forgiveness-grudge IAT produced some unexpected results, significantly *negatively* correlating with the WTF. Although not significant, correlations between the forgiveness-grudge IAT and the other forgiveness scales also all trended in the negative direction, with all of these correlation coefficients being significantly different to those reported in Study 3. The fact that this IAT variant correlated negatively with the self-report scales may explain why the combined IAT scores did not significantly correlate with the attitude scales.

As there is no theoretical reason that the forgiveness-grudge IAT should be correlating negatively with the explicit forgiveness measures, a further analysis of the data was carried out to explore an explanation for this relationship. IAT scores are a product of performance on two different sorting tasks, typically referred to as the 'compatible' and 'incompatible' blocks, with difference scores calculated between reaction times on these blocks (after controlling for errors). For the forgiveness-grudge

self-concept IAT, the compatible task (blocks 3 and 4) required sorting 'forgiveness' and 'self' stimuli to one side, and 'grudge' and 'other' stimuli to the other side of the screen. The incompatible task paired 'grudge' and 'self' on the same side of the screen, and 'forgiveness' and 'others' on the opposing side. Theoretically, more dispositionally forgiving individuals should respond faster (lower reaction times) on the compatible (forgiveness-self) IAT blocks, whereas they should take longer to respond (higher reaction times) on the incompatible (grudge-self) blocks. The significant negative relationship found between the forgiveness-grudge IAT and the WTF should not be possible if performance on both the compatible and incompatible tasks are in the theoretically expected directions.

To evaluate this, correlations were computed for the forgiveness-grudge IAT (N=66), examining the relationships between the forgiveness scales and reaction times on the compatible (blocks 3 and 4) and incompatible IAT tasks (Table 4.7).

Table 4.7

Correlations between Forgiveness Attitude Scales and Reaction Times for Critical IAT Blocks (Forgiveness-Grudge IAT)

	ATF	TTF	FIS	HFS	WTF
1. Block 3 (Forgive-Self)	-.03	.17	-.05	.12	.27*
2. Block 4 (Forgive-Self)	-.15	.12	-.04	.02	.16
3. Block 6 (Grudge-Self)	-.21	-.04	-.05	-.01	-.04
4. Block 7 (Grudge-Self)	-.20	.03	-.04	.00	.05

* $p<.05$.

The only significant correlation was a significant positive relationship between reaction time on one of the compatible blocks with the WTF scale. This correlation was in the opposite direction to expected, such that those scoring higher on the WTF found it more difficult (slower response) to associate 'forgiveness' with 'self': higher explicit forgiveness was associated with lower implicit forgiveness, which could account for the unexpected negative correlations between the WTF and the forgiveness-grudge IAT.

These correlational analyses were repeated for the forgiveness-revenge and forgiveness-justice IATs but revealed no significant relationships.

4.5.3.3.2 Using a forgiveness identification scale to improve IEC

The FIS was not significantly correlated with any of the IAT variants, and therefore did not improve IEC.

4.5.3.3.3 Using relative explicit measures to improve IEC

The present study included three different types of relative explicit measures of forgiveness: dichotomous (identical to those used in Study 4), single-dissociation difference scores, and double-dissociation (including a social comparison) difference scores. The forgiveness-revenge IAT, but not the forgiveness-justice or forgiveness grudge IATs, significantly correlated with several of these relative measures.

The double-dissociation measures showed the highest overall convergence with the forgiveness-revenge IAT. These correlations were slightly higher (*rs* >.35) than those observed between the IAT and standard attitude scales (*rs* <= .30), however these differences are mostly negligible. Both the forgiveness-grudge and forgiveness-combined

measures converged with the IAT at a significantly greater level than did the WTF (*ps*=.04 and .03 respectively), but no significant differences were reported over the ATF, TTF or HFS. In general, the forgiveness IATs did not show greater convergence with the relative measures compared to standard forgiveness scales.

4.5.3.4 Replicating study 3: Differences between IAT variants

Consistent with the findings of Studies 3 and 4, there were no significant differences between the IAT versions[15], $F(2,194)=1.26$, $p=.29$, $\omega=.11$.

4.5.4 Discussion

The results of this study largely replicated those found in Study 4. Using a self-concept IAT to assess forgiveness resulted in improvements in IEC for two out of the three IAT versions when compared with the attitude IAT used in Study 3. In contrast, although relative explicit measures of forgiveness showed greater convergence with the IAT than those used in Study 4, they still did not offer any significant improvement in IEC over that of the standard forgiveness attitude scales.

The forgiveness-revenge and forgiveness-justice self-concept IATs improved IEC in terms of both the number of significant IE relationships found, with the forgiveness-justice IAT also showing a significant improvement in the magnitude of one of these relationships. This is in stark contrast to the relationships observed for the forgiveness-

[15] 98.3% Confidence intervals for the three comparisons: IAT(revengeXgrudge) CI range = -.10 to .22; IAT(revengeXjustice) CI range = -.21 to .11; IAT(grudgeXjustice) CI range = -.29 to .07

grudge IAT. Correlations between the forgiveness-grudge IAT and all four of the forgiveness scales were significantly different from those reported in Study 3, with all of these relationships unexpectedly trending in the negative direction. One of these correlations – with the WTF – reached statistical significance, again in the opposite direction to that which had been expected.

To more fully understand why this negative relationship had occurred the IAT was separated into its two critical tasks – the compatible and incompatible blocks, with correlational analyses indicating that the source of this trend arose from the compatible (forgiveness-self and grudge-other) block. Specifically, there was a significant positive relationship between scores on the WTF and reaction time for this block, suggesting that those who reported themselves to be more forgiving actually found it more difficult to associate 'forgiveness' with the 'self'. However, there is potentially a reasonable explanation for this. The WTF scale is not a measure of forgiveness *self-concept*, but a measure of how generally – across situations – a person is likely to forgive others. Thus, if the WTF is considered at purely an association level, it gauges the extent to which a person associates 'forgiveness' with 'others' equally as much as – if not more than – the extent to which 'forgiveness' and 'self' are related. For this reason, the significant positive correlation between these measures makes sense: for those who are more willing to forgive *others*, sorting 'forgiveness' and 'others' to opposing sides becomes a more difficult task.

This effect may have been more pronounced on the forgiveness-grudge IAT (versus forgiveness-revenge or forgiveness-justice) because of the more benign nature of the contrast category. On an IAT task, increased compatibility between 'forgiveness' and

'other' also necessarily requires 'grudge' and 'self' to become more compatible, which may not be especially difficult owing to the relatively benign nature of the 'grudge' category. Ideas of revenge and punishment are arguably more severe than those of grudge-holding, and may be less easy to associate with oneself, whereas holding a grudge may have less severe connotations or consequences, and may be more easily associated with the self. Arguably, most people can concede that they can hold grudges, yet it is less easy to perceive ourselves as inherently vengeful. The difficulty associating 'self' and 'revenge/punishment' may have been strong enough to over-ride any potential forgiveness-other associations that may have occurred for the non-grudge IAT versions.

In addition to replicating the findings of Study 4, the present study attempted to build on these findings by introducing a broader range of relative self-report measures, assessing forgiveness-revenge/grudge/justice attitudes not just as a dichotomies but also differences that reflected both single and double dissociations. Consistent with Study 4, the dichotomous forgiveness items did not provide any additional improvement in IEC than that already found for standard attitude scales, although unlike in Study 4, two of these three measures *did* significantly correlate with one of the IATs (forgiveness-revenge variant). Additionally, although the single-dissociation difference scores did not correlate with any of the IATs, all of the double-dissociation difference scores significantly correlated with the forgiveness-revenge IAT. Furthermore, some of these correlations were slightly greater in magnitude than those between the forgiveness-revenge IAT and some of the standard forgiveness scales, although these differences were mostly non-significant. Specifically, they all out-performed the WTF in terms of convergence with the IAT. Thus, explicit measures that mirror the double-dissociation structure of the IAT, but

not single-association or dichotomous measures, may be a useful complement to standard forgiveness attitude scales for research investigating IAT-measured forgiveness.

4.6 General Discussion

4.6.1 Overview of findings

The findings from studies 4 and 5 suggest that the forgiveness-self-concept IAT – particularly the versions comparing forgiveness with either revenge or justice – may be more useful than an attitude IAT for assessing forgiveness at the implicit level. Furthermore, results from these studies have theoretical implications for forgiveness research, both in terms of the relationship between explicit and implicit forgiveness associations, and the way in which forgiveness is represented implicitly.

4.6.2 Improving structural fit

The present work provides some evidence that IEC may be at least partially moderated by structural similarity of the measures, extending the findings of Payne et al. (2008) by demonstrating that structural fit also has specific relevance for the IAT. It was theorised that a self-concept IAT may be more conceptually similar to explicit attitude scales than an attitude IAT is, as explicit attitudes are derived through relative social comparisons (Festinger, 1954), not absolute judgements about preference. In light of this, it was hypothesized that a forgiveness-self-concept IAT would show greater convergence with self-reported attitudes, which is indeed what was found. In both studies 4 and 5, the self-concept (self-other) IATs significantly correlated with more of the explicit forgiveness measures, and correlated more strongly with these measures, than

the equivalent attitude (pleasant-unpleasant) IATs did in Study 3. These results build on previous findings that altering the evaluative/attribute dimension of the IAT can have an impact on IEC (Houben et al., 2010), although this prior work only examined different versions of an attitude IAT. The present studies provide the first direct comparison of an attitude and self-concept IAT for the same content domain, and suggest that utilising self-concept IATs may have the added benefit of improving IE convergence. It is recommended that research on implicit forgiveness proceeds with a self-concept, rather than attitudinal, IAT.

Although modifying the structure of the IAT improved IEC, modifying explicit measures was less successful, but not futile, in enhancing convergence between the measures. Dichotomous items that contrasted forgiveness with one of revenge/grudge/justice either showed no significant correlations (Study 4) or correlations that were lower in magnitude (Study 5) than more traditional attitude scales (ATF, TTF, etc). Single-dissociation difference measures – assessing difference in preferences between forgiveness and the three nominated counterparts – were no more effective, showing no significant correlations with any of the three IATs. Double-dissociation measures, however, fared much better, correlating with the forgiveness-revenge IAT at least as well as other explicit measures. In fact, two of these double-dissociation measures performed significantly better than one of the established forgiveness scales – the WTF. It is perhaps unsurprising that the double-dissociation measures performed best of the relative measures, as these most closely mirrored the structure of the forgiveness self-concept IATs. Notably, the IEC found between these measures and the forgiveness-revenge self-concept IAT (mean r=.38) is among the highest found thus far for

the forgiveness IAT, and is equivalent to the average IEC found among studies that exclusively use relative explicit measures (Nosek, 2005). Thus, using double-dissociation explicit measures may have some utility in complementing the forgiveness IAT.

The fact that the double-dissociation forgiveness measures performed equally as well as the ATF, TTF and HFS - but significantly better than the WTF - may reveal additional insight regarding the impact of structural fit on IE convergence. Across the three studies (studies 3, 4, & 5), the WTF was largely unrelated to all of the IATs. In fact, the only significant IAT-WTF relationship across the three studies was the unexpected negative relationship found in Study 5. In contrast, the ATF and HFS were the best performed of the forgiveness scales, always correlating with at least one of the IATs, with effect sizes up to r=.45 and r=.40 respectively. One possible explanation for this is that the IAT may be more conceptually similar to the ATF and HFS than it is to the WTF. On face value this would make sense. The ATF and HFS assess forgiveness at the *general* level, capturing general thoughts/feelings about forgiveness that are devoid of context. The WTF, in contrast, is context-dependent, providing respondents with a series of hypothetical situations and requiring a decision about whether or not they would actually forgive that person. The IAT operates at the association rather than decisional level, and as such, explicit measures that assess general attitudinal associations rather than behavioural intentions might be reasonably expected to show greater IEC. This association-decision distinction may be an important structural consideration when selecting explicit measures to complement the IAT.

4.6.3 Appropriate contrast categories for the forgiveness IAT

In addition to structural fit, studies 4 and 5 also sought to replicate Study 3, using a forgiveness-self-concept IAT in the place of an attitude IAT. Specifically, the studies aimed to explore the impact of contrasting forgiveness with several other constructs and evaluating which of these was most appropriate for use as a contrast category in the forgiveness IAT. In terms of the magnitude of IAT scores, the choice of contrast category had a negligible impact, with no significant differences in D scores found between the three IAT versions in either of the studies (consistent with Study 3). However, an examination of the differential patterns of convergence among the three IAT versions suggests that the contrast category did matter, with some versions out-performing others in terms of IEC. Specifically, both revenge and justice showed promise as suitable IAT complements to forgiveness, whereas grudge did not.

The forgiveness-revenge and forgiveness-justice self-concept IATs were relatively comparable in terms of their level of convergence with explicit forgiveness scales, both correlating with explicitly-reported forgiveness in both studies 4 and 5. Thus, it appears that the relationship between forgiveness and each of 'revenge' and 'justice' may be implicitly conceptualised in similar ways. Theoretically this is perhaps unsurprising, as both concepts contain ideas of retribution and punishment, although they may differ in terms of control (individual vs society) and legitimacy (Fitness & Peterson, 2008; McCullough, 2008). Prior research has suggested that motivations for retributive justice and revenge are essentially one and the same, with the former merely a socially acceptable permutation of the latter (McKee & Feather, 2008). The present finding adds

further support to this argument, suggesting that this relationship between revenge and justice may also operate at the implicit level.

In contrast to revenge and justice, grudge was not an appropriate candidate to contrast with forgiveness, especially in the self-concept IAT. The forgiveness-grudge attitude IAT (Study 3) initially showed some convergence with explicit forgiveness, significantly correlating with the ATF. However, the forgiveness-grudge self concept IAT (studies 4 and 5) did not significantly correlate with any of the explicit forgiveness measures in the expected (positive) direction. Furthermore, this IAT variant significantly correlated with the WTF scale in the negative (unexpected) direction (Study 5), with other relations between the forgiveness-grudge IAT and explicit measures also trending in this direction. The likely mechanism for these negative relationships appears to be an artefact of two processes: (1) the relatively benign valence of the concept of 'grudge', when compared with revenge and retributive justice, and (2) the greater salience of forgiveness-other associations in the WTF scale compared to the other forgiveness scales. This finding has several implications for the application of the forgiveness IAT. Specifically, it provides scope for potentially using the forgiveness IAT to not only measure attitudinal or dispositional forgiveness, but to also assess situation-specific forgiving behaviour at the implicit level.

This finding suggests that if the self-concept IAT is to be used as a measure of forgiveness, then care must be taken to ensure that the impact of these forgiveness-other associations is minimised. This can be achieved in two key ways. The first is to not use 'grudge' as the contrast category for forgiveness, as this unexpected correlation between the IAT and explicit forgiveness was only observed for the forgiveness-grudge IAT. Using

'revenge' and 'justice' as the contrast categories not only eliminated this effect, but actually produced correlations with explicit measures that were in the expected directions, possibly because the incompatibility of 'revenge/punishment' and 'self' was able to override the unwanted associations between 'forgiveness' and 'other'. The second option is to only use explicit measures that assess forgiveness at a purely dispositional (no context) level, and avoid those that include information about context or specific 'others' (hypothetical scenarios).

Across three studies both revenge and retributive justice appear to be appropriate contrast categories for the self-concept IAT, with each displaying reasonable convergent validity with self-report scales. However, the forgiveness-revenge IAT is recommended over the forgiveness-justice IAT as the most suitable way of operationalising forgiveness within the IAT. This is for two reasons. The first reason is its performance in regards to IEC. The forgiveness-revenge IAT reported the highest percentage of significant correlations with standard forgiveness scales across the three studies. Importantly, the forgiveness-revenge IAT was also the only IAT version that significantly correlated with any of the relative explicit measures. The second reason is theoretical. Opting for 'revenge' is consistent with much of the theorising that already exists in the forgiveness literature: revenge is already established as a credible opposite for forgiveness, being included in existing forgiveness definitions (Enright & The Human Developmental Study Group, 1991; McCullough et al., 1997; Worthington, 1998) and measures (McCullough et al., 1998; Rye et al., 2001).

Studies 3, 4 and 5 addressed validity issues concerning the categories used in the forgiveness IAT, both at the target (revenge/grudge/justice) and attribute (pleasant-

unpleasant vs self-other) category levels. At the target category level, revenge appears to be the most appropriate category, while at the attribute category level a self-concept IAT was able to improve structural fit with explicit measures. Overall, the structural fit account appears to be useful for explaining divergence between explicit and implicit measures of forgiveness. At best, IEC for forgiveness reached *r*=.45 (this is especially so when using the forgiveness-revenge IAT), explaining approximately 20% of the variance in the implicit-explicit relationship. This means that we can have some confidence that the two measures are assessing the same construct. However, there is still much variance in this implicit-explicit relationship that remains unaccounted for, and – as noted earlier – there is no consensus on what constitutes an ideal or acceptable level of IE convergence. Thus, IEC has only limited utility in terms of assessing the validity of an IAT measure. In light of this, additional methods of assessing an IAT's validity must be utilised.

4.6.4 Moving beyond IEC: Predictive validity

Perhaps the most compelling method for validating an IAT is to demonstrate that it has predictive validity. As already noted in this chapter, the IAT has been shown to predict a wide variety of behaviour, often remaining a unique predictor even after controlling for explicit attitudes towards the same construct (Greenwald et al., 2009). Furthermore, IATs can reliably predict behaviour, even when they show low correlations with self-report measures (Hofmann et al., 2005a). An obvious next step in the development of a forgiveness IAT is to demonstrate that it can predict unique variance in actual forgiving behaviour, above that already explained by explicitly reported forgiveness attitudes.

Chapter 5:

Using the Forgiveness-Revenge IAT to predict forgiveness of a recalled transgression

5.1 Chapter overview

The forgiveness-revenge IAT has now been assessed against a range of criteria, including resistance to socially desirable responding, immunity to valence-driven strategic recoding, and convergence with self-reported measures of forgiveness. However, none of these criteria provide information about the utility of the forgiveness-revenge IAT: it is yet to be determined if a forgiveness-revenge IAT is any more useful for assessing dispositional and attitudinal forgiveness than the (self-report) measures that already exist. Specifically, it has not yet been established whether the forgiveness-revenge IAT can predict unique variance in forgiving behaviour that is not already captured by self-report scales. This chapter presents data from three studies (studies 6, 7, & 8) that assessed the predictive (and incremental) validity of the forgiveness-revenge IAT in regards to real-life recalled transgressions.

5.2 Introduction

5.2.1 Predicting forgiving behaviour

There are a number of factors which can predict forgiving behaviour, with situational factors tending to be better predictors than dispositional factors (for a review see Fehr, Gelfand, & Nag, 2010). Forgiveness is predicted by situationally-dependent factors such as apology (Darby & Schlenker, 1982; Eaton et al., 2006; McCullough et al., 1998), perceived severity of the transgression (Fincham et al., 2005), perceived intentionality (Struthers, Eaton, Santelli et al., 2008; Young & Saxe, 2009), relationship commitment (Finkel, Rusbult, Kumashiro, & Hannon, 2002), empathy (McCullough et al., 1997) and rumination (Burnette et al., 2007; McCullough et al., 2007). A recent meta-

analysis revealed that forgiveness was most strongly related to transgressor-related factors such as apology (*r*=.42) and perceived intent (*r*=-.49), and victim-focused states such as empathy (r=.51) and anger (*r*=-.41) (Fehr et al., 2010).

Nonetheless, dispositional variables can also play an important role. Forgiveness is related to Big Five traits such conscientiousness (Balliet, 2010), agreeableness and neuroticism (Brose et al., 2005; McCullough & Hoyt, 2002), trait anger (Lawler-Row, Karremans, Scott, Edlis-Matityahou, & Edwards, 2008), and trait empathy (Toussaint & Webb, 2005a). However, of these dispositional factors, none have been found to be more predictive than *direct* measures of dispositional forgiveness. In their meta-analysis of 175 forgiveness studies, Fehr et al. (2010) found dispositional forgiveness to be the strongest trait-level predictor of forgiving behaviour. Specifically, across the 30 studies that measured forgiveness at both the dispositional and situational levels (combined N=5685), there was a mean correlation of .30 between the two. Of particular note, is that this relationship was stronger than that found for several of the situational predictors of forgiveness, including offense severity, relationship closeness and relationship commitment.

5.2.2 Using the IAT to predict behaviour

If the forgiveness-revenge IAT is a useful measure of forgiveness attitudes/disposition, then it should be expected to predict state-level forgiveness. Furthermore, to be considered a useful measure, the forgiveness-revenge IAT should be able to predict variance in forgiving behaviour that is not already captured by existing measures of forgiveness attitudes. The literature suggests that the forgiveness-revenge

IAT has the potential to meet both of these requirements, with IATs showing both predictive and incremental validity across a wide variety of (non-forgiveness) domains.

For example, political preference IATs predicted voter behavior in the 2005 Italian parliamentary elections (Arcuri, Castelli, Galdi, Zogmaister & Amadori, 2008); a Swedish-Arab IAT predicted workplace hiring decisions of persons belonging to these ethnic groups; implicit conscientiousness has been found to predict the academic performance of Italian university students (Vianello, Robusto, & Anselmi, 2010); implicit aggressiveness has been shown to predict aggressive/punishment responses to provocation (Richetin, Richardson, & Mason, 2010); and scores on a life-death self-concept IAT have been found to significantly predict prospective suicide attempts (next 6 months) among psychiatric patients (Nock, Park, Finn, Deliberto, Dour, & Banaji, 2010). In fact, the relationship between IAT measures and behaviour is fairly robust, and not dissimilar to that observed between explicit attitude and behavioural measures. In a meta-analysis across 184 independent samples (N=14900), Greenwald et al. (2009) found that implicit measures were predictive of behaviour at a mean *r* of .27, compared with a mean *r* of .36 for parallel explicit measures, with each possessing superior predictive utility in specific construct domains.

Of particular relevance to the present work is the fact that IAT measures can sometimes be *superior* predictors of behaviour than their explicit counterparts. In a laboratory setting, Jung and Lee (2009) found that an honesty-deception IAT significantly predicted whether participants would cheat on a task to increase their monetary reward, whereas explicit measures of honesty/deception did not predict cheating behaviour. Similar results have been found in applied contexts such as pilot risk-taking (Molesworth

& Chang, 2009) and teacher prejudice (Van den Bergh, Denessen, Hornstra, Voeten, & Holland, 2010).

Nonetheless, studies in which IAT measures are superior to self-report measures in predicting behaviour appear to be the exception, rather than the rule. In their meta-analysis, Greenwald et al. (2009) found that the relative predictive utility of each type of measure was largely a function of the particular construct domain. Explicit measures tend to fare better in most domains, especially in clinical applications, and for political and consumer preferences. Across studies, IAT measures out-performed explicit measures in intergroup research contexts, especially prejudice research. The authors proposed that this effect was moderated by socially desirable responding, and found evidence to support this claim. Specifically, social sensitivity of the attitude domain significantly correlated with the size of the IAT-behaviour relationship (r=-.18). This finding suggests that IATs may be superior predictors of behaviour in domains where there is more motivation for an individual to conceal their true attitudes.

As argued in Chapter 1, forgiveness may be one such domain in which a person is motivated to conceal their true attitudes, as forgiveness is commonly seen as a 'pro-social' and desirable attribute/behaviour (Bono et al., 2008; Enright & North, 1998a). Evidence from the present work supports this claim, with measures of socially desirable responding regularly correlating with explicit measures of forgiveness attitudes. The forgiveness-revenge IAT has the potential to overcome these concerns.

5.3 Study 6

5.3.1 Study overview

In this study, forgiving behaviour was investigated retrospectively using a standard approach that is pervasive in the forgiveness literature (see McCullough et al., 2000). Participants are asked to reflect on a personally experienced interpersonal transgression, and report the extent to which they had forgiven on the basis of avoidance, revenge and benevolent motivations towards the transgressor.

The study was concerned with both the *predictive* and *incremental* validity of the forgiveness-revenge IAT. In terms of predictive validity, the study aimed to determine if the IAT could predict forgiveness of a personally-experienced transgression. The study also aimed to determine if the forgiveness-revenge IAT could predict any additional variance in behaviour over and above that predicted by explicit measures of forgiveness.

Hypotheses were tested using three separate multiple regression models, regressing avoidance, benevolence and revenge [16] behavioural motivations on trait/attitude forgiveness scales and the forgiveness-revenge IAT. As there is evidence that socially desirable responding motivations may moderate the attitude-behaviour relationship, SDR was controlled for in the first step of each of the analyses.

Specifically, after controlling for SDR, it was hypothesised that:

i. Explicit forgiveness attitudes would positively predict forgiveness (greater benevolence, less avoidance and revenge) of a recalled transgression

[16] Avoidance, revenge and benevolence are well-established indicators that forgiveness has taken place (see McCullough et al, 1998; 2006).

ii. The forgiveness-revenge IAT would positively predict forgiveness (greater benevolence, less avoidance and revenge) of a recalled transgression

iii. The forgiveness-revenge IAT would continue to significantly predict forgiving behaviour after controlling for self-reported forgiveness attitudes

5.3.2 Method

5.3.2.1 Design

This study utilised a single factor correlational design.

5.3.2.2 Participants

Participants were 147 (94 female, 53 male) first year undergraduate psychology students at the University of Adelaide, with a mean age of 20.54 years (*SD*=4.90). Participation was in exchange for course credit.

5.3.2.3 Materials

5.3.2.3.1 IAT Design and Structure

A forgiveness-revenge self-concept IAT was used, with the category labels 'forgiveness', 'revenge', 'self' and 'other'. Stimuli for these categories were identical to those described for the forgiveness-revenge IAT used in studies 4 and 5. All other aspects of the IAT design, instructions to participants, and administration, were identical to those described for the studies in the previous chapter.

5.3.2.3.2 Explicit measures of dispositional forgiveness/forgiveness attitudes

Three measures of dispositional forgiveness were included: the ATF and TTF (Brown, 2003), as well as the "other" subscale of the HFS (Thompson et al., 2005). Based on recommendations from the previous chapter, the WTF (DeShea, 2003) was omitted, and the two items used to construct the double-dissociation measure of forgiveness relative to revenge were again included (although the dichotomous and single-dissociation measures were not).

5.3.2.3.3 Explicit measures of situational forgiveness/forgiving behaviour

Forgiving behaviour was assessed by asking participants to recall (and write a brief summary of) a personally experienced transgression in which they were the victim, and then complete the TRIM-17 (McCullough et al., 2006).

Participants were provided with the following instructions to prompt them to think about an interpersonal transgression from their own life:

> *"This study requires that you think of a time when someone did something to you which hurt you in some way. This person could be someone you were close to, or someone you did not know very well at all. The offence that they committed against you could be something quite minor, or it could be something quite serious/severe. It may have happened a long time ago, or it might be something that happened quite recently. No matter what the event, please try and remember it in as much detail as possible.*

For a moment, visualise in your mind the event and the person who hurt you, and try to recall what happened. Below is a set of questions about the person who hurt you and the event. Specifically, try and reflect on how it felt to be hurt in this way, and by this person."

These instructions were followed by a text box which could accommodate responses of up to 1000 characters. A separate text box was provided for participants to report how long ago this offense took place.

Situation-specific forgiveness was assessed using the TRIM-17 (McCullough et al., 2006). The TRIM-17 consists of three subscales which assess post-offense behavioural motivations towards a transgressor: revenge, avoidance and benevolence motivations. The revenge subscale consists of five items, such as "I'll make him/her pay", while the avoidance subscale includes seven items such as "I keep as much distance between us as possible". The five item benevolence subscale includes items such as "I want us to bury the hatchet and move forward with our relationship". Each item is measured on a 5 point Likert-type scale, ranging from "strongly disagree" (1) to "strongly agree" (5). Totals were computed for each subscale, with each one acting as a proxy for forgiveness: less avoidance and revenge and greater benevolence represents greater situation-specific forgiveness.

5.3.2.3.4 Social Desirability Measures

The present study introduced an alternative measure of socially desirable responding: the Balanced Inventory of Desirable Responding (BIDR: Paulhus, 1986). The BIDR consists of two subscales – Impression Management (IM) and Self Deceptive

Enhancement (SDE) - each comprised of 20 items. Responses are recorded on a 7 point Likert-type scale, indicating agreement with a series of statements, with half of the items from each subscale negatively keyed. Total scores for the two-subscales are based on the number of extreme responses, with item scores of 6 or 7 equating to a score of 1 for that item, and ratings of 5 or below equating to 0. Thus, each subscale has a possible range between 0 and 20.

The BIDR seems to be an appropriate choice for use alongside the IAT because it may be able to disentangle some of the differential processes associated with SDR that the IAT attempts to address. Specifically, it has been hypothesized that the IAT measures attitudes that an individual may be either *unwilling* or *unable* to report on self-report scales (Greenwald & Banaji, 1995) – the two subscales of the BIDR each represent one of these. The IM scale assesses self-presentation tendencies or 'other-deception' – the extent to which an individual is *unwilling* to report their true attitudes. The SDE scale, in contrast, measures 'self-deception' – the extent to which individuals are *unable* to report their true attitudes, largely due to ego-enhancing positivity bias in their self-judgements. The IM subscale includes items such as "I never cover up my mistakes" and "I always obey laws, even if I'm unlikely to get caught", while the SDE subscale includes such items as "I always know why I like things" and "I rarely appreciate criticism".

Due to the separation of the different component of desirable responding, the BIDR allows for a more nuanced analysis of SDR than either of the SDS-17 (Stober, 2001) or Marlowe-Crowne Short Form (MCSF: Ballard, 1992). Based on this fact, and the low reliability found for both the SDR and MCSF in studies 1 to 4, each of these previously-used measures of SDR were omitted from the present study.

5.3.2.3.5 Internal consistency reliabilities

Cronbach's alpha for the scales were borderline acceptable for the ATF (α=.68), TTF (α=.67) and good for the HFS (α=.81). The three TRIM subscales all demonstrated very good internal consistency, on avoidance (α=.94), benevolence (α=.91), and revenge (α=.84). Both subscales of the BIDR also showed an acceptable level of internal consistency: α=.75 for SDE, and α=.76 for IM.

5.3.2.4 Procedure

The study began with the invitation for participants to recall and write a summary of the transgression that they had experienced, after which they completed the TRIM.

The remaining part of the study was divided in to two conditions, which were counter-balanced by random allocation. Following the TRIM, one group of participants completed the forgiveness-revenge IAT *before* the explicit measures (forgiveness attitudes, SDR, demographics), while the other group completed the IAT *after* the explicit measures. The only difference between groups was presentation order – all participants completed all of the same measures.

5.3.3 Results

5.3.3.1 Data preparation

IAT scores were computed using the revised D scoring algorithm (Greenwald et al., 2003) following the same procedures already outlined in the present work. This resulted in 23 participants, for whom more than 10% of IAT trial responses were faster than

300ms, being excluded from the analysis (new N=124). The mean D score was .52 (*SD*=.30).

5.3.3.3 Correlational analyses

Correlations are presented in Table 5.1. The IAT was only significantly related to one of the state-level indicators of forgiveness: avoidance.

Table 5.1

Intercorrelations between IAT D Scores, Forgiveness Attitude Scales, and State-level Forgiveness

	1	2	3	4	5	6	7	8	9
1. IAT									
2. ATF	.08								
3. TTF	.03	.35***							
4. HFS	.06	.50***	.48***						
5. F-R difference	.01	.41***	.38***	.52***					
6. TRIM-avoidance	.18*	-.16	-.25**	-.28**	-.18*				
7. TRIM-revenge	.13	-.18*	-.11	-.30**	-.07	.58***			
8. TRIM-benevolence	-.12	.30**	.20*	.38***	.21*	-.83***	-.65***		
9. BIDR-SDE	.11	.21*	.15	.09	.06	-.05	-.04	.04	
10. BIDR-IM	.11	.37**	.15	.37**	.39***	-.05	-.09	.11	.26**

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

5.3.3.4 Predicting forgiving behaviour

Three hierarchical multiple regressions were performed to investigate the relative contributions of explicit and implicit forgiveness attitudes in predicting forgiving responses (avoidance, revenge, and benevolence), after controlling for SDR and presentation order of measures. Control variables (BIDR and order) were entered at the first step, followed by explicit forgiveness scales (step 2) and the forgiveness-revenge IAT (step 3). The outcomes of these regression analyses are presented in Table 5.2.

Table 5.2

Hierarchical Multiple Regressions of Self-reported and IAT-measured Forgiveness Attitudes on Forgiving Response (Avoidance, Benevolence, Revenge), Controlling for Social Desirability and Order of Measures

	Avoidance			Revenge			Benevolence		
	β	R^2	ΔR^2	β	R^2	ΔR^2	β	R^2	ΔR^2
Step 1:		.04	.04		.02	.02		.03	.03
BIDR-SDE	-.03			-.01			.08		
BIDR-IM	-.00			-.07			.02		
Order	.19*			.11			-.15		
Step 2:		.11	.07*		.09	.07*		.17	.14**
ATF	-.01			-.06			.16		
TTF	-.14			.05			-.02		
HFS	-.20			-.29*			.32**		
Step 3:		.16	.05**		.12	.03^		.20	.03^
IAT	.23**			.17^			-.17^		

* $p<.05$. ** $p<.01$. ^ $p<.07$.

As predicted, explicit forgiveness attitude measures significantly predicted forgiveness of a recalled transgression, accounting for between 7% and 14% of the variance in avoidance, revenge and benevolence. Of the explicit measures, the HFS was the only significant independent predictor.

Also consistent with hypotheses, the forgiveness-revenge IAT uniquely predicted avoidance, contributing an additional 5% of variance above that explained by explicit attitude measures. Similar effects were observed for revenge (3%, p=.061) and benevolence (3%, p=.053), although these were statistically marginal. However, contrary to prediction, these effects were in the opposite direction to expected. The forgiveness-revenge IAT *negatively* predicted forgiveness (greater avoidance and revenge, less benevolence). In contrast, explicit forgiveness attitudes predicted behaviour in the expected direction for all three TRIM subscales.

One possible explanation for this finding is that the presentation order of measures impacted scores on the IAT. Presentation order was a significant predictor of avoidance such that those who completed the IAT *before* explicit attitude measures were significantly higher in avoidance motivations. As we would expect, an independent t-test supports this, with the two groups differing significantly on avoidance, $t(122)=2.21$, p=.03, but not on revenge or benevolence. However, presentation order should not have influenced scores on the TRIM, as all participants completed this part of the study before counter-balancing took place. Therefore, the impact of presentation order on avoidance appears to have occurred by chance, owing to naturally occurring differences in forgiveness between the two groups.

Although order effects can be ruled out in explaining mean differences in TRIM scores, scores on the measures that *were* counter-balanced may have still been susceptible to priming. Further comparison of differences between groups indicates that there were also significant differences on other variables, including the IAT, $t(122)=2.11$, $p=.04$; the ATF, $t(122)=2.45$, $p=.02$; and the HFS, $t(122)=3.53$, $p<.01$. These differences were such that participants who completed the IAT before the explicit attitude measures (IAT $M=.57$, $SD=.29$; ATF $M=30.75$, $SD=4.40$; HFS $M=31.03$, $SD=5.77$) were less forgiving overall than those who completed the explicit measures first (IAT $M=.46$, $SD=.29$; ATF $M=28.60$, $SD=5.37$; HFS $M=27.47$, $SD=5.47$) – on both attitude and situational measures. However, this overall difference in both dispositional and situational forgiveness still does not account for the negative *relationship* found between the IAT and TRIM-measured forgiveness.

To explore whether presentation order significantly moderated the relationship between the IAT and TRIM, the three multiple regressions (avoidance, revenge, benevolence) were repeated for each order condition separately. Outcomes of these analyses are presented in Tables 5.3 and 5.4.

Table 5.3

Hierarchical Multiple Regressions of Self-reported and IAT-measured Forgiveness Attitudes on Forgiving Response (Avoidance, Benevolence, Revenge), Controlling for Social Desirability, for Participants Who Completed the IAT Before the Forgiveness Attitude Scales (N=60)

	Avoidance			Revenge			Benevolence		
	β	R^2	ΔR^2	β	R^2	ΔR^2	β	R^2	ΔR^2
Step 1:									
BIDR-SDE	-.08			-.01			.06		
BIDR-IM	.08			.00			-.12		
		.01	.01		.01	.01		.02	.02
Step 2:									
ATF	.06			.05			.19		
TTF	-.26^			-.00			.08		
HFS	-.33*			-.36*			.35*		
		.20	.19*		.15	.14*		.23	.21**
Step 3:									
IAT	.41**	.35	.15**	.26^	.20	.06^	-.36**	.34	.12**

* $p<.05$. ** $p<.01$. ^ $p<.07$.

Table 5.4

Hierarchical Multiple Regressions of Self-reported and IAT-measured Forgiveness Attitudes on Forgiving Response (Avoidance, Benevolence, Revenge), Controlling for Social Desirability, for Participants Who Completed the IAT After the Forgiveness Attitude Scales (N=64)

	Avoidance			Revenge			Benevolence		
	β	R^2	ΔR^2	β	R^2	ΔR^2	β	R^2	ΔR^2
Step 1:									
BIDR-SDE	-.07			-.08			.04		
BIDR-IM	-.01			-.14			.15		
		.01	.01		.03	.03		.03	.03
Step 2:									
ATF	-.08			-.20			.14		
TTF	-.00			.07			-.20		
HFS	-.11			-.12			.37*		
		.03	.02		.09	.06		.16	.14*
Step 3:									
IAT	.07	.03	.01	.06	.10	.00	.03	.17	.00

* $p<.05$. ** $p<.01$. ^ $p=.07$.

As can be seen from these analyses, the observed negative relationship between the IAT and TRIM subscales only occurred in the condition in which participants completed the IAT *before* the explicit attitude measures (Table 5.3). In this condition, the IAT explained an additional 15% in avoidance and 12% in benevolence motivations above that accounted for by self-reported forgiveness attitudes, but still in the opposite direction to hypothesised. There was also a marginal effect ($p=.059$) of the IAT on revenge, accounting for an additional 6% of the variance. No significant relationships with the IAT were observed in the other condition (Table 5.4).

5.4 Further analysis

Contrary to prediction, the forgiveness-revenge IAT did not significantly positively predict any unique variance in forgiveness of a real situation. In fact, the opposite effect was observed, with the IAT significantly *negatively* predicting forgiving behaviour. However, presentation order moderated this relationship, suggesting that the effect may have been a methodological artefact.

The observed difference between the two counter-balanced conditions suggests that IAT responses may have been primed by situational factors. The observed negative relationship occurred in the condition in which participants completed the IAT *immediately after* recalling a transgression and completing the TRIM. It is plausible that asking people to think about a transgression, a transgressor, and forgiveness in relation to these, interfered with people's ability to respond to the IAT. Specifically, and similar to the effect observed with the WTF in Study 5, the self-concept IAT may behave differently when a situational context is provided, such that thinking about a specific transgressor (an 'other') may prime an increased implicit association between 'forgiveness' and 'other'. This explanation is consistent with work showing that IAT effects can be derived from increased performance when the most salient categories are paired, regardless of whether or not these capture the intended nominal associations (Rothermund & Wentura, 2004; Rothermund et al., 2005). This *salience asymmetry* account of IAT effects has already been discussed in Chapter 2 of the present work (see section 2.2.1). Thus, the presence of a 'forgiveness-other' prime (thinking of a person who hurt you) may interfere with the intended 'forgiveness-self' association that the IAT was designed to assess, such that either combination can represent greater forgiveness.

If the negative relationship between the IAT and state forgiveness was a function of contextual priming, then it is curious that it only occurred in one of the counter-balanced conditions. In both conditions the IAT was completed *after* the TRIM, meaning that all participants had been asked to reflect on a forgiveness context prior to completing the IAT. The difference may have resulted due to the length of time that elapsed between completing the two measures. In the condition where the effect was observed, the IAT had been completed *immediately after* the TRIM scales, but the second (unaffected) condition required completion of the explicit attitude scales in between. Completing these explicit measures may have either acted as (a) a filler task, during which the IAT priming effects wore off, or (b) a counter-prime (reinforcing forgiveness-self associations), which cancelled out the effects of the initial prime without completely overcoming/reversing them.

One way to explore these potential priming effects is to divide the IAT in to its separate blocks, and compare reaction times and error rates across counter-balanced conditions. The "incompatible" blocks (where forgiveness-other and revenge-self are paired) are critical here, as it is the forgiveness-other association that should differ as a function of priming. That is, the task should have been easier (evidenced in reaction times and error rates) in the condition where the IAT immediately followed the TRIM, as recalling a transgression should make forgiveness-other associations more salient. Specifically one should find that:

i. Reaction times for the forgiveness-other IAT blocks (blocks 6 and 7) would have been significantly faster in condition 2 (priming) than in condition 1 (no priming).

ii. Reaction times for the forgiveness-self IAT blocks (blocks 3 and 4) would not have differed significantly as a function of counterbalancing condition.

iii. Error rates for the forgiveness-other IAT blocks (blocks 6 and 7) would have been significantly lower in condition 2 than condition 1.

iv. Error rates for the forgiveness-self IAT blocks (blocks 3 and 4) would not have differed significantly as a function of counterbalancing condition

Data examining these first two hypotheses, addressing reaction time, are reported in Table 5.5 (below).

Table 5.5

Differences in Reaction Times (in Milliseconds) for Compatible and Incompatible IAT Blocks Across Counterbalanced Conditions.

Block	Cond	Mean	*SD*	*t*-test
Block 3 (compatible)	1	837.13	164.27	$t(122) = .82$, $p = .41$
(forgiveness-self)	2	814.77	135.97	
Block 4 (compatible)	1	755.06	111.88	$t(122) = -.11$, $p = .91$
(forgiveness-self)	2	757.30	116.44	
Block 6 (incompatible)	1	1110.45	320.03	$t(122) = 2.08$, $p = .04$*, $d = .38$
(forgiveness-other)	2	1005.83	229.36	
Block 7 (incompatible)	1	919.49	216.42	$t(122) = .34$, $p = .73$
(forgiveness-other)	2	907.28	177.73	

* $p<.05$. Conditions: 1 = IAT after self-report scales, 2 = IAT before self-report scales

Reaction times in block 6 differed in the expected direction, whereas reaction times did not differ significantly in blocks 3 and 4 as a function of presentation order. Means for block 7 were in the expected direction but this difference was not significant. The most likely explanation for this difference between block 6 and 7 is that participants may have been getting bored towards the end of the IAT and not taking it as seriously in either condition. Error rate data (reported in Table 5.6) support this supposition – error rates for block 7 are much higher than for the other critical blocks, suggesting decreased concentration on that block.

The second table (Table 5.6) shows error rates for the four IAT blocks, revealing that contrary to predictions:

i. There were no differences between conditions for the forgiveness-other blocks (blocks 6 and 7)

ii. There were significant differences between conditions for the forgiveness-self block (blocks 3 and 4)

This means that, in the condition where the IAT was primed, participants either found it more difficult to associate 'forgiveness' with 'self', or found it more difficult to associate 'revenge' with 'other'. Both of these possibilities are still consistent with the priming explanation.

Table 5.6

Differences in Number of Errors for Compatible and Incompatible IAT Blocks Across Counterbalanced Conditions.

Block	Cond	Mean	SD	*t*-test
Block 3 (compatible)	1	1.06	1.15	$t(110.17) = -2.56$, $p = .01$*, $d = .46$
(forgiveness-self)	2	1.68	1.51	
Block 4 (compatible)	1	3.38	2.51	$t(122) = -2.17$, $p = .03$*, $d = .39$
(forgiveness-self)	2	4.37	2.57	
Block 6 (incompatible)	1	2.92	2.06	$t(122) = -.56$, $p = .60$
(forgiveness-other)	2	3.12	2.07	
Block 7 (incompatible)	1	6.00	4.15	$t(122) = .68$, $p = .50$
(forgiveness-other)	2	5.52	3.76	

* *p<.05.* Conditions: 1 = IAT after attitude scales, 2 = IAT before attitude scales

An alternative way of investigating the contextual priming account is to examine the *correlations* between state (TRIM) forgiveness and performance on the four relevant IAT blocks. Results showed small but significant positive correlations between avoidance and revenge (*r*s of .20 and .24 respectively) with the number of errors made on the first forgiveness-other block (block 6), indicating that the more a person had forgiven a specific other the less difficult they found it to associate "forgiveness" with "other". This finding supports a salience asymmetry priming account in producing the unexpected relationship between the TRIM and forgiveness IAT. The TRIM was not related to performance on any of the other three blocks.

The IAT reaction time and error rate data suggest that IAT scores were primed by thinking about a personally-relevant transgression. Specifically, thinking about forgiveness as it related to an 'other' interfered with participants' abilities to associate forgiveness with the 'self'. This poses a problem for using the IAT to predict forgiveness of a real situation. However, there are several ways in which this problem may be overcome. One approach would be to revert to using an attitude (pleasant-unpleasant) rather than a self-concept (self-other) IAT. The proposed mechanism by which the IAT is being primed is through temporary changes in a person's ability to associate forgiveness with either 'self' or 'other'. An IAT that does not use these two category labels should theoretically be immune to this type of priming, as associations between forgiveness and pleasant/unpleasant should not be expected to vary in the same way between dispositional and situational levels.

A second, and equally valid, way to circumvent this priming effect would be to alter the presentation order of the measures. In the present study, all participants completed the measures relating to the real-life transgression *first*. That is, a forgiveness context – in which 'forgiveness-other' associations may have been primed – was always induced before participants completed the IAT. One simple way to overcome this would be to administer the IAT at the very beginning of the study, before completing other measures.

The remainder of this chapter will examine these two strategies. Study 7 will replicate the present study using an attitude, rather than self-concept, IAT. Study 8 will use the self-concept IAT, but present this *before* asking participants to reflect on an actual transgression. It is anticipated that each of these strategies should eliminate the negative

relationship found in the present study between the IAT and state-level forgiveness. In the absence of these unwanted effects, these studies should also serve as a more useful test of the IAT's ability to predict forgiveness of a real situation.

5.5 Study 7

5.5.1 Study Overview

The present study aimed to determine if the unexpected findings of Study 6 replicated when using an attitude (pleasant-unpleasant) IAT in place of the self-concept (self-other) IAT. Specifically it was hypothesised that if the negative relationship found between the forgiveness-revenge IAT and situation-specific forgiveness was caused by priming increased salience of 'forgiveness-other' associations, then the same effect should not be observed for an IAT that does not include a 'self-other' dimension.

5.5.2 Method

5.5.2.1 Design

This study utilised a single factor correlational design.

5.5.2.2 Participants

Participants were 75 (51 female, 24 male) first year undergraduate psychology students at the University of Adelaide, who participated in exchange for course credit. The mean age for the sample was 20.29 years (*SD*=4.33).

5.5.2.3 Materials

All measures were identical to those described for Study 6 (αs ranged from .66 to .91), with the exception of a modification to the attribute categories of the IAT. The self-other category pairing was replaced with pleasant-unpleasant, with these categories comprised of the same stimulus set used in Study 3. 'Pleasant' was represented by gorgeous, beautiful, fantastic, brilliant, marvelous, and magnificent. 'Unpleasant' was comprised of terrible, ugly, horrible, nasty, dreadful, and awful. Stimulus words for the forgiveness and revenge categories were identical to those used in Studies 3, 4, 5 and 6.

5.5.2.4 Procedure

The procedure for this study was identical to that of Study 6, but with two exceptions. The first was that the self-concept IAT used in Study 6 was replaced with an attitude IAT. The second was that there was no counterbalancing of presentation order. As the priming effect observed in Study 6 only occurred in the condition where the IAT came before the explicit attitude measures, this was the only condition of interest in determining if this effect would replicate with a different IAT variant. All participants first completed the details about the transgression that they had experienced, then the TRIM, followed by the IAT, and finishing with the explicit attitude questionnaire.

5.5.3 Results

5.5.3.1 Data preparation

IAT scores were computed using the revised D scoring algorithm (Greenwald et al., 2003) following the same procedures already outlined in the present work. This resulted

in 6 participants - for whom more than 10% of IAT trial responses were faster than 300ms - being excluded from the analysis (new N=69). Mean D score for the sample was .74 (*SD*=.35)

5.5.3.3 Replication of Study 6

Three separate hierarchical multiple regression analyses were conducted to determine if the negative relationship between the forgiveness-revenge self-concept (self-other) IAT and TRIM-measured forgiveness found in Study was replicable using forgiveness-revenge attitude (pleasant-unpleasant) IAT. As anticipated, this effect did not replicate: the IAT did not significantly predict revenge, avoidance or benevolence motivations after controlling for SDR and self-reported forgiveness attitudes. These analyses are summarised in Table 5.7.

5.5.3.4 Predicting forgiveness of a real transgression

Consistent with the findings of Study 6, after controlling for social desirability factors, explicit forgiveness attitude measures significantly predicted state-level forgiveness, accounting for 24% of variance in benevolence motivations, 18% of variance in avoidance motivations and 17% of variance in revenge motivations.

The IAT did not significantly explain any variance in forgiving behaviour.

Table 5.7

Hierarchical Multiple Regressions of Self-reported and IAT-measured Forgiveness Attitudes on Forgiving Response (Avoidance, Benevolence, Revenge), Controlling for Social Desirability.

	Avoidance			Revenge			Benevolence		
	β	R^2	ΔR^2	β	R^2	ΔR^2	β	R^2	ΔR^2
Step 1:									
BIDR-SDE	-.16			-.17			.05		
BIDR-IM	-.08			-.24^			.16		
		.04	.04		.12	.12*		.03	.03
Step 2:									
ATF	-.13			-.09			.21		
TTF	.06			.09			-.16		
HFS	-.39*			-.42*			.43*		
		.22	.18**		.29	.17**		.27	.24**
Step 3:									
IAT	.05	.22	.00	-.17	.31	.02	.01	.27	.00

* $p<.05$. ** $p<.01$. ^ $p=.07$.

5.6 Study 8

5.6.1 Study Overview

This study aimed to determine if the negative relationship found in Study 6 between the self-concept IAT and TRIM-measured forgiveness would replicate if the measures were completed in a different order. A plausible explanation for why this effect occurred is that reflecting on a transgression/transgressor primed an association between

'forgiveness' and 'other' for those who had forgiven more, which influenced subsequent performance on the self-concept IAT. If this explanation is true, then asking participants to complete the IAT *before* reflecting on a transgression should eliminate this effect. The present study tested this assumption.

5.6.2 Method

5.6.2.1 Design

This study utilised a single factor correlational design.

5.6.2.2 Participants

Participants were 70 (55 female, 15 male) first year undergraduate psychology students at the University of Adelaide, who participated in exchange for course credit. Mean age for the sample was 19.69 years (*SD*=4.10).

5.6.2.3 Materials

All measures were identical to those described for Study 6 (αs ranged from .67 to .96). This study used the forgiveness-revenge *self-concept* IAT.

5.6.2.4 Procedure

This study was a replication of Study 6, with the measures presented in an alternative order. All participants completed the IAT first, followed by explicit attitude measures. Following this they completed details about their interpersonal transgression,

and then the TRIM. Unlike in Study 6, there was no counter-balancing of presentation order.

5.6.3 Results

5.6.3.1 Data preparation

IAT D scores were, with two participants - for whom more than 10% of IAT trial responses were faster than 300ms - being excluded from the analysis (new N=68). Mean D score for the sample was .46 (*SD*=.25).

5.6.3.3 Replicating Study 6

Consistent with hypothesis, presenting the IAT before the transgression-related measures resulted in no significant negative association between the IAT and TRIM-measured forgiveness (see Table 5.8).

5.6.3.4 Predicting forgiving behaviour

In contrast to the previous two studies, explicit forgiveness attitudes only significantly predicted revenge motivations, but not avoidance or benevolence motivations. Consistent with studies 6 and 7, the IAT did not positively predict forgiveness of a real transgression.

Table 5.8

Hierarchical Multiple Regressions of Self-reported and IAT-measured Forgiveness Attitudes on Forgiving Response (Avoidance, Benevolence, Revenge), Controlling for Social Desirability.

	Avoidance			Revenge			Benevolence		
	β	R^2	ΔR^2	β	R^2	ΔR^2	β	R^2	ΔR^2
Step 1:									
BIDR-SDE	.13			.13			-.13		
BIDR-IM	-.15			-.28*			-.00		
		.02	.02		.06	.06		.02	.02
Step 2:									
ATF	.23			-.16			-.05		
TTF	-.26			-.19			.24		
HFS	-.06			-.16			.02		
		.10	.08		.24	.18**		.08	.06
Step 3:									
IAT	-.08	.10	.00	.04	.24	.00	.11	.09	.01

* $p<.05$. ** $p<.01$.

5.7 Discussion (Studies 7 and 8)

Results from Study 6 found that the forgiveness-revenge self-concept IAT significantly *negatively* predicted forgiveness of a real transgression. This finding was unexpected, but interpretable as an artefact of the self-concept IAT being primed by contextual factors. Studies 7 and 8 sought to determine if this negative relationship would replicate after manipulating two of the factors that were thought to have allowed

this priming to occur: IAT structure, and presentation order of forgiveness measures. Results confirmed that there was no negative relationship between the IAT and situation-specific forgiveness measures when an attitude IAT was used, nor when the IAT was presented before context-related measures. Both of these studies support the hypothesis that the negative relationship observed in Study 6 was a function of IAT priming.

However, neither Study 7 nor Study 8 found evidence for the predictive validity of the forgiveness-revenge IAT. Study 6 was initially developed to assess the IAT's ability to significantly *positively* predict forgiveness of a real transgression, but the unexpected priming effects made this aim difficult. Replicating this study while controlling for these priming concerns allowed the IAT's predictive validity to be tested. Contrary to hypothesis, the IAT did not significantly predict – either uniquely or incrementally – forgiveness of a real transgression, in either Study 7 or Study 8.

5.8 General Discussion

In Studies 6, 7, and 8 it was hypothesized that the forgiveness-revenge IAT would be significantly associated with forgiveness of an actual, recalled transgression. There was little support for this hypothesis across the three studies. Although the IAT's predictive validity may have been masked in Study 6 by contextual priming, the hypothesis was still unsupported in the two subsequent studies, in which this type of priming should not have been possible.

5.8.1 Malleability of the forgiveness self-concept IAT

The priming that occurred in Study 6 was unexpected, given the findings in the broader IAT literature that the order of implicit and explicit measures is generally unimportant (Hofmann et al., 2005a; Nosek, 2005). However, it is likely that priming occurred in Study 6 not because an explicit measure was completed before the IAT, but because of the specific directive to reflect on a real context (recalling the transgression). Although explicit measures have generally been found to not influence subsequent IAT performance, implicit measures have been shown to be malleable in other ways (for reviews see Blair, 2002; Gawronski & Bodenhausen, 2006). Implicit attitudes have been shown to be affected by context-dependent factors such as mood (DeSteno, Dasgupta, Bartlett, & Cajdric, 2004), antidepressants (Price, Nock, Charney, & Mathew, 2009) and even food deprivation (Seibt, Hafner, & Deutsch, 2007).

Of particular relevance to the present work are findings from several studies, which have shown that completing specifically-designed tasks, in which new associations are temporarily 'learned', can impact subsequent IAT performance. IAT priming can occur in response to relatively passive tasks such as viewing word pairs (Karpinski & Hilton, 2001), images (Dasgupta & Greenwald, 2001; van Quaquebeke & Schmerling, 2010), or videos (Wittenbrink, Judd, & Park, 2001), as well as in response to more active tasks such as playing violent video games (Bluemke, Friederich, & Zumbach, 2010; Uhlmann & Swanson, 2004). For example, Wittenbrink et al. (2001) had participants complete a race IAT, watch one of two short videos, and then again complete the race IAT (post-test). One of these videos depicted African-Americans who were members of a criminal gang, whereas the other video depicted African-Americans relaxing at a friendly barbecue. As

predicted, there was a significant difference between groups as a function of this manipulation, with those exposed to the positive-stereotype video recording a larger decrease in IAT effect (more pro-black attitudes) than the negative stereotype group.

The abovementioned studies all share a common mechanism for manipulating IAT performance: they increased the salience of a particular pair of categories, and subsequent IAT performance reflected associations that were consistent with this salient pairing. This mechanism may explain why reflecting on a real-life transgression makes a forgiveness-other association more salient (Study 6). However the present finding may add another element to our understanding of the malleability of the IAT. Of the previous work that has primed the IAT, the majority of studies have relied on participants 'learning' an association that is not necessarily their own. That is, participants have been primed with *extrapersonal* information. In the present work, priming occurred in response to thinking about a *personal* experience. Although not specifically designed to modify participants' associations, reflecting on a transgression seems to have served the same kind of priming function as a learning task. This has implications for IAT research in terms of the level of control that a researcher has over extraneous influences on IAT performance. Up until now, knowing that presenting people with material that 'taught' participants new associations could impact IAT performance also meant that not presenting this information should result in IAT performance that is not affected by context. However, results of Study 6 suggest that it may not be possible to control for the effects of context on the IAT, as the contextual priming information may be contained within – as well as outside – the person.

5.8.2 Incongruent priming of the IAT

The other unique feature of the priming that occurred in the Study 6 is that it resulted in *incongruent* IAT performance: the IAT was primed in an unexpected direction. In contrast, studies that have examined the malleability of the IAT have typically shown that the IAT can be primed in a direction that is *congruent* with explicit attitudes. The majority of these studies use attitude (pleasant-unpleasant) IATs and show that 'learning' more positive associations towards a particular group leads to a more positive endorsement of that group on an IAT measure (Dasgupta & Greenwald, 2001; Karpinski & Hilton, 2001; Wittenbrink et al., 2001). Similarly, the few studies which have demonstrated malleability of IATs with self-other category pairings have shown *congruent* priming: the preceding task increases the association between the target category and *self* (Baccus, Baldwin, & Packer, 2004; Bluemke et al., 2010; Uhlmann & Swanson, 2004). The present finding is unique because it is the first study to show that priming can increase the implicit association between the target category and *other*. As such, the priming is not just a result of context, but occurs due to an interaction between context and the IAT categories themselves.

Recent evidence suggests the interaction between context and categories exerts an important influence on implicit measures. Using a lexical decision task, Casper, Rothermund and Wentura (2010) asked participants to evaluate whether a series of targets were real words or not, with each target specifically chosen to reflect stereotypical information about both category and context. Each target word was presented over a background of four (2 x 2) possible category/context background combinations, comprised of either a compatible/incompatible category and

compatible/incompatible context. For example, the target word “emotional” appeared over a combination of one of the categories “women” (compatible) or “obese people” (incompatible) and one of the contexts “tissues” (compatible) or “golf ball” (incompatible). They hypothesised that the target would be sorted quickest (lowest reaction time) when both context and category were compatible. This interaction was significant, such that the effect of the nature of the categories (compatible/incompatible) was only significant when the target was presented in front of a compatible context background.

Although Casper et al. (2010) did not investigate the IAT specifically, their work still demonstrates that the interaction between category and contextual information may be an important predictor of performance on implicit, reaction time measures. However, they only showed an interaction between the context and *one* category dimension at a time. As the IAT is a double-dissociation task, it necessarily measures responses to *two* category pairs. The priming that occurred in the present study appears to represent a three way interaction between the target categories (forgiveness/revenge), the attribute categories (self-other) and a forgiveness-relevant context (forgiveness of an ‘other’). Further attempts to understand the mechanisms operating in *incongruent* priming of the IAT should address the relative impact of each of these three factors in influencing IAT scores.

5.8.3 A possible application: using the forgiveness-revenge IAT to measure forgiveness of a specific transgression

The forgiveness-revenge self-concept IAT's sensitivity to contextual information may have another important application. Specifically, although the IAT has thus far been used as an attitudinal/trait measure, it may have some utility as a measure of forgiveness that is more *situation-specific.* That is, the forgiveness IAT may be useful for assessing the extent to which forgiveness and revenge are currently mentally associated with a specific transgressor. However, further research is needed before the forgiveness-revenge IAT can be used in this way.

First, the findings from Study 6 need to be replicated. Second, effort must be put in to understanding the conditions under which the IAT can reliably assess context-specific associations. For example, the observed priming effect in the present study only occurred when the IAT was completed immediately after the contextual prime. The reason it did not occur in the other condition may have been because the effect had dissipated with time, or because completing explicit attitude measures acted as a counter-prime, or a combination of these factors. Similarly, the present data provides no information about whether the priming occurred as a result of merely thinking about a transgression, or if the TRIM and other transgression-related questionnaire items were also necessary to increase the salience of the forgiveness-other association. Further research is needed to address these questions, and to better understand the mechanisms through which the IAT can capture context-specific attitudes, preferences, or self-concept.

Despite this new possibility for the forgiveness-revenge IAT, exploring the potential of the IAT as a state-level measure of forgiveness is beyond the scope of the present work. Rather, the present work is focused on using the IAT as a trait-level measure of forgiveness. The findings of Study 6 suggest that the IAT was only sensitive to context when that context was explicitly primed *before* the IAT was completed. Based on this finding, it is recommended that further research using the IAT as a trait-level measure of forgiveness requires participants to complete the IAT *before* any tasks that relate to a specific forgiveness-relevant context.

5.8.4 The forgiveness-revenge IAT did not predict forgiveness motivations in response to a recalled offense

Even after contextual priming concerns were accounted for (studies 7 and 8), neither the forgiveness-revenge self-concept IAT nor attitude IATs significantly predicted forgiveness of a recalled transgression. In contrast, explicit forgiveness attitudes significantly predicted revenge motivations in all three studies, and significantly predicted avoidance and benevolence in two out of the three studies. There are several conclusions that could be drawn from these results. One conclusion is that the forgiveness-revenge IAT is simply not useful for predicting forgiveness of a real transgression. An alternative interpretation is that the forgiveness-revenge IAT was unable to predict forgiving behaviour because of the way in which this behaviour was operationalised: as self-reported responses to a (recalled) past transgression. There are, of course, several ways in which state-level forgiveness might be operationalised, including retrospective, hypothetical, and laboratory-based approaches (for a review of these refer back to

Chapter 1). Given the growing evidence that implicit and explicit measures may predict different *types* of behaviour (Asendorpf et al., 2002; Egloff & Schmukle, 2002; Perugini, 2005; Wilson et al., 2000), it is possible that the forgiveness-revenge IAT may predict forgiveness of a real situation when that forgiveness is measured in one of these alternative ways.

Specifically, dual attitude approaches to the relative predictive validity of implicit and explicit measures have demonstrated that explicit attitudes may be better predictors of deliberative, controlled, reflected behaviour, whereas implicit attitudes may better predict more impulsive, spontaneous, automatic behaviours (Asendorpf et al., 2002; Egloff & Schmukle, 2002; Perugini, 2005). For example, Perugini (2005) found that a fruits-snacks IAT significantly predicted participants' choice to take either a piece of fruit or a snack home with them, but did not predict self-reported consumption of either, whereas explicit fruit-snack preferences significantly predicted self-reported behaviour but not the actual behavioural choice. The retrospective approach taken to examine forgiving behaviour in the present study was arguably more *deliberative* than *spontaneous*, which may have explained why the IAT did not predict this behaviour.

The forgiveness-revenge IAT may be more suited to predicting behaviour that is less reflective, and more automatic. The retrospective approach taken in the present chapter did not allow for this hypothesis to be tested. Further research is required to determine if the IAT can predict forgiving behaviour that is more impulsive, automatic, and/or spontaneous.

Chapter 6:

Using the Forgiveness-Revenge IAT

to predict automatic forgiving behavior

6.1 Chapter Overview

The current chapter presents a study which attempts to assess the forgiveness-revenge IAT's utility in predicting more spontaneous types of forgiving behavior (as opposed to the deliberative processes that presumably underlie self-reported forgiveness of a recalled transgression). Automatic forgiveness responses will be assessed in two main ways. First, forgiving responses to brief hypothetical scenarios will be recorded, whilst restricting the cognitive resources that participants have available to deliberate on these responses. Second, an iterated trust game will be used to transgress against participants in real time, and forgiving responses will be assessed at both deliberative and spontaneous levels.

6.2 Introduction

6.2.1 Predictive models of implicit and explicit attitudes

In attempting to understand the way in which implicit and explicit measures differentially predict behaviour, Perugini (2005) summarised three predictive models: the additive, multiplicative/interactive, and double dissociation models.

The additive model is based on a representation of implicit and explicit measures tapping the same single attitude, but capturing different elements of it (see Fazio & Olsen, 2003). Based on this assumption, implicit and explicit measures should predict unique variance in behaviour. For example, additive approaches have been used in assessing implicit anxiety, with studies demonstrating that IAT-measured anxiety predicts unique variance in anxious behavior and task performance (e.g. giving an impromptu

speech) beyond that accounted for by explicit anxiety measures (Schnabel, Banse, & Asendorpf, 2006; Egloff & Schmukle, 2002).

The multiplicative/interactive model suggests that implicit and explicit measures interact to influence behavior. For example, Jordan, Spencer, Zanna, Hoshino-Browne, & Correll, (2003) demonstrated an interaction between implicit and explicit self-esteem in predicting defensive behavior. Specifically, for participants with high (but not low) explicit self-esteem, implicit self-esteem was significantly negatively related to behavioural measures of defensiveness (narcissism, in-group bias, and dissonance reduction).

The double dissociation model is based on dual attitude theory (Wilson et al., 2000) and posits that implicit and explicit measures predict different types of behaviour. Explicit responses are more controlled and so will therefore predict behaviour that is more controlled, deliberate and reflective. Implicit responses are not easily controlled, so will therefore better predict behavioural responses that are more automatic, reflexive, spontaneous, and for which there is little time for reflection or cognitive elaboration.

6.2.3 Explaining how implicit and explicit measures predict *different* types of behaviour: The double dissociation model

The additive and interactive models are both useful in explaining variance in a single type of behaviour. However, these models do not account for the fact that behaviour can take a variety of forms, occurs under a range of conditions, and can be measured in several different ways. The double dissociation model addresses these points, and has received particular recent empirical attention.

The double-dissociation model suggests that explicit measures will better predict behaviour that is more controlled/deliberative, whereas implicit measures will better predict behaviour that is more automatic/spontaneous (Fazio, 1990; Perugini, 2005; Wilson et al., 2000). What, then, constitutes an "automatic" or "spontaneous" behaviour? The MODE model (Fazio, 1990; Fazio & Olsen, 2003) suggests that automatic processes occur when there is either little *motivation* or *opportunity* to cognitively elaborate. Studies which have explored the predictive validity of the IAT within a double dissociation framework have tended to focus on the *opportunity* component of this explanation. Specifically, behaviour has been viewed as more automatic (and less deliberative) when an individual is thought to have less *control* over it. Studies have examined behaviours for which there should naturally be less controllability such as body language (Asendorpf et al., 2002; Egloff & Schmukle, 2002; McConnell & Leibold, 2001) and physiological response (Van Bockstaele et al., in press), as well as experimentally interfering with the cognitive resources that people have available to exert a controlled response (Friese et al., 2008; Friese, Wänke, & Plessner, 2006; Ranganath et al., 2008).

6.2.3.1 Non-verbal behaviours

Perhaps the most common method for assessing automatic behaviour in the IAT literature has been to examine how IATs and explicit measures differentially predict body language. In one of the first studies of its kind, McConnell and Leibold (2001) demonstrated that a race (black-white) preference IAT could predict non-verbal

behaviours, whereas explicit measures of racial prejudice did not[17]. At separate stages during the experiment, participants interacted with both a white and black confederate (both female). These interactions were videotaped and later rated on 16 different non-verbal behaviours, including eye contact, body positioning, and general levels of friendliness. Difference scores were calculated between the ratings for the two encounters. A pro-white IAT preference predicted more favourable non-verbal behaviour towards the white confederate on several of these indices, including speaking time (r=.51), speech errors (r=.42), smiling (r=.39) and hesitation (r=.35), whereas explicit measures of racial prejudice did not significantly predict any of these behaviours.

The relationship between the IAT and non-verbal behaviour has received particular attention in the domains of anxiety and shyness. Asendorpf et al. (2002) paired participants with a physically attractive (opposite sex) confederate and then coded non-verbal behaviour from videos of the interactions. Actions such as speech were classed as more controlled behaviour while non-verbal cues such as body tension were perceived as more spontaneous. They found that a shy (versus non-shy) self-concept IAT better predicted spontaneous behaviours whereas explicit shyness ratings better predicted more controlled behaviours. Similarly, Egloff and Schmukle (2002) examined nervous body language cues (e.g. mouth, eye and hand movements) during a stressful situation:

[17] There is an ongoing debate in the literature about the conclusions that can be drawn from this particular study. Blanton et al. (2009a) reanalysed McConnell and Leibold's (2001) original data and suggested that there was little evidence for their conclusion that the race IAT predicted spontaneous anti-black behaviour. McConnell and Leibold have contested this reinterpretation, with debate continuing (Blanton et al., 2009b; McConnell & Leibold, 2009).

an assessed oral presentation for which participants were given little time to prepare. An anxiety-calmness IAT significantly predicted nervous hand positioning/movements (*r*=.39) and global ratings of nervous behaviour (r=.38), whereas explicit trait anxiety did not predict any of the observed behaviours. In contrast, the IAT did not predict self-reported state-anxiety, whereas explicit trait anxiety did (r=.42).

6.2.3.2 Limiting cognitive resources

An alternative approach to examining automatic behaviour has been to experimentally limit the cognitive resources that participants have available to make a controlled, reflective decision. For example, Friese et al. (2008) showed that IAT-measured food preferences were more predictive of behaviour when cognitive or affective resources for making these behavioural decisions had been depleted. In one study, participants completed a battery of measures (implicit and explicit) which assessed their preference for chocolate relative to fruit. Following this, they were told to select five snack items from a table containing a mix of chocolates and fruits. Whilst making this choice, one group of participants was instructed to remember a one digit number (low cognitive interference), while the other group was told to remember an eight digit number (high cognitive interference). Results were consistent with the double dissociation model. When cognitive interference was low – providing greater opportunity to control behaviour – only the explicit food preference significantly predicted the actual foods chosen (*r*=.60). When cognitive interference was high – providing less opportunity to control behaviour – only the IAT was significantly predictive of food choice (*r*=.45).

Perhaps one of the simplest methods for limiting an individual's cognitive resources is to limit the amount of time that the person has to make a particular decision or complete a specific task. In another study using their food choice research paradigm, Friese and colleagues assessed whether the IAT and explicit measures would differentially predict food choice as a function of whether this choice was made under time restrictions (Friese et al., 2006). Participants completed an IAT and explicit measures that assessed preference for well-known relative to generic brands, and were then given the opportunity to select one of two combinations of food products to take home with them. One of these arrangements was comprised exclusively of well-known brands, and the other only contained generic branded products. In one condition, participants had unlimited time to make this decision, compared with a five second response window in the other condition. When implicit and explicit measures converged, explicit measures were highly predictive of behaviour, in both conditions of the time-constraint manipulation. However, when the two measures diverged, they differed in their prediction of behaviour across conditions. When participants had unlimited time to make their choice, 90% of participants chose the product combination that was consistent with their explicit preference. However, when forced to decide in less than 5 seconds, only 38% of choices were consistent with explicit preferences, with implicit preference being more strongly predictive.

Additional evidence that placing participants under time pressure can produce responses more consistent with implicit attitudes comes from a study on IE correspondence. For two preference domains (gay/straight and pop/jazz) Ranganath, Smith, and Nosek (2008) collected data on three types of measures: an IAT, standard self-

report, and a speeded/timed self-report measure. The speeded self-report measure required participants to respond to items on a four point scale, allowing seven seconds for each item. In both attitude domains, structural equation modelling revealed that the best model fit occurred when the speeded self-report loaded on the same factor as the IAT, with the standard explicit measure loading on a separate factor.

6.2.4 Applying the double dissociation model to forgiving behaviour

Together, the studies described above suggest that the double dissociation model is a useful tool – at least in some domains - for understanding the conditions under which the IAT may best predict behaviour. Perhaps forgiveness is another such domain. Despite forgiveness being largely viewed as a conscious, freely-chosen, and deliberative process (e.g. Enright et al., 1998; Worthington, 1998), there is recent evidence to suggest that at least some aspect of forgiveness may operate at the automatic level (Karremans & Aarts, 2007). Karremans and Aarts (2007) theorised that forgiveness is part of a "relational schema", which is evoked automatically when thinking of a close other. Specifically, they demonstrated that making participants think about a close (versus non-close) other led to a greater generalised tendency to forgive, and a greater likelihood of spontaneously generating forgiveness-related words on a word completion task. Moreover, these effects did not require the participant to be consciously aware that they had been primed to think about a close other. Thus, there is some evidence that forgiveness may be more automatic in the context of close relationships.

6.2.4.1 Measuring automatic forgiving behaviour

Conceptualisations of forgiveness as a largely conscious, deliberative process have impacted the ways in which forgiving behaviour has commonly been measured. Forgiveness research has typically taken two main approaches to examining behaviour as it relates to specific situations: retrospective/recall (e.g. McCullough et al., 1998; Subkoviak et al., 1995) and hypothetical scenario (e.g. Berry et al., 2001) approaches. Although both of these have proven useful in increasing understandings of forgiveness and its correlates, they are both reflective – rather than spontaneous – methods of assessing behaviour.

Studies using the recall approach have found trait forgiveness to be a reasonable predictor of forgiveness of a past transgression (Fehr et al., 2010; mean r=.30, 26 independent samples). Studies 6, 7 and 8 of the present work used this method and found that self-reported trait forgiveness – but not the forgiveness-revenge IAT - significantly predicted forgiveness of a recalled offense.

However, the double dissociation model suggests this finding should have been expected. The retrospective approach is by its very nature *reflective* and *controlled*. First, it specifically requests that participants *consciously* and *deliberatively* reflect on the situation and their behaviour. Second, the transgression always occurred in the past, thus any automatic or spontaneous reactions towards it would have already dissipated. Finally, it requires the self-reporting of behavior, which should capture more deliberative responses. These factors would suggest that retrospective approaches are not ideal for assessing the predictive validity of the forgiveness-revenge IAT.

An alternative to the retrospective approach is to gauge behavioural intentions towards a hypothetical transgression. Hypothetical scenarios overcome some of the temporal problems associated with recall approaches. Specifically, scenario approaches make it possible to gauge an individual's immediate (and more automatic) reaction to the transgression, compared with recall approaches that sometimes see participants reporting transgressions which took place months (or even years) ago. In regards to the depth of deliberation, it could be argued that hypothetical approaches should also evoke less deliberation about the transgression, as the individual is privy to significantly fewer details on which to deliberate, and is less able to ruminate over an offense that they have not actually experienced.

However, hypothetical approaches still measure forgiveness at a more deliberative than spontaneous level, as they still require consciously reported responses (i.e. imagining oneself in a situation and then making a conscious decision as to how one would act). Furthermore, there is some evidence to suggest that responses to hypothetical scenarios may be *more* considered than those relating to real-life events (Walker, Pitts, Hennig, & Matsuba, 1995; Wygant, 1997). Similarly, prospective transgressions may be evaluated more negatively than equivalent retrospective transgressions (Caruso, 2010). Finally, hypothetical approaches still require self-reporting of behavioural intentions.

The IAT literature suggests that more automatic behavioural responses can be elicited by manipulating the conditions under which these responses take place. Specifically, it has been shown that limiting an individual's cognitive resources can generate more automatic kinds of behaviours (Friese et al., 2008; Ranganath et al., 2008).

Placing these types of restrictions on cognitive resources may be a promising angle for examining more automatic forgiving behaviour. Specifically, it should be possible to use a hypothetical approach to assess participants' immediate forgiving responses to a transgression, while at the same time minimising the opportunity they have for deliberately reflecting on these responses. Study 9 will adopt this approach.

6.2.5 An alternative paradigm: The iterated trust game

Although it seems plausible that a time-limited hypothetical approach may be successful in eliciting more automatic behavioural forgiving responses, this method still has some shortcomings. Specifically, hypothetical scenarios operate at an abstract level: participants respond to a transgression that they have not actually experienced. Furthermore, there is evidence to suggest that scenario methodologies result in forgiveness being framed at a more cognitive (rather than affective) level than do methodologies focusing on real-life transgressions (Fehr et al., 2010). Each of these factors suggests that forgiving responses to hypothetical transgressions may lack a degree of psychological realism, which may influence the extent to which these findings are generalisable to the real world. Herein lies the predicament: hypothetical approaches gauge immediate reactions but lack psychological realism, whereas recall approaches possess this realism but do not allow for assessing immediate/automatic reactions.

Perhaps the most obvious way to assess forgiveness behaviour that is both (a) immediate, and (b) related to a personally-experienced transgression, is to actually transgress against participants in real time. There is some evidence that IATs may be more predictive than self-reports in response to real-time transgressions. Richetin,

Richardson, and Mason (2010) had participants complete an aggressive (harmful-harmless) self-concept IAT, after which an experimenter entered the room to provide feedback. In one condition this feedback was neutral, whereas in the other condition the experimenter was rude and insulting, claiming that the participants' slow performance had inconvenienced the experimenter. At the conclusion of the experiment, participants in both conditions were given a chance to evaluate the experimenter under the assumption that the evaluations could influence the experimenter's chances of keeping their job the following semester. The aggressive IAT significantly predicted evaluations reflecting punishment motivations following the transgression (β=.24), but was not a significant predictor of these evaluations in the neutral feedback condition. Thus, the presence of a real time transgression appeared to be the catalyst needed for an aggressiveness IAT to predict punishment behaviour. Perhaps the forgiveness-revenge IAT will be equally predictive in the context of a real transgression.

As noted in Chapter 1, attempts to measure forgiveness in response to real-time transgressions have been relatively scarce, probably owing to the difficulty in generating an appropriate research paradigm. Approaches such as that mentioned above (Richetin et al., 2010), as well as others used specifically to measure forgiveness (e.g. Zechmeister et al., 2004), are limited because the transgressor is an authority figure (experimenter/confederate) rather than a peer. This difference in status between victim/transgressor is a potential confound of the forgiveness process, given that status has a meaningful influence on the construal of a person's words and actions (Holtgraves, Srull, & Socall, 1989). Approaches that attempt to overcome this status differential are required to asses forgiving behaviour that is more naturalistic.

Perhaps the greatest difficulty in devising a suitable peer-on-peer transgression is achieving a delicate balance of severity. On the one hand, a researcher must be ethically responsible, ensuring that ultimately no harm is done to participants, and that the level of harm is justifiable by the perceived benefits of the research. Practically, this means that transgressions are effectively limited to causing participants some *discomfort*, rather than actual *harm*. On the other hand, this discomfort must be severe enough that forgiveness is actually relevant. One promising means for constructing a real time transgression is by using game theory. Researchers in economics and consumer psychology regularly use multiplayer computer-based money-making games such as the iterated prisoner's dilemma task (Axelrod, 1984) to assess how people respond in response to competitive or cooperative decisions by their opposing player. Prisoner's dilemma games have already found some application in forgiveness research, but this application has tended to address the experiencing of receiving – rather than granting – forgiveness (e.g. Struthers, Eaton, Shirvani et al., 2008; Wallace et al., 2008).

One game that may be especially useful for assessing forgiving behaviour is the 'trust game' (Berg, Dickhaut, & McCabe, 1995), which has found utility across a range of disciplines, including conflict management (e.g. Olekalns & Brett, 2008), business research (e.g. Tomlinson & Mayer, 2009) and psychology (e.g. Lount, Zhong, Sivanathan, & Murninghan, 2008; DeRue, Conlon, Moon, & Willaby, 2009). In the original version of the game there are two players, each positioned in separate rooms. The first player is given some money (e.g. $10), with which they can make two choices. The first is to keep the money, at which point the game ends. Alternatively, they can choose to "invest" some (or all) of this money by passing it to the other player, at which point that original sum of

money increases by a pre-determined factor (e.g. it is tripled, to $30). The second player can then decide what percentage they will return to the first player. Once player 2 has made this decision the game ends and each player keeps whatever money they are left with. The amount of money that player 1 is willing to pass to player 2 is interpreted as representing their level of trust.

The original 'one-shot' version of this game is limited because it provides no real incentive for either player to 'trust' the other. Player 2 knows that there will be no repercussions for not returning any money to player 1. Similarly, player 1 knows that they have no mechanism for ensuring that player 2 will return any money to them. For this reason, applications of the trust game have tended to use iterated (or repeated) versions of the game, in which there are several rounds (e.g. Güth, Ockenfels, & Wendel, 1997). In an iterated trust game, there *are* repercussions. Player 2 knows that failure to return a fair share of money back to player 1 will decrease the likelihood that player 1 will pass on any points to them in the following round. Similarly, player 1 knows that they can take a chance by initially passing points to the other player, and that they have the tools to punish that player for unfair behaviour.

The process that unfolds in the iterated trust game presents a significant opportunity for the study of forgiveness in real time. Interpersonal transgressions represent a betrayal of trust, with the trust literature focusing on the means by which trust can be restored following a transgression (Kim, Ferrin, Cooper, & Dirks, 2004; Korsgaard, Brodt, & Whitener, 2002). If the transgression may be seen as a *betrayal of trust*, then forgiveness may signal the *restoration of trust* that follows a transgression. Using trust recovery as a means for examining forgiveness is not a new idea. Trust has

been defined as an integral part of the forgiveness process (Rusbult et al., 2005; Veenstra, 1992)[18]. Moreover, one of the most common means for assessing transgression-specific forgiveness – the TRIM – uses trust as one indication that forgiveness has taken place (McCullough et al., 1998; 2006). Thus, post-transgression trust recovery may be a useful indicator of forgiveness.

6.2.6 The present research

This chapter presents a study which aimed to measure forgiving behaviour that is more automatic/spontaneous, and assess the utility of the forgiveness-revenge IAT in predicting this type of behaviour. This was achieved in one task by examining spontaneous forgiving behavioural decisions in response to a series of brief hypothetical scenarios. A second task utilised an iterated trust game to 'transgress' against participants in real-time in the laboratory, and then assessed spontaneous forgiving responses to the transgressor.

In both tasks, 'spontaneous' forgiving behaviour was defined as decisions that took place when cognitive resources for making such decisions were limited. Specifically, limitations were placed on resources by imposing time limits on behavioural decisions, following similar procedures to those used by Friese et al. (2006) and Ranganath et al.

[18] An alternative perspective is offered by those theorists who believe that there is a clear distinction between forgiveness and reconciliation (e.g. Enright & Zell, 1989; Fincham, 2000). For these theorists, trust is a necessary component of reconciliation, but forgiveness can occur without trust restoration. In the present work, it is argued that the distinction between forgiveness and reconciliation is trivial, and that trust is an important component of the forgiveness process.

(2008). In the first task, participants were given a seven second window to respond (on a computer) to a series of yes/no type forgiveness decisions, ranging from the relatively benign (e.g. "Whilst driving, someone cuts you off. Beep your horn at them?") to the more severe (e.g. "Your partner cheats on you. Forgive them?"). In the second task, the same seven second response window was applied to a series of reward/punishment (benevolence/revenge) decisions that were directed at a partner who had just transgressed against the participant (in the trust game). Specifically, this task allowed participants to modify (both positively and negatively) the amount of points that the transgressor could keep, which would potentially impact on whether or not they would receive a $10 prize.

Consistent with the double-dissociation model of explicit-implicit interaction, it was anticipated that IAT-measured forgiveness would fare better than explicit measures of forgiveness attitudes in predicting spontaneous, time-pressured forgiving behaviour.

In response to the brief hypothetical scenarios, it was hypothesised that:

i. The forgiveness-revenge self-concept IAT would better predict spontaneous forgiving behaviour than would self-reported forgiveness attitudes

Responses to the iterated trust game transgression were measured at both the automatic (time-pressured) and controlled (self-reported) levels. It was hypothesized that:

i. The forgiveness-revenge self-concept IAT would better predict spontaneous forgiving behaviour (time-pressured reward/punishment decisions) than would self-reported forgiveness attitudes

ii. Self-reported forgiveness attitudes would better predict controlled (self-reported) forgiving behaviour than would the forgiveness-revenge self-concept IAT

The rationale for using the iterated trust game to examine forgiveness relied on the premise that a transgression represents a *betrayal* of trust. As such, levels of self-reported post-game trust towards the transgressor should also provide an indication of the extent to which forgiveness had taken place. It was hypothesised that:

iii. The forgiveness-revenge self-concept IAT would significantly predict trust

iv. Self-reported forgiveness attitudes would significantly predict trust

v. Given that trust was self-reported (with no time restriction), it was predicted that explicit forgiveness attitude measures would be a greater predictor of trust than would the forgiveness-revenge IAT.

6.3 Study 9

6.3.1 Method

6.3.1.1 Design

The study adopted a single factor correlational design.

6.3.1.2 Participants

Participants were recruited for four ostensibly unrelated experiments on a diverse range of topics, for a series of studies entitled "Responses to Conflict, Decision Making Under Time Pressure, and Cooperation Using an Investment Game". Recruitment of participants occurred through poster advertisements at two South Australian university

campuses, with participants compensated for their time with a payment of $10. The sample was comprised of 86 participants (56 female, 30 male), with a mean age of 22.91 (*SD*=5.45).

6.3.1.3 Materials

6.3.1.3.1 The IAT

The forgiveness-revenge self-concept IAT was used. The structure and sequence of this measure was identical to that used in studies 4, 5, 6 and 8.

6.3.1.3.2 Explicit Attitude Measures

Explicit measures of forgiveness attitudes and socially desirable responding were identical to those used in studies 6, 7 and 8, and included the ATF (α=.50) and TTF (α=.53; Brown, 2003), the HFS 'other' subscale (α=.60; Thompson et al., 2005), two markline items used to construct a double-dissociation difference measure of forgiveness relative to revenge, and the BIDR (SDE, α=.71; IM, α=.73; Paulhus, 1986). The questionnaire concluded with demographic items assessing age and gender.

Internal consistency reliabilities for the ATF and TTF were poor. Deleting two (of six) items from the ATF improved reliability to α=.61. Reliability for the TTF was not improved by deleting items.

6.3.1.3.3 Forgiveness go/no-go task

Automatic/spontaneous forgiveness behaviour was measured using a go/no-go procedure. On a computer screen, participants were presented with a series of brief hypothetical decisions such as "Your partner cheats on you. Forgive them?" or "A classmate spreads some nasty rumours about you. Get even?". For each decision task, the participant was allowed 7 seconds to respond[19]. If the participant agreed with the statement and believed that they would take that action, they were instructed to click their mouse on a button marked "continue" (a "go" response). If they disagreed with the statement and did not want to take that action, they were instructed to wait for 7 seconds to pass (a "no-go" response). The participant was then presented with the next decision task.

In total, the task consisted of 34 decisions, 9 of which were relevant to forgiveness. The remaining 25 were distracter items, designed to conceal the true nature of the task, and consisted of such items as "You have an assignment due in 4 weeks. Put it off until the night before?", "You have a choice of beer or soft drink. Have a beer?", and "A second-hand textbook is $20 cheaper than the new one. Buy it second-hand?". These items also functioned as practice tasks, with five of these appearing before the first forgiveness-relevant item. None of the distracter items were used in the analyses. The nine forgiveness-relevant scenarios/decision tasks were designed to represent a range of

[19] Pilot testing (N=10) indicated that an initial response window of 5 seconds was too quick for some participants to accurately perform the task. The response window was therefore modified to 7 seconds, which was enough time for the majority of participants to accurately complete most of the tasks.

situations in which forgiveness may be relevant, differing in severity and closeness to offender. These items are presented below:

- *"You catch your partner reading through your private journal. Forgive them?"*
- *"A classmate spreads some nasty rumours about you. Get even?"*
- *"Your mum forgets your birthday. Forgive?"*
- *"You break a promise you had made to your best friend. Apologise?"*
- *"Whilst driving, someone cuts you off. Beep your horn at them?"*
- *"A good friend doesn't invite you to their birthday party. Will you invite them to yours?"*
- *"Your partner cheats on you. Forgive them?"*
- *"A friend copies your assignment and claims it as their own. Expose them?"*
- *"A co-worker bullies you. Avoid him/her?"*

Items were reverse-scored as required. These go/no-go decisions provided two types of information: (1) Whether or not the participant actually made the decision, and (b) How *quickly* they made the decision. Each of these pieces of information was treated separately in the analyses.

6.3.1.3.4 Iterated Trust Game

The iterated trust game used in the present study was structurally similar to the one-shot trust game developed by Berg, Dickhaut and McCabe (1995), and the repeated/iterated cousin of this same game (see Güth et al., 1997). However, it differed from the standard game in two main ways: (1) rather than being paired with another

human player, participants played against a computer-programmed set of responses (although they were led to believe that it was a real person)[20], and (2) there was opportunity for communication between players, taking the form of typed messages[21]. These modifications were made in order to create a plausible transgression that was standard across participants.

Participants were first given instructions informing them that they would be playing a computer-based investment game in which they would be paired with a student in another lab, however in reality the student was playing against a pre-programmed computer opponent. An instruction page informed the participant that if a player achieved a certain score by the end of the study (a score equal or greater to 100 points) then that player would receive a bonus payment of $10 (in addition to the $10 payment that they were already receiving for participating in the study). Participants were also informed that the best strategy for achieving this score was by cooperating with the other player. After reading thorough instructions and being shown demonstrations of how the investment game works, participants were informed that they had been randomly allocated to the role of "investor", and that their partner had been allocated the role of

[20] Similar to the approach recently adopted in research on trust recovery (Haselhuhn et al., 2010; Schweitzer et al., 2006)

[21] Haselhuhn, Schweitzer and Wood (2010) also adopted similar modifications to the iterated trust game. However, these authors utilised the messaging function to *apologise* following untrustworthy behaviour. In contrast, the present study uses a messaging function to *compound* a transgression, rather than apologise for it.

"stockbroker". In reality, all participants completed the game in the "investor" role, and the "stockbroker" role was a set of pre-programmed responses.

Participants were informed that the game would be comprised of a random number of rounds, ranging between 1 and 12. In actuality, the game always lasted for fice rounds. At the start of each round the participant (investor) was allocated 10 points, with which they could choose to do one of the following:

(1) Keep all points

(2) Split points evenly

(3) Invest all points

If they chose option (1) 10 points were added to their score and the other player received nothing. If they chose option (2) 5 points were added to both players' scores. In each case the other player (computer) did not receive a turn and the next round began. If the player selected option (3) then the points were quadrupled (i.e. there were now 40 points) and the turn moved to the other player (stockbroker), who could now choose to keep them all or return some (options ranged 5 to 40 points, presented in increments of 5).

Participants were encouraged to select option (3) through two incentives. The first was by a perceived monetary incentive – participants were told that if they achieved a score of 100 then they would qualify for the bonus $10 cash prize. It was explicit to participants that the best strategy for achieving the most points was to select option (3) and hope that the other player returned half of the points (20) back to them. The second incentive was through an in-game messaging function designed to build trust that if an investor did in fact invest all the points, that the other player would actually return some

to them. Before the game began the participant was informed that each player now had the opportunity to send the other player a message, and was then prompted to do so. Regardless of whether or not the participant chose to send a message, they were informed that the stockbroker had sent them the following message “hi! If u invest I will split points with u ok?”.

The investment game consisted of five rounds, and the first four rounds were spent “building trust”. That is, if the participant chose to invest all points in those rounds, the stockbroker would always return half to the participant. Thus, if the participant always chose to invest all points, then at the start of round 5 (final round) both players would each be on scores of 80 points. If both players continued with the same response pattern (participant invests all and stockbroker returns half) then they would both reach the 100 point threshold, guaranteeing them both the $10 prize. This is where the “transgression” took place. In the final (fifth) round, if the participant invested all points, the stockbroker (computer) would return nothing, leaving the participant with a score of 80 and the stockbroker with 120. This concluded the game.

Following the completion of the trust game, another opportunity to exchange messages was presented to participants. Irrespective of whether the participant chose to take this opportunity they received a message from the stockbroker that read “ha ha ur screwed!”. This message was designed to stop participants from justifying or rationalizing the other player’s behaviour – the message makes it very clear that other player betrayed their trust and that his or her actions were unjustified.

6.3.1.3.5 Spontaneous forgiving behaviour

The dependent measure of 'automatic' forgiveness was again a go/no-go type task. Immediately following the investment game, participants were informed that they would be participating in another game where they would have the chance to influence both their partner's and their own score. Instructions to participants explained that this task would follow the same procedure as the decision making under time pressure (go/-no-go) task they undertook earlier: they would be presented with a series of options (one at a time), and to select an option they click the "continue" button, or to not take an option they wait for 7 seconds. This time, however, the options presented allowed for modifying either player's score by adding/subtracting a set amount of points. For example, the participant was presented with options such as "Add 5 points to Player 1's score", "Deduct 10 points from Player 2's score", and "Double Player 2's points". These decisions were designed to act as a proxy for the extent to which the participant had forgiven the other player at the automatic level. Adding points to the opponent's score was interpreted as *benevolent* behaviour, and deducting points represented *vengeful* behaviour. Items allowing for the modification of one's own score were included in an attempt to mask the true nature of the task. The complete list of options presented are listed below:

"Add 5 points to Player 2's score"

"Add 5 points to Player 1's (your) score"

"Deduct 5 points from Player 2's score"

"Deduct 5 points from Player 1's (your) score"

"Add 3 points to Player 1's (your) score"

"Add 10 points to Player 2's score"

"Deduct 50 points from Player 2's score"

"Deduct 50 points from Player 1's (your) score"

"Add 50 points to Player 2's score"

"Deduct ALL of Player 2's points"

6.3.1.3.6 Deliberative forgiving behaviour

Deliberative forgiveness behaviour was assessed in two ways: self-rated forgiveness and self-rated trust. Upon completion of the investment game, participants were asked to indicate whether they believed that their partner had "done the right thing by (them)" in the game, on a 5 point Likert-type scale. Participants who reported scores less than 3 (indicating disagreement with the item) were asked several follow-up questions, all rated on 5 point scales with 5 signalling the greatest agreement. The first of these items assessed forgiveness directly: *"I would be willing to forgive my team mate for what they did during this study"*

In addition to directly assessing forgiveness, four items were included to assess levels of *trust* towards the partner:

- *"I would be comfortable having my team mate work on a task or problem that was important to me, even if I could not monitor their actions"*
- *"I would be comfortable having my team mate make decisions that critically affect me"*
- *"I would keep an eye on my team mate"*

- *"If I had my way, I would not let my team mate have any influence over issues that were important to me"*

6.3.1.4 Procedure

The study took place in a computer laboratory, with participants sitting in separate corralled cubicles. All tasks were completed on a standard PC. Upon arrival, participants were provided with a cover story, being informed that they would be participating in a series of unrelated studies which assessed the broad themes of speed of thought and decision-making, cooperation when communication is limited, and general social attitudes. They were also informed – in passing – that half of the participants for the study had been sent to a different room. Verbal instructions were provided asking participants to let the experimenter know after they had completed each task, and that he would then set up the next study for them. The verbal instructions also informed participants that because one of studies was interested in communication when opportunity for communication was limited, they would be paired with someone in the other room for that particular study. Once participants were ready to begin, the experimenter made a bogus phone call to a fictitious experimenter in the "other room", to "check" that those participants were also ready to begin. Participants were then instructed to begin.

The Forgiveness-Revenge IAT was completed first. In order to not arouse suspicion that the study was about forgiveness, a cover story was provided to participants that they were about to complete a timed-word sorting task, designed to assess category sorting ability under time pressure. The instructions stated that participants would be

randomly allocated to one of five different IATs, containing one of the following category pairs: (1) Male/Female, (2) Forgiveness/Revenge, (3) Arrogance/Humility, (4) Calm/Panic, (5) Effort/Laziness. In actuality, all participants completed the Forgiveness-Revenge IAT.

Following the IAT, participants completed the first go/no-go task (i.e. the hypothetical scenarios). As participants finished the go/no-go task, they were asked to wait patiently while the experimenter checked that one of the participants in the other room had completed the first two tasks and was ready for the investment game. After another bogus phone call – in which the experimenter pretended to inform the confederate of the participant's computer number – the participant was instructed that they could now begin the study. This task comprised the investment game, followed by the second go/no-go task, and self-reported forgiveness and trust of the partner.

The final component of the study was the questionnaire battery, assessing explicit forgiveness attitudes, socially desirable responding, and demographic information. Following completion of the questionnaire, participants were then immediately debriefed about the true nature of the study, and paid for their time.

6.3.2 Results

6.3.2.1 Data preparation

After computing D scores, four participants - for whom more than 10% of IAT trial responses were faster than 300ms - were excluded from the analysis (new N=82). Mean D score for the sample was .58 (*SD*=.30).

6.3.2.3 Predicting (time-pressured) forgiveness of brief hypothetical scenarios

It was hypothesized that the forgiveness-revenge self-concept IAT would better predict time-pressured forgiving behaviour than would self-reported forgiveness attitudes, in response to the brief hypothetical scenarios. To investigate this hypothesis, correlations were calculated between the predictor variables, SDR, and each of the nine forgiveness-relevant responses on the go/no-go task (Table 6.1). The pattern of correlations suggests that the hypothesis was not supported. The forgiveness-revenge IAT was not significantly related to any of the go/no-go behavioural responses. In contrast, there were significant correlations between some of these behaviours and the explicit forgiveness attitude measures. The ATF, TTF and HFS were all significantly related to what was perhaps the most benign and automatic behavioural decision – whether to beep your horn at a car that had "cut you off" in traffic. Likewise, all three of these scales were related to a decision about whether or not to forgive a cheating partner.

To further investigate the hypotheses, two multiple regression analyses were performed. For one of these, the nine go/no-go forgiveness responses were combined to create an index of how automatically forgiving participants were across the nine situations. After reverse-coding responses to items 6 through 9, the items were summed to create a score ranging from 0 to 9 (actual range 3 – 9).

Table 6.1

Intercorrelations between IAT D Scores, Forgiveness Attitude Scales, and State-level Forgiveness

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. IAT															
2. ATF	.03														
3. TTF	.04	.39***													
4. HFS	.32**	.33**	.46***												
5. F-R difference	.05	.09	.12	.15											
6. BIDR-SDE	.14	.05	.26*	.28*	.14										
7. BIDR-IM	.02	.11	.24*	.19	.02	.49***									
8. G/NG1: Partner reads journal, forgive?	-.02	.04	.09	.10	.11	-.05	-.00								
9. G/NG2: Mum forgets birthday, forgive?	-.09	.17	.20	-.01	.17	.20	.14	-.00							
10. G/NG3: Partner cheats on you, forgive?	.10	.29**	.25*	.21^	.11	.27*	.29**	.31**	.17						
11. G/NG4: Friend doesn't invite you, invite them?	-.14	.12	.10	.14	.12	.04	.11	.20^	.22*	.14					
12. G/NG5: You break a promise, apologise?	.01	.19	.13	.01	-.06	.05	.02	.08	.26*	.08	.02				
13. G/NG6: Co-worker bullies you, avoid?	-.19	.05	.06	-.14	-.03	-.06	.04	-.09	.13	-.06	.08	.20			
14. G/NG7: Classmate spreads rumours, avenge?	-.16	-.26*	-.21^	-.03	-.18	-.17	-.06	-.04	.03	-.21^	-.13	.07	.01		
15. G/NG8: Car cuts you off, beep horn?	.06	-.22*	-.32**	-.25*	-.09	-.18	-.17	.15	.16	-.13	-.08	-.05	-.16	.26*	
16. G/NG9: Friend copies assignment, expose?	-10	-19	-.11	-.21^	-.06	.02	.07	-.07	-.01	-.06	-.01	.18	.04	.21^	-.02

* $p<.05$. ** $p<.01$. *** $p<.001$. ^ $p<.07$.

G/NG = go/no-go response. Each of these items consists of a two-level categorical response, such that "go" responses are recorded as a value of 1 and "no-go" responses are coded as 0.

The second regression analysis examined *speed* of response. In addition to whether or not participants chose to forgive, each go/no-go item recorded information about how quickly they reached this decision. For those choosing a "go" response, a reaction time ranging (theoretically[22]) from 0 to 6999 milliseconds was recorded, with a "no-go" response recording a default of 7000 milliseconds. A second index was created using this reaction time data. After reverse-coding appropriate items, scores were summed to produce a total score with a theoretical range from 0 to 63000 (actual range 21840 – 50480), with lower scores indicating that less time was taken to make a forgiving response. For ease of interpretation, this score was then reverse-coded so that higher scores represented greater forgiveness. The two indices were then regressed on explicit and implicit trait forgiveness measures, after controlling for SDR. The two subscales of the BIDR were entered at the first step, with the ATF, TTF, HFS and IAT entered simultaneously at step 2. The results of these analyses are presented in Table 6.2 (below).

[22] In actuality, at least 1000ms would have been reasonably required to read and process the item before responding. Reaction times faster than 1000ms were deemed to be due to participant error, and were treated as missing values in the analyses.

Table 6.2

Hierarchical Multiple Regressions of Self-reported and IAT-measured Forgiveness Attitudes on Forgiving Response (Indices of Go/no-go Responses), Controlling for Social Desirability

	Go/no-go Index			Go/no-go Reaction Time Index		
	B	R^2	ΔR^2	β	R^2	ΔR^2
Step 1:						
BIDR-SDE	.19			.17		
BIDR-IM	.09			.01		
		.06	.06		.03	.03
Step 2:						
ATF	.26*			.32**		
TTF	.16			.12		
HFS	.12			.14		
IAT	.03			-.04		
		.17	.17**		.22	.19**

* p<.05. ** p<.01.

Results from the regression do not provide any support for the hypothesis that the IAT would better predict time-pressured responses to forgiveness scenarios. The forgiveness-revenge IAT did not significantly predict any variance in forgiving responses. In contrast, one of the explicit measures – the ATF – did predict these responses, in terms of both *whether* participants chose to behave in a particular way, and how quickly they were able to arrive at this decision.

6.3.2.4 Predicting Forgiveness in response to the iterated trust game transgression

6.3.2.4.1 Manipulation checks and data exclusion

Of the 82 participants who completed the study, 67 chose to invest all of their points on all five rounds of the trust game. Data for the remaining 15 participants were excluded from the analyses, as not choosing to invest resulted in there being no opportunity for these participants to experience a transgression, which in turn meant that behavioural measures of forgiveness were redundant. An additional five participants indicated during study debriefing that they were confident that they had been playing the game against a computer opponent, rather than another student. These participants were also excluded from the analyses (remaining N=62).

6.3.2.4.2 Predicting automatic forgiving responses

It was hypothesised that the forgiveness-revenge IAT would be a better predictor of automatic forgiving responses than would explicit forgiveness attitude measures. Automatic forgiveness was operationalised as the time-pressured behavioural decisions directed at the other player following the transgression. This behaviour took two forms: benevolence and revenge. Benevolence was measured with three options that allowed participants to *add* points (5, 10, 50) to their opponent's score. Revenge was measured with three options allowing participants to *deduct* points (5, 50, all) from their opponent's score.

Correlations suggest that there may be some limited support for the hypothesis. The forgiveness-revenge IAT was significantly related to two of the three *benevolence*

responses, with a greater implicit preference for forgiveness related to an increased chance of adding 5 points (r=.34) and 10 points (r=.30) to the transgressor's score. In contrast, none of the explicit forgiveness measures were significantly related to the three behavioural measures of benevolence.

IAT-measured and self-reported forgiveness preferences were both largely unrelated to responses to the three behavioural measures of *revenge.* The only exception was a significant negative correlation between the HFS and the decision to deduct 5 points from the other player's score (r=-.30), with those who self-reported as more forgiving less likely to make this choice. This decision, however, was unrelated to the three other explicit forgiveness measures.

Four hierarchical multiple regressions were conducted to further examine the roles of implicit and explicit forgiveness preference in predicting automatic benevolent and vengeful behaviour. These analyses (separately) regressed two indices of benevolent behaviour, and two indices of vengeful behaviour, on implicit and explicit trait forgiveness (after controlling for SDR). Benevolence and revenge were each represented by an index of *whether* or not the participant chose that option, as well as an index of *how quickly* they made the decision. Again, reaction time indices were reverse-coded for ease of interpretation, with higher scores representing a faster endorsement of revenge/benevolent behaviour. The two subscales of the BIDR were entered in the first step of each model, with the ATF, TTF, HFS and IAT all entered as simultaneous predictors at step 2. Results are presented in Table 6.3 (below).

Table 6.3

Hierarchical Multiple Regressions of Self-reported and IAT-measured Forgiveness Attitudes as Predictors of Indices of Benevolence and Revenge, Controlling for Social Desirability

	Benevolence (go/no-go)			Benevolence (reaction time)			Revenge (go/no-go)			Revenge (reaction time)		
	B	R^2	ΔR^2	β	R^2	ΔR^2	β	R^2	ΔR^2	β	R^2	ΔR^2
Step 1:												
BIDR-SDE	.12			.12			-.19			-.22		
BIDR-IM	-.21			-.27			.13			.11		
		.04	.04		.06	.06		.03	.03		.04	.04
Step 2:												
ATF	-.07			-.07			-.04			-.06		
TTF	.23			.22			.15			.12		
HFS	-.04			-.01			-.33*			-.31*		
IAT	.29*			.25^			.14			.12		
		.15	.12		.16	.10		.12	.09		.13	.08

* $p<.05$. ** $p<.01$. ^ $p<.07$.

Results of the regression analyses suggest that the hypothesis – that the IAT would be a better predictor of automatic forgiving behaviour than would explicit forgiveness scales – was partially supported. The IAT was the only significant predictor of overall benevolence, accounting for 8.4% of the variance in go/no-go benevolent response. Furthermore, the IAT marginally predicted 6.3% of variance in the speed at which participants reached their benevolent decision, with greater IAT-measured forgiveness preference predicting less hesitation in acting benevolently towards the transgressor (p=.069) .

In contrast, the IAT did not significantly predict automatic vengeful responses, whereas one of the explicit forgiveness scales did. The HFS significantly accounted for 10.9% of variance in go/no-go revenge responses, such that those who self-reported as more forgiving were significantly less vengeful towards the transgressor. The same relationship was mirrored in the reaction time data, with the HFS accounting for 9.6% of variance in the speed at which participants reached a decision to punish: greater forgivers were slower to punish.

6.3.2.4.3 Predicting controlled forgiving responses

In contrast to automatic forgiving behaviour, it was hypothesised that controlled forgiving behaviour would be better predicted by explicit forgiveness scales, rather than the IAT. Controlled forgiving behaviour was assessed in two ways: (1) an item directly asking how much the participant would be willing to *forgive* their partner, and (2) a scale comprising four items assessing *trust* in the partner.

An initial examination of correlations suggests that there was little support for the hypotheses in relation to either forgiveness or trust. None of the attitude/trait forgiveness measures – implicit or explicit – significantly predicted the direct measure of forgiveness, despite a marginal effect for the ATF (r=.25, p=.063). Additionally – and contrary to hypothesis – the IAT was the only predictor variable that was significantly related to trust (r=.26).

To examine the hypotheses further, forgiveness and trust were both regressed on the implicit and explicit forgiveness measures, after controlling for SDR. The results of these two hierarchical regressions are presented in Table 6.5. (below). The two subscales

of the BIDR were entered at step 1, with the predictor variables – ATF, TTF, HFS, and IAT – entered together at step 2.

Table 6.4

Hierarchical Multiple Regressions of Self-reported and IAT-measured Forgiveness Attitudes on Forgiving Response (Indices of Go/no-go Responses), Controlling for Social Desirability

	Forgiveness			Trust		
	B	R^2	ΔR^2	β	R^2	ΔR^2
Step 1:						
BIDR-SDE	-.04			-.23^		
BIDR-IM	-.13			.03		
		.03	.03		.05	.05
Step 2:						
ATF	.25			.07		
TTF	.13			.06		
HFS	-.24			-.16		
IAT	.26^			.31**		
		.16	.14		.14	.09

* $p<.05$. ** $p<.01$. ^ $p<.07$.

The results of the regression analyses reveal that the IAT was the only significant predictor of controlled forgiving responses. After controlling for SDR, the IAT significantly predicted trust towards the transgressor, accounting for 9.6% of the variance in this relationship. There was also a marginal effect for the IAT on forgiveness, accounting for 6.7% of variance (p=.063). In contrast, none of the ATF, TTF or HFS were significant

predictors of either forgiveness (*p*s > .08) or trust (*p*s > .20). The hypothesis was not supported.

6.4 Discussion

Study 9 was designed to assess the utility of the forgiveness-revenge IAT in predicting behaviour that was more spontaneous and automatic. There was only minor support for the hypotheses. Time-pressured responses to hypothetical transgressions – contrary to prediction – were not predicted by the forgiveness-revenge IAT. Responses to an iterated trust game transgression suggested that the IAT may better predict benevolent – but not vengeful – behaviour following a transgression. However, the forgiveness-revenge IAT was also unexpectedly more effective than explicit forgiveness measures in predicting levels of deliberative post-transgression trust.

6.4.1 Predicting automatic forgiving responses to hypothetical transgressions

There was no support for the hypothesis that the forgiveness-revenge IAT – rather than explicit forgiveness attitude measures – would better predict automatic forgiving behavioural responses to hypothetical transgressions. Based on recent work with the IAT (Friese et al., 2006; Ranganath et al., 2008), it was anticipated that placing time pressure on participants' decisions should have elicited responses that reflected more automatic behaviour, and that the IAT should have predicted this type of behaviour. However, it may have been that the approach taken – time-pressured responding to a hypothetical transgression – resulted in responses that were still more controlled than they were automatic. That is, it did not adequately discourage effortful processing.

The methods used by Friese and colleagues (2006, 2008) – such as choosing snack foods – were successful because they required relatively little effortful processing. In contrast, responding to a hypothetical scenario (regardless of time pressure) arguably requires *more* – not less – cognitive effort. If an individual has actually experienced a transgression, then all they need to do is know how they feel about it. Responding to a hypothetical scenario, on the other hand, is an abstract mental exercise, requiring one to use their reasoning abilities to make an (arguably) objective decision. In fact, there is evidence to suggest that hypothetical scenarios can evoke more thoughtful and considered responses than those concerned with real-world problems (Wygant, 1997). Thus it is possible that the null relationship between the forgiveness-revenge IAT and go/no-go responses to hypothetical forgiveness scenarios occurred because the scenario methodology – even under time constraint – was not automatic *enough*.

6.4.2 Predicting automatic forgiving responses to a real transgression

In contrast to the responses to hypothetical forgiveness scenarios, those made in relation to the trust game betrayal provided at least some support for the hypothesis: that the forgiveness-revenge IAT (rather than forgiveness attitude scales) would better predict automatic forgiving behaviour. As hypothesised, the IAT significantly predicted time-pressured decisions to add points to a transgressor's score, in terms of both whether to make these decisions, and how quickly they were made. In contrast, none of the explicit forgiveness attitude measures significantly predicted these decisions. However, this finding was only limited to one of the two indicators of forgiveness: benevolence. Time-pressured decisions about deducting points from the transgressor were not

predicted by the IAT. However, one of the explicit measures – the HFS – did significantly predict these vengeful responses.

Why did the IAT and explicit attitude measures differentially predict benevolent and vengeful behaviour? One possibility is that explicit reasoning was constrained by logic, whereas automatic reasoning was not. Before beginning the go/no-go reward/punishment task, participants were already consciously aware that the other player had a score of 120 points: 20 more than what was needed to receive the $10 prize. Thus, rationally, choosing to add points would have had no meaningful impact on outcomes for the other player. In contrast, if the participant chose to *deduct* varying amounts of points, then there was a reasonable expectation that the other player's score would fall below 100, therefore depriving them of the $10 prize. At a rational level, revenge-orientated decisions were efficacious, whereas benevolence-orientated decisions were not.

Practically, this may have meant that a large proportion of participants – regardless of whether they explicitly reported pro-forgiveness or anti-forgiveness attitudes – did not place much emphasis on whether or not to *add* points to their opponent's score. However, at the automatic level, this kind of reasoning should not have been operating; implicit associations should have predicted more impulsive behavioural responses. Post-hoc analyses would appear to support this hypothesis: a series of independent *t*-tests finds no significant differences between those who did not add any points (N=51) and those who took at least one of the adding options (N=11) on all of the explicit measures, whereas IAT scores were significantly greater for those who chose to take at least benevolent decision, $t(60)=2.84$, $p=.006$.

An alternative explanation for the relationship between revenge and explicit measures may have been to do with the order of measures. Whereas the IAT was completed at the beginning of the study, the explicit forgiveness attitude measures were administered at the end, after the transgression and behavioural measures had taken place. It is possible that there may have been an effect of participants inferring their attitude from the way in which they behaved towards the other player, consistent with self-perception theory (Bem, 1967). For example, a participant who chose to deduct all of their opponent's points may have then felt compelled to report that they are not a very forgiving person, even if they had previously believed that they were. The (unusually) poor reliabilities reported for some of these explicit measures suggest that participants did have difficulty completing these measures in a consistent way. Unfortunately, addressing this potential confound is difficult, as completing forgiveness attitude scales before behavioural measures may likewise have unwanted priming effects on behavioural responses. Additionally, this account does not explain why participants would not have also inferred their forgiving attitudes from their benevolent behaviour.

6.4.3 Predicting controlled forgiving responses

Based on the double dissociation model, it had been hypothesised that forgiveness attitude scales should have been better predictors of more deliberative measures of forgiveness: self-reported forgiveness and trust towards the transgressor. Contrary to hypothesis, none of the forgiveness attitudes scales significantly predicted either forgiveness or trust. However, unexpectedly, the forgiveness-revenge IAT did

significantly predict self-reported post-transgression trust. Furthermore, the IAT also marginally predicted self-reported forgiveness of the transgressor.

These unexpected finding raise two important questions. First, why should implicit forgiveness attitudes/associations predict *controlled* forgiving behaviour but explicit attitudes not? Second, if the forgiveness-revenge IAT can predict controlled forgiving behaviour, then why did it not also do so in the previous chapter, in response to recalled transgressions? Each of these questions may share a common answer. As noted earlier, existing attempts to apply the double-dissociation model to implicit/explicit measurement research have tended to focus on the conditions under which participants have little *opportunity* to control their behaviour, by observing non-verbal cues or experimentally limiting an individual's cognitive capacity. However, this may not be addressing the complete picture. According to Fazio and Olsen (2003), the extent to which behaviour will be influenced by automatic processes is dependent not just on opportunity, but also on *motivation*. In Study 9, it is plausible that participants may have been implicitly motivated to behave in a particular way towards the transgressor, even though they were not explicitly conscious of this.

Specifically, participants may have been implicitly motivated by fear of the potential threat of recidivism, operating at a non-conscious level. The automatic nature of fear responses is well-documented, predicting initial behaviour before more conscious, rational processes can take over (Mineka & Öhman, 2002; Arne Öhman, 2005). It is plausible that decisions about whether or not to trust a person who has just harmed you could be driven by a fear response, even if this is not operating at the conscious level. The four behavioural items used to measure trust all predominantly addressed the

participant's willingness to allow the transgressor to have control over future outcomes for the participant, such as "making decisions that critically affect" or "working on a task or problem that was important to" the participant. At the implicit level, a person who has less favourable attitudes toward forgiveness may be more likely to perceive a transgressor as a threat, resulting in lower trust. No such connection should exist at the explicit level, as the participant is consciously aware that there is no real threat, as the study has now finished.

6.4.4 Implications and further research

Irrespective of the mechanisms by which the present findings occurred, they represent an important step forward for research on forgiveness and the IAT. Together, findings from the present study provide the first evidence that the forgiveness-revenge IAT may be useful for predicting specific kinds of real world behaviour. Of particular importance is the finding that the IAT may be useful for predicting behaviour that cannot be accounted for by explicit self-report measures of forgiveness attitudes and/or dispositions: specifically benevolence and trust in response to a real-time transgression.

Forgiveness is generally seen as a conscious, controlled and deliberated process, with investigations of forgiveness at the automatic level almost non-existent. The one study that has thus far demonstrated the automatic of forgiveness has only shown that forgiveness can be evoked *more* automatically when thinking about a close – relative to non-close – other (Karremans & Aarts, 2007). The present work extends these findings by demonstrating that forgiveness can also operate automatically when the transgressor is a non-close (unknown) other.

Of course, the links between the forgiveness-revenge IAT and spontaneous behaviour have only been demonstrated in one study, and need to be replicated. This need for replication is heightened by the poor reliabilities reported for some of the explicit forgiveness measures, which may have impacted on this study's ability to accurately assess these constructs. Additionally, in attempting to create forgiving responses that were more automatic, the present work only used a single methodology: placing time restrictions on responses. As noted earlier, automatic behaviour may be observed in several ways, including coding non-verbal behaviours, and by depleting self-regulatory cognitive resources. A necessary next step for research on implicit forgiveness research is to determine if a forgiveness IAT can predict other types of automatic forgiving behaviour, using other types of transgressions, and other types of automatic behaviour.

What the present work does achieve is the provision of preliminary evidence that a forgiveness-revenge IAT is useful for predicting forgiveness of an actual transgression. With further refinement, the IAT shows promise as another tool to help further develop current understandings of forgiveness, especially in relation to some of the more automatic components of the forgiveness process.

Chapter 7:

General Discussion

7.1 Overview

The aim of the present work was to determine if an Implicit Association Test (Greenwald et al., 1998) could be used effectively to measure forgiveness attitudes. Developing an IAT suitable for forgiveness measurement addressed a significant limitation of forgiveness research to date: its near-exclusive reliance on self-report methodologies (Hoyt & McCullough, 2005). Additionally (with one exception) studies on forgiveness have been typically confined to exploring the conscious, reflective aspects of the forgiveness process (*cf:* Karremans & Aarts, 2007), which potentially limits current understandings of the construct. The forgiveness-revenge IAT – as an indirect measure of automatic attitudes – directly addresses both of these concerns. Data from nine studies provide initial evidence that the IAT may be a valid and reliable tool for measuring forgiveness. Furthermore, there is some evidence that IAT-measured forgiveness associations can predict behavioural responses to a transgression that are not explained by self-reported forgiveness attitudes or trait forgiveness. The findings from these studies, along with their implications, limitations of the present work, and recommendations for future research, will be discussed.

7.2 Summary of findings

The thesis has presented findings from nine studies which developed several variants of a forgiveness IAT. Chapter 2 presented two studies which assessed the appropriateness of the words used to represent the forgiveness category in the IAT, with a particular emphasis on the valence of these words. Chapters 3 (Study 3) and 4 (Studies 4 and 5) shifted the focus from the level of individual stimulus words to the over-arching

categories: first examining the impact of the target categories (forgiveness-revenge/grudge/justice), and then the attribute categories (pleasant-unpleasant compared with self-other) on IAT scores. Chapters 5 (Studies 6, 7, and 8) and 6 (Study 9) evaluated the effectiveness of a forgiveness IAT in predicting actual forgiving behaviour: first by assessing incremental validity in self-reported responses to recalled transgressions, and then by assessing the IAT's utility in predicting automatic forgiving behaviour, in response to both hypothetical and lab-based transgressions. Throughout these studies, the validity of the forgiveness IAT was assessed against a range of criteria: the impact of both stimuli and categories used to represent forgiveness and its IAT counterpart; resistance to socially desirable responding; convergence with self-reported forgiveness attitude measures; and predictive validity. The main findings from these studies are presented below.

7.2.1 The Forgiveness-Revenge IAT measures forgiveness associations

The forgiveness-revenge IAT seems to measure an individual's forgiveness-related associations, rather than capturing methodological noise associated with heuristic processing. Data presented in studies 1 and 2 demonstrated that the forgiveness IAT was not impacted by a potential confound of IAT effects: that is, strategic recoding based on valence of stimuli. This finding is important, given that several previous studies had reported that some IATs could be significantly affected by the valence of stimuli exemplars (Govan & Williams, 2004; Mitchell et al., 2003). Importantly, scores on a forgiveness-revenge IAT did not differ significantly based on the valence of the words used to represent the forgiveness category, inspiring confidence that these IAT scores

were representing actual forgiveness-related associations. Of course, this finding does not rule out other extraneous influences on IAT performance, but it does provide initial evidence that the IAT is suitable for forgiveness measurement.

More convincing evidence that the forgiveness IAT measures actual forgiveness preference comes from its convergence with explicit measures of forgiveness attitudes. Meta analysis of implicit-explicit correspondence across the nine studies reveals that the IAT correlates significantly and modestly with two measures of forgiveness attitudes: the ATF (r=.14, p<.001) and the 'other' subscale of the HFS (r=.15, p<.001)[23]. The magnitude of this implicit-explicit convergence is not dissimilar to that typically found in the IAT literature: on average lower than r=.25 (Greenwald et al., 2009; Hofmann et al., 2005a). Moreover, these effect sizes are comparable to those found for more socially sensitive domains, such as racial prejudice (r=.12; Greenwald et al., 2009). Forgiveness may be another such domain. Additionally, Greenwald et al.'s (2009) meta-analysis reported that implicit-explicit correspondence was lowest for studies which had used the IAT in close relationship contexts (mean r of .09 across 10 such studies), which is again relevant for the forgiveness IAT (given the salience of forgiveness in close relationships; Karremans & Aarts, 2007). Taken together, these prior studies suggest that the convergence found between the forgiveness IAT and explicit measures of forgiveness attitudes is at a level which should be expected.

However, although the forgiveness IAT was significantly related to explicit measures of forgiveness *attitudes*, it was not, for the most part, related to explicit

[23] These effect sizes increase marginally if Study 6 (in which the IAT was primed) is excluded from the analysis.

measures that are *dispositional* in nature. Meta-analysis revealed non-significant relationships between the IAT and the TTF (r=.04, p=.16, 9 samples), and the IAT and WTF (r=-.01, p=.75, 3 samples). The TTF assesses trait forgiveness in a generalised sense (i.e. how much a person *tends* to forgive across situations), whereas the WTF specifically asks participants to indicate how much they would forgive a series of hypothetical transgressions. This is in contrast to the attitudinal measures - the ATF and HFS - which ask participants to make an *evaluation* of forgiveness in terms of whether it is good/bad. Perhaps it is this distinction between evaluations and behavioural tendencies that accounts for the IAT's differential convergence with these measures. The IAT is – by its very design – an evaluative tool, so it may be reasonably expected to relate to explicit measures with an evaluative emphasis.

Irrespective of the explanation for why the IAT converged with some forgiveness scales and not others, the key point is that it *did* converge with some measures of self-reported forgiveness. This finding suggests that forgiveness measured using the IAT is at least partially related to explicit forgiveness attitudes, which reinforces the construct validity of the forgiveness IAT.

7.2.2 "Revenge" and "justice" are equally useful contrast categories in a forgiveness IAT

The choice of category used to contrast with forgiveness had a negligible impact on performance of the forgiveness IAT, although *revenge* and *retributive justice* appeared to be more suitable categories than *grudge*. The IAT is a relative measure, always requiring a target category to be compared with another relevant, complementary

construct. Consequently, both proponents and critics of the IAT have warned that selection of this contrast category is critical, and can have important implications for the construal of the implicit association (De Houwer, 2003; Nosek et al., 2007; Lane et al., 2007). Across three studies, there were no significant differences between IATs contrasting forgiveness with one of revenge, grudge, or retributive justice. The three IAT variants did differ somewhat in their correlations with explicit forgiveness measures, with both revenge and justice outperforming grudge in terms of convergent validity. These findings suggest that both revenge and retributive justice are equally useful contrast categories for a forgiveness IAT. However, given lay and theoretical conceptualizations of revenge as the natural opposite to forgiveness, it is recommended that future research using a forgiveness IAT proceeds with revenge as the contrast category.

7.2.3 The Forgiveness IAT appears resistant to socially desirable responding

Across the nine studies, the forgiveness IAT was not significantly associated with socially desirable responding, as measured by three different SDR scales. This finding is important, given that one rationale for developing a forgiveness IAT was to provide an alternative to self-report forgiveness measures, which would not be affected by the same sources of error. The IAT was specifically chosen because it purports to be resistant to self-presentation distortions.

In contrast to the IAT, self-reported forgiveness attitudes *were* significantly related to SDR measures. These relationships are summarised in Table 7.1 (below), which presents a meta-analysis of the effect sizes for relationships between the three explicit

forgiveness measures that were included in all nine studies – the ATF, HFS, and TTF – with the three measures of SDR.

Table 7.1

Meta-analysis of Correlations Between Self-report Measures of Forgiveness Attitudes and Social Desirability Scales Across the Nine Studies

	SDS-17	Ballard MCSF	BIDR-SDE	BIDR-IM
ATF	.06	.12**	.08	.22***
HFS	.20***	.31***	.13*	.24***
TTF	.21***	.30***	.15**	.13*

* $p<.05$. ** $p<.01$. *** $p<.001$. SDS-17: 5 studies, N=831; Ballard MCSF: 4 studies, N=695; BIDR: 4 studies, N=343

These findings suggest that while SDR is associated with self-reported forgiveness, it is not with the IAT. The absence of evidence for SDR affecting the IAT does not, of course, definitively preclude the possibility that SDR may impact on the forgiveness IAT. SDR was only assessed using a scale approach, and experimental approaches to assessing the measure – under a range of socially sensitive conditions – may be needed to further understand this relationship. For the moment, consistent null findings across nine studies, using three different SDR scales, suggest that the forgiveness IAT is relatively robust to self-presentation factors.

7.2.4 The Forgiveness-revenge IAT can predict behaviour

Under certain conditions, and for some types of transgressions, the forgiveness-revenge IAT may be useful for predicting behaviour. The present work assessed the utility

of the IAT in predicting forgiving behaviour in response to a range of transgressions: recalled, hypothetical, and laboratory-based. This forgiving behaviour was also assessed in a variety of forms – benevolence, revenge, avoidance, and trust – and with varying degrees of controlled cognitive processing. These approaches yielded mixed results for the predictive validity of the forgiveness-revenge IAT.

Initial explorations suggested that the IAT was not useful for assessing deliberative, controlled behavioural responses to a recalled transgression (Studies 6, 7, and 8). Consistent with the literature (e.g. Fehr et al.'s, 2010 meta-analysis), explicit forgiveness measures significantly predicted forgiveness motivations: r=-.31 for avoidance, r=-.35 for revenge and r=.38 for benevolence across Studies 6, 7, and 8. However, the IAT did not significantly explain any variance in these motivations, either uniquely or additionally. The forgiveness-revenge IAT was not useful for predicting forgiveness of a recalled transgression.

The forgiveness-revenge IAT performed no better in predicting behaviour in response to hypothetical offenses. In Study 9, participants were presented with a series of brief hypothetical transgressions, varying in severity and closeness of relationship with the transgressor. Decisions to forgive or not were made under time pressure. Previous research has found that forgiveness attitudes/dispositions predict forgiving behaviour in response to hypothetical scenarios at a mean of r=.34 (Fehr et al., 2010). Likewise, imposing time restraints on one's behavioural decisions can lead to more automatic behaviour, which the IAT should predict better than would self-reported attitudes (Friese et al., 2006). Based on these two literatures, it was expected that the forgiveness-revenge IAT should have predicted these kinds of forgiving responses. However, forgiving

decisions were significantly predicted by explicit (ATF) – but not IAT-measured – forgiveness attitudes.

The forgiveness-revenge IAT performed equally as well as – and in some respects better than – explicit forgiveness measures in predicting forgiving responses to an actual real-time transgression. Consistent with a double-dissociation model (Fazio, 1990; Perugini, 2005; Wilson et al., 2000), it was anticipated that the forgiveness IAT would predict forgiving responses that were less amenable to controlled processing, such as those made under time constraints. Consistent with hypothesis, utilising an iterated trust game methodology to engineer a transgression (Study 9), the IAT did predict time-pressured *benevolent* responses better than explicit measures. However, the opposite effect was found for time-pressured *revenge* responses: explicit measures were better predictors of these. The hypothesis was only partially supported.

Transgressing against participants using an iterated trust game provided an additional – and surprising – finding: the IAT was not only useful for predicting forgiveness responses that were more *automatic*, but it was also the only significant predictor of more *controlled* forgiving responses. The forgiveness-revenge IAT significantly predicted self-reported post-transgression trust towards a transgressor, with the IAT also marginally predicting explicit reports of whether participants would be likely to forgive the transgressor. None of the explicit forgiveness measures significantly predicted these direct reports of transgressor-directed trust or forgiveness. This finding was unexpected, given that the double-dissociation model suggests that explicit attitudes should have outperformed the IAT in predicting this type of response.

Although there is evidence that the forgiveness-IAT can be useful for predicting both automatic and controlled responses to a real-time transgression, this has only been demonstrated in a single study, and requires replication. Nevertheless, this finding provides initial promise that the IAT may be useful for predicting actual behaviour. Together, these findings suggest that the IAT is a valid, reliable, and useful tool for the measurement of forgiveness. There are several implications of this work – mostly theoretical, but some practical – for understanding forgiveness and for how it is measured. Furthermore, some of these findings also have implications for the IAT, and the broader literature on implicit measurement. These implications will be discussed next.

7.3 Implications for Forgiveness

7.3.1 Addressing the "mono-method" bias

One goal of the present work was to address what Hoyt and McCullough (2005) refer to as a "mono-method" bias in forgiveness research, which may limit current understandings of forgiveness. To address this bias, the authors advocated taking a multi-method approach to forgiveness measurement. Specifically, they identified a need for additional modes of measurement that were not self-report scales: different enough in structure that they did not share the same kinds of 'bias' variance, but similar enough that they still assessed the same basic construct. The forgiveness IAT adequately addresses both of these needs.

One of the more common sources of bias variance in self-reported attitude scales is socially desirable responding – variance that is not shared with the IAT. Generally, the

IAT claims to be resistant to self-presentation strategies, with evidence suggesting that individuals are typically not able to fake their responses without prior instruction on how to do so (Kim, 2003; Steffens, 2004). Findings from the present work are consistent with this claim: the forgiveness IAT was not associated with social desirability factors, whereas explicit forgiveness measures typically were.

The IAT was also designed to overcome another potential source of error variance in self-report measures: the fact that individuals may be unable to accurately report their attitudes because they do not necessarily always have conscious access to them. The IAT is not limited in the same way, as it is not a requirement of the IAT that an individual be consciously able to introspect on their inner cognitive processes (Greenwald et al., 1998). Convergent/divergent and predictive validity findings from the present work suggest that the forgiveness IAT may have been assessing forgiveness attitudes of which participants were not consciously aware. The IAT and forgiveness attitude measures converged, suggesting that they are at least partly assessing a similar construct. However, the two types of measures were generally more divergent than convergent, suggesting a disparity between the automatic associations people make about forgiveness, and what they are consciously able to report[24]. Moreover, each of the types of measures significantly predicted forgiving behaviour, and in some cases the IAT was a more useful predictor

[24] Of course discriminant validity alone is insufficient evidence for concluding that the IAT assesses unconscious forgiveness: the divergence may result from self-presentation or other motivational factors, or from a range of other individual difference or methodological factors. However, interpreting this evidence in conjunction with the IAT's unique predictive utility suggests that the IAT is assessing non-conscious processes.

than the attitudes that participants were able to consciously report. Together, these findings suggest that the forgiveness IAT may be a useful tool for addressing the mono-method bias in forgiveness research.

Nonetheless, despite the forgiveness IAT showing promise in overcoming some of the sources of error common to self-report scale measures, it is not recommended that it be used in isolation. In highlighting the mono-method bias, Hoyt and McCullough (2005) were not advocating the development of new methods as *alternatives* to replace self-report measures. Rather, these authors were suggesting that forgiveness research adopt *multiple* measures within a single study design. Thus, the forgiveness-revenge IAT is not presented as an alternative to existing measures of forgiveness attitudes, dispositions, or self-concept, but instead offered as a *companion* to these tools. The IAT, too, has its own sources of bias variance that are not problematic for self-report measures. For example, the IAT has been shown to be influenced by situational factors (Blair, 2002; Han, Czellar, Olsen, & Fazio, 2010), and is less stable over time than self-reported attitudes (Schmukle & Egloff, 2005). Using the IAT as a standalone measure would not be addressing the mono-method bias: it would merely be creating a different kind of mono-method study.

As demonstrated in Study 9, using both explicit and implicit forgiveness measures provided insight in to the forgiveness process that was more nuanced than either measure could have provided in its own right. If the explicit forgiveness attitude scales had been omitted, then there would have been no information about predicting revenge under time constraints, nor about predicting time-pressured hypothetical forgiveness decisions. Similarly, had the forgiveness-revenge IAT not been included then there would have been no evidence connecting forgiveness attitudes to post-transgression trust

towards a transgressor. Thus, merely using one of these methods would have significantly altered the conclusions that were drawn from the findings. Together, these findings support Hoyt and McCullough's (2005) recommendation that no single approach should be used in isolation. Specifically, the IAT should be viewed as a companion to, not a replacement of, existing methods.

In addition to producing a more refined understanding of forgiveness processes - particularly the links between forgiveness attitudes and behaviour - using a forgiveness IAT alongside other forgiveness measures may have particular utility in contexts where individuals may be motivated to conceal their true attitudes towards forgiveness. For example, Christians have been found to report more positive attitudes towards forgiveness, even though actual forgiving behaviour tends to be no different between Christians and non-Christians (McCullough & Worthington, 1999)[25]. At the explicit level, Christians may consciously believe that they are more forgiving people, or may be motivated to present themselves as more forgiving. However, it is possible that examining forgiveness attitudes at the implicit level may reveal different patterns of differences between Christians and non-Christians, and assist in further illuminating this relationship.

7.3.2 New perspectives on forgiveness

The forgiveness IAT has potential in helping to offer further insight in to the way in which forgiveness is defined and understood. As noted in Chapter 1, despite the growing

[25] Although see Tsang et al. (2005) for an alternative explanation of this finding.

research focus on forgiveness within psychology, there is still much contention about how exactly forgiveness should best be defined. The present work provides some new information which may help to resolve this definitional debate.

7.3.2.1 Forgiveness operates at the automatic level

Perhaps one of the most telling contributions of the present work is the evidence it provides for forgiveness operating automatically. In the literature, forgiveness is almost exclusively understood as a conscious and deliberate process (e.g. Enright et al., 1998; Worthington, 1998). Karremans and Aarts (2007) do provide some evidence that forgiveness may occur more readily when individuals are primed to think of a close other than a non-close other, which may indicate that forgiveness is part of automatically activated "relational schemas" (Baldwin, 1992). However, the present work is the first to specifically and directly measure the automatic components of forgiveness, both in their attitudinal and behavioural forms. Findings from studies 6 through 9 suggest that forgiveness measured at the more automatic level – using the IAT – is distinct from more consciously reported forgiveness, and may predict different types of forgiving behaviours. For example, self-reported forgiving responses to both recalled and hypothetical transgressions were better predicted by consciously deliberated forgiveness attitudes, perhaps owing to the level of cognitive effort required to report on transgressions in this way. In contrast, forgiving behaviour in response to a lab-based transgression was better predicted by automatic forgiveness (IAT) preferences than by more consciously controlled attitudes.

Furthermore, some aspects of forgiveness may operate more automatically than others. Following the lab-based transgression, participants engaged in a timed-decision task which contained both benevolent (add points) and vengeful (subtract points) options. It had initially been hypothesised that the forgiveness IAT would be the better attitudinal predictor of both of these types of behaviour, owing to the fact that they were both time-pressured tasks, which should have produced more automatic behaviour. However, despite both representing more automatic behaviours, they were each predicted by a different type of attitude measure: the IAT predicted automatic benevolence but not revenge, whereas an explicit measure (HFS) predicted revenge but not benevolence. Although the reasons for this are unclear, this finding suggests that it is important to consider forgiveness as both a conscious *and* automatic process.

Additionally, these findings extend those of Karremans and Aarts (2007) in two important ways. First, the present work demonstrates that forgiveness may not only operate more automatically in the context of close relationships, but may do so more generally. In Study 9, the transgressor was always a non-close other (a person unknown to the participant) yet implicitly-measured forgiveness preferences significantly predicted forgiving behaviour towards this transgressor. Automatic forgiveness appears to therefore be more universal than previously demonstrated, and cannot be completely explained as a routine component of one's 'relational schema'.

Second, Karremans and Aarts (2007) acknowledged that they had only demonstrated automatic forgiveness in response to a transgressor who was "psychologically present", and urged further research to assess the automaticity of forgiveness in more real life contexts. The present work has done precisely that, by

demonstrating that automatic forgiveness preferences were the best predictors of forgiving behaviour in response to a real-time transgression.

The finding that forgiveness may occur automatically represents a significant contribution to the forgiveness literature, and may have practical implications. Across the nine studies, participants overwhelmingly implicitly preferred forgiveness to revenge, grudge or retributive justice. At a practical level, if automatic forgiveness can drive behaviour, and if people implicitly prefer forgiveness to the alternatives, then this bodes well for a positive society. McCullough (2008) describes forgiveness as an *instinct,* a capacity with which human beings come pre-programmed, and cites an array of (predominantly) evolutionary and anecdotal evidence to support this claim. According to McCullough, this forgiveness instinct is what stops the human race from destroying itself through endless cycles of retaliation and revenge. If forgiveness is an evolutionary adaptation then it stands to reason that it should at least partially operate at an automatic level. The present work provides behavioural evidence in support of this.

7.3.2.2 Defining what forgiveness is not

The present work may be useful for further understanding the constructs that may or may not be related to forgiveness. Forgiveness theorists have made specific claims about what forgiveness is not: that it is distinct from discounting/minimising concepts such as pardoning, excusing, condoning or accepting (Enright & North, 1998a; Fincham, 2000). However, results from the present work suggest that (at least at the association level) these concepts may be important components in the way in which people conceptualise forgiveness.

Studies 1 and 2(b) directly compared IATs that used different words to represent forgiveness. In each of these studies, one IAT variant represented forgiveness as a *prosocial* construct. A second IAT represented forgiveness using *minimising* constructs (e.g. words like excusing, overlooking, condoning). A comparison of these two IAT variants revealed no significant differences in mean IAT scores, in either of the two studies, suggesting that the distinction between prosocial and minimising aspects of forgiveness may not be as relevant as is sometimes claimed. Both of these components appear to be relevant to people's understandings of forgiveness.

7.2.3.3 Forgiveness as other-focused

Forgiveness research has tended to focus on just one party: the victim (*cf*: Kelln & Ellard, 1999; Struthers et al., 2008). Similarly, studies which have directly asked people about their understandings of forgiveness have revealed that people generally conceptualise it as self-focused, with the most common responses indicating that people forgive for their own health and happiness (Younger et al., 2004), and that the transgressor need not even know that they have been forgiven (Kanz, 2000). However, the present findings suggest that the association between *forgiveness* and *other* may be equally important as the association between *forgiveness* and the *self*.

The first indication of the importance of the implicit forgiveness-other association was in the unexpected negative relationship found between the forgiveness-grudge IAT and the WTF scale in Study 5. Detailed analysis of this relationship, by examining reaction times on individual IAT blocks, indicated that those who scored higher on the WTF (i.e. were more forgiving of specific situations and people) found it more difficult to associate

forgiveness with *self*. A similar albeit marginal effect was observed between the forgiveness-justice IAT and the WTF in Study 3. In contrast, in both of these studies the explicit measures of forgiveness that were *context-independent* (ATF and HFS) produced relationships with the IAT in the expected direction, such that those with more positive attitudes toward forgiveness found it *easier* to associate forgiveness and self. These findings suggest that forgiveness attitudes may elicit more forgiveness-self associations, but that forgiveness in the context of an actual transgression may generate stronger associations between forgiveness and other.

Further, and perhaps more convincing, evidence for the context-dependent associations between forgiveness and other was found in Study 6. Asking participants to reflect on a personally-experienced transgression from their own past immediately before completing the forgiveness-revenge IAT resulted in IAT scores that were related to forgiving behaviour, but in the opposite direction to that which had been predicted. Specifically, results indicated that the more a participant had forgiven the transgressor, the more *forgiveness* was associated with *other,* rather than *self*. Again, this suggests that in the context of an actual transgression – at least at the automatic level – forgiveness is other-focused.

These findings demonstrating an "other" focus of forgiveness may have important theoretical implications for current understandings of forgiveness. Lay conceptualisations of forgiveness have emphasised the self-focus of forgiveness: that it is a largely internal process for the benefit of the victim (Kanz, 2000; Younger et al., 2004). Similarly, forgiveness has also been framed theoretically as a coping mechanism, as a form of self-healing (Strelan & Covic, 2006; Worthington & Scherer, 2004). Findings from the present

work suggest that it may be insufficient to merely focus on the relationship between forgiveness and the self, but rather (at least at the automatic level) associations between forgiveness and other are an integral component of the mechanisms underlying the forgiveness process.

Furthermore, the IAT may have some specific application in further developing an understanding of what forgiveness means to a transgressor. The knowledge that people form intricate implicit patterns of how forgiveness relates to both the self and others could provide valuable insights in to forgiveness from the transgressor's perspective.

7.3.3 Measuring Forgiving Behaviour

In addition to providing a new perspective on the measurement of forgiveness *attitudes*, the present work also provides insight into measuring forgiving *behaviour*. The forgiveness-revenge IAT significantly predicted forgiving behaviour – both automatic and controlled – in response to a standardised laboratory-based transgression, but did not predict behavioural responses to either recalled or hypothetical transgressions. Inversely, self-reported attitude scales successfully predicted responses to recalled and hypothetical transgressions, but were less predictive of responses to a lab-based transgression than was the IAT. Together, these findings suggest that different kinds of mental processes may be operating, purely as a function of the approach taken to measure forgiving behaviour. The relatively strong relationships between explicit attitudes and behaviours in response to recalled transgressions may suggest that this methodology assesses forgiveness at a relatively *controlled* level. In contrast, generating a standard

transgression in the laboratory elicited forgiving responses that relied on both controlled *and* automatic processes.

Intuitively, it makes sense that behavioural responses to a recalled past offense may be qualitatively different than responses which immediately follow a transgression. Responses to a recalled transgression are less reliant on automatic processes, simply as a function of the amount of time that has elapsed since the offense took place. The less recent the transgression, the more time has been available for conscious, controlled deliberation. Additionally, the very act of consciously reflecting on or trying to remember the details of a transgression means that responses require some degree of cognitive elaboration (and greater *awareness*), and will therefore be less automatic in nature (Bargh, 1994).

However, results from the present work using the iterated trust game suggest that conscious, controlled processes may be insufficient for understanding actual forgiving behaviour. If automatic processes are important predictors of forgiving behaviour, and neither recall nor hypothetical approaches are appropriate for examining these automatic processes, then current understandings of forgiveness may be limited. Given that the majority of psychological research on forgiveness utilises either recall or hypothetical methodologies, this then represents a significant problem for forgiveness research, and further reinforces the need to develop new methods for measuring forgiving behaviour.

7.4 Implications for the Implicit Association Test

Although this was not the goal of the present work, these findings may also have several implications for the IAT literature, as well as research on implicit measurement more generally.

7.4.1 IAT scores are influenced by structural factors

The present work provides two (related) contributions to the growing body of work exploring the methodological factors influencing IAT performance, and specifically convergence between IAT and explicit measures.

First, this work provides the first direct evidence that implicit-explicit convergence can be improved by modifying IAT and self-report measures to be more structurally similar. Payne et al. (2008) had previously demonstrated this using an alternative 'implicit' measurement procedure (the IAP) but not with a reaction-time based implicit procedure. Specific to the IAT, correlational evidence had suggested that structurally and conceptually similar explicit and IAT measures would produce greater IE correspondence (Nosek, 2005), but this had not been tested directly.

This finding has significant implications for an IAT literature that is still grappling with a fundamental question: how much should implicit and explicit measures be expected to converge? Fazio and Olsen's (2003, p.304) answer to this question – that "it depends" – has generated much subsequent research and debate into the factors that "it depends" on. That is, what are the moderators of implicit-explicit correspondence? Structural similarity of measures may be one such important moderator.

Following from the findings addressing structural fit, the present work is also the first to directly compare two different types of IATs for assessing the same construct. Studies 3, 4, and 5 compared forgiveness attitude (pleasant-unpleasant) with forgiveness self-concept (self-other) IAT variants, finding that the forgiveness self-concept IAT was superior in its convergence with explicit forgiveness attitude measures. This has implications for implicit measurement research more broadly, highlighting the importance of selecting the appropriate kind of IAT, which may have a meaningful impact on the results of a study. The choice of IAT variant may be another such moderator of the implicit-explicit relationship.

Of course, opting for an alternative IAT variant may not necessarily be useful nor practical for many of the constructs typically investigated in IAT research, particularly in the domain of prejudice. For example, using a self-concept IAT would be inappropriate for assessing racial prejudice using a black-white target pairing, as difficulty associating 'black' with 'other' is more indicative that a (white) participant does not belong to a particular group (black), rather than possessing negative associations towards blacks. Similarly, research on implicit anxiety typically uses a self-concept IAT, as an attitude IAT would be inappropriate: most people find anxiety *unpleasant*, irrespective of their own levels of trait anxiety. That is, *attitudes* towards anxiety are relatively unhelpful for examining individual differences in anxiety.

The choice between using an attitude or self-concept IAT appears then to be most appropriate for constructs which can be viewed as both an attitude *and* a trait. Besides forgiveness, the IAT has been applied to a range of other such constructs, such as humility (Powers et al., 2007), honesty/deception (Jung & Lee, 2009), political ideology (Choma &

Hafer, 2009), suicide (Nock et al., 2010), smoking (Robinson, Meier, Zetocha, & McCaul, 2005), and alcohol (Houben, 2007; Wiers et al., 2002) - all of which could be assessed using either attitude or self-concept IATs. Further research in these domains should investigate the impact of using different types of IATs, which may improve convergence between these IATs and their equivalent explicit measures.

7.4.2 Malleability of implicit associations: Incongruent priming effects

The present work found mixed evidence for the malleability of the IAT. On the one hand, features of the IAT stimuli – specifically, valence – did not have a meaningful bearing on forgiveness IAT scores. On the other hand, thinking about forgiveness in context – by either recalling a real offense (Study 6), or considering hypothetical transgressions/transgressors (Study 5) – can lead to increased associations between *forgiveness* and *other* on subsequent IAT tasks. This finding represents a significant contribution to current understandings of the malleability of implicit measures.

The idea that IAT measures may be malleable is not new. There is now ample evidence that, given the right conditions, IAT scores can reflect associations that have only been learned immediately before completing the IAT task (e.g. Dasgupta & Greenwald, 2001; Han et al., 2010; Wallaert, Ward, & Mann, 2010). However, research examining the malleability of the IAT has typically focused on just one aspect – valence. Specifically, it has been demonstrated that inducing participants to feel more positively towards a particular target can result in a more pleasant evaluation of that target on subsequent IAT measures. Similarly, research focusing on the malleability of the self-concept IAT has demonstrated that priming can induce stronger associations between the

target and *self* (Baccus et al., 2004; Bluemke et al., 2010; Uhlmann & Swanson, 2004). However, the present work is the first to demonstrate that this priming can take place in an *incongruent* direction: that priming can result in stronger associations not just between target and self categories, but also between the target and *other* categories. Furthermore, priming of an IAT can occur without learning *new* associations: simply directing participants to think about a pre-existing association was enough for this association to be reflected in IAT measures.

These findings have several implications for implicit measurement research. First, evidence for malleability of the forgiveness IAT further augments a growing body of evidence suggesting that IATs are somewhat unstable. Large reviews of IAT studies have suggested that the ordinal position of the IAT relative to other tasks has no significant impact on IAT performance, in terms of either IE correspondence (Hofmann et al., 2005a; Nosek, 2005), or predictive validity (Greenwald et al., 2009). However, at least in the case of forgiveness, the order of tasks *does* seem to matter. If completing a behavioural task prior to the IAT can influence scores on that IAT, then this finding potentially has implications for the (re)interpretation of several other published studies exploring the predictive validity of the IAT. For example, Jung and Lee (2009) found a significant relationship between scores on a deception-honesty IAT and actual deceptive behaviour in a computer game. However, the IAT was administered *after* the game, which means that it is impossible to determine whether these IAT scores reflect stable attitudes/dispositions or if they were influenced by participants' behaviour in the game.

Fortunately, the priming observed in the present work is easily overcome, by placing the IAT at the beginning of a study. A replication of Study 6 with the IAT

completed before any of the other tasks (Study 8) found no evidence of priming effects. It is recommended that future research using the IAT alongside behavioural or context-dependent tasks positions the IAT measure at the beginning of the study.

7.5 Limitations and future research

7.5.1 Methodological constraints of taking a scale approach to measuring socially desirable responding

Throughout the nine studies there were no significant correlations between the forgiveness IAT and any of the measures of socially desirable responding – a finding which has thus far been used as evidence that the forgiveness IAT is a valid instrument. However, these studies may have been limited by the approach taken towards accounting for SDR concerns.

SDR was only measured using self-report scales, which may have been problematic for several reasons. There is an irony that a thesis which aimed to address some of the deficiencies of self-report measures (attitude scales), would do so by using a self-report measure (SDR scales) as a validating tool. SDR scales may share common variance with attitude scales simply because they also share a common methodology, which may explain their greater convergence with explicit rather than IAT-measured forgiveness attitudes.

There may have also been some more specific problems with using scales to assess SDR. First, two of the three SDR scales used throughout the studies consistently demonstrated either poor (mean α=.59 for the Ballard, 1992 measure) or borderline (mean α=.65 for the SDS-17) internal consistency reliability. However, perhaps more

troubling is the possibility that these measures may not have been exclusively measuring socially desirable responding tendencies, but may have conflated these tendencies with another forgiveness-related personality construct. There was an unusual finding in that forgiveness (and revenge) measures seemed to covary with SDR scales, independent of actual SDR. For example, results from Study 2(b) found that not only did SDR measures positively correlate with forgiveness attitude scales, but they also *negatively* correlated with an explicit measure of revenge attitudes. One plausible explanation for this is that these SDR scales measure generalised prosocial tendencies: if a person is more likely to engage in socially desirable behaviours such as putting litter in the bin then perhaps that person is also more likely to engage in other prosocial acts, such as forgiving others.

A second explanation for these findings concerns social inhibition and control. In reviewing the application of SDR scale measures, Uziel (2010) proposes a reconceptualisation of what these scales actually measure. Specifically, Uziel argues that these measures are more indicative of an individual's ability to exercise self-control, particularly in social contexts. If this is the case, then it is unsurprising that a measure of revenge would negatively correlate with SDR scale measures: those individuals who can exercise greater self-control – particularly in regards to emotions and impulses – should logically also self-report less vengeful attitudes.

Given the limitations of a scale approach in assessing social desirabilty factors, alternative methods are required for evaluating the forgiveness IAT's robustness to self-presentation concerns. Ideally, these approaches would adopt experimental methodologies, exploring the whether the forgiveness IAT is more predictive of behaviour

under different types of socially sensitive conditions. However, taking an experimental approach to SDR was beyond the scope of this thesis.

7.5.1 Using the IAT to predict other types of behaviour

The iterated trust game approach had both its advantages and disadvantages. The trust game was useful because it allowed for transgressing in real time, with the inclusion of both cognitive and affective aspects of a transgression. At the cognitive level, the game allowed for an outcome to occur that was objectively unfair, with this trust violation representing an *injustice*. Affectively, the game had an in-built messaging system, which allowed for unwarranted insults and the hurt to be *emotionally experienced*. Together, these two features represent aspects common to many typical transgressions, ensuring that participant responses should have been relatively representative of behaviours in the real world.

However, although effort was made to ensure that the trust game transgression was as psychologically real as possible, it also differed from a typical transgression in several key ways. Most importantly, the transgression took place between two individuals who had had no prior contact. In contrast, the majority of transgressions for which forgiveness is relevant occur between two parties who are involved in relatively close interpersonal relationships. Thus, it is unknown whether findings using the trust game will generalise to close relationships. Second, the trust game represented a transgression that was de-contextualised, which meant that the participant (victim) was equipped with no prior knowledge of the transgressor on whom they could base their judgments or behaviours. This absence of context is an important consideration given

evidence that a partner's past behaviour is a significant predictor of future forgiveness of that partner (Effron & Monin, 2010).

Additionally, the fact that the transgression was de-contextualised also meant that there was no scope for future reciprocity or recidivism. The participant was aware that their relationship with the transgressor would end within half an hour of the transgression taking place, which means that the participant's behaviour would have no long-term consequences. In contrast, interpersonal relationships in the real world typically involve mutual exchanges, with forgiveness/punishment decisions often having more long-range repercussions (Baumeister et al., 1998; Murphy, 2005). Finally, the trust game also differed from most real life transgressions in that the entire interaction occurred on a computer: there was no face-to-face interaction. Transgressions which occur face-to-face may operate differently, as they provide for richer communication (visual cues such as non-verbal behaviour), and may elicit different types of behavioural responses. Attempts to further validate the forgiveness IAT should seek to move beyond game theory, utilising approaches that possess even greater ecological validity.

Future work directed at understanding automatic forgiveness processes must focus not only on different types of *transgressions*, but must also examine different types of forgiving *behaviour*. The forgiveness IAT may potentially be a good predictor of automatic forgiving behaviour, but a challenge for future research is to identify ways in which this behaviour can be accurately measured. This task is further complicated by ongoing theoretical disagreement about how automaticity may best be defined: there is no consensus on what constitutes an automatic process (Moors & De Houwer, 2006). How, then, can forgiving behaviour be examined at the automatic level? Bargh (1994)

suggests that automaticity is characterised by four criteria: awareness, intentionality, efficiency, and controllability. The time-pressured decisional approach adopted in the present work only addressed on one of these aspects – efficiency – attempting to force more efficient processing by limiting the time available to make these decisions.

Alternative approaches may examine forgiving behaviour in regards to these other dimensions of automaticity (or combinations of these). For example, one fruitful area for future research could be an investigation of more non-verbal behaviours, such as body language. Non-verbal behaviour potentially addresses several of Bargh's criteria, including awareness, controllability and intentionality. Combined with a recall approach, participant body language could be coded while they are giving a verbal recollection of a past transgression. Similarly, non-verbal behaviour could be observed in response to a laboratory based transgression, such as the trust game or a negative feedback paradigm such as those used by Zechmeister et al. (2004) and Wilkowski et al. (2010). Such approaches may reveal new insight in to the nature of implicit forgiveness processes.

7.6.2 The interaction of implicit and explicit processes

A final way in which the findings of the present work could be meaningfully extended is by exploring the *interactions* between implicitly and explicitly measured forgiveness. Throughout this work, implicit and explicit forgiveness attitudes and trait forgiveness were largely treated as independent processes. Chapter 5 adopted an additive approach to understanding the unique and incremental contributions the two types of measures could make towards explaining variance in a single type of behaviour. Chapter 6 shifted to a double-dissociation approach, treating implicit and explicit

preferences as independent constructs, that would each differentially predict different types of behaviour. However, it has been well argued that implicit and explicit processes may be largely *interdependent* and interactive. That is, behaviour is simultaneously influenced by both automatic and controlled processes, which are in turn influencing each other (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Wegner & Bargh, 1998). Thus, there may be much to learn about forgiveness through the way in which explicit and implicit processes interact.

Practically, one way in which to investigate the interaction between the two processes is to apply the multiplicative model (Perugini, 2005). For example, Greenwald et al. (2009) found that when implicit and explicit measures were more convergent, they were also both better predictors of behaviour. Research using the IAT has also uncovered more nuanced interaction effects. For example van Goethem et al. (2010) found that implicit attitudes towards bullying were predictive of bullying behaviour, but only for children who also reported relatively favourable explicit attitudes towards bullying. Similarly, Jordan et al. (2003) demonstrated that implicit self-esteem was predictive of defensive behaviour, but only for those who also reported high levels of explicit self-esteem. It is possible that implicit and explicit measures of forgiveness may interact in a similar way: future research should attempt to investigate these interactions.

7.7 Final comments and conclusions

The aim of this thesis was to develop an Implicit Association Test that is suitable for the measurement of forgiveness. This work has demonstrated that forgiveness IAT is

a valid and reliable tool, with the potential to provide new and unique insights in to the forgiveness process.

The major contributions of the present work are two-fold. First, the forgiveness IAT addresses a well documented gap in the forgiveness literature: the need to move beyond mono-method studies, most of which rely on self-reported information (Hoyt & McCullough, 2005; McCullough et al., 2000). Specifically, the IAT appears suitable for addressing one of the common sources of bias variance that applies to explicit measures: socially desirable responding. Moreover, the forgiveness IAT is able to predict variance in behavioural responses that are not adequately captured by explicit attitude measures. Used together with self-report measures of forgiveness attitudes, the forgiveness IAT holds promise for further developing current conceptualisations of forgiveness.

The second key contribution of the present work is that it highlights the role of some of the automatic processes that may drive forgiving behaviour. Until now, forgiveness theorists and researchers have focused almost exclusively on forgiveness as a deliberative, consciously-directed, controlled process (*cf*: Karremans & Aarts, 2007). Nowhere has this emphasis on the controlled elements of forgiveness been more clear than in models used to explain the forgiveness process (Enright & Coyle, 1998; Worthington, 1998). This thesis presents the first direct evidence that forgiveness may – in some cases – operate at a more automatic level. Furthermore, implicit forgiveness associations may predict different types of behavioural responses than those predicted by standard measures of forgiveness attitudes. The forgiveness IAT provides a means for further examining the role of automatic processes as they relate to forgiveness.

It has now been more than five years since Hoyt and McCullough identified forgiveness researchers' heavy reliance on self-report measures, yet this mono-method bias still remains. This thesis takes significant steps to address this bias, by providing forgiveness researchers with a new tool to add to their measurement toolbox. Importantly, this tool has demonstrated that it can provide additional insight in to the forgiveness process, beyond the measures that already exist. There are still many avenues to be explored, including further developing understandings of how the forgiveness IAT may predict behaviour in response to different types of transgressions, how it may predict more automatic types of behaviour, and how implicit and explicit forgiveness processes may interact. However, the findings of this thesis are encouraging enough that exploring these avenues should prove to be a fruitful exercise.

Bibliography

Ahadi, B., & Ariapooran, S. (2009). Role of self and other forgiveness in predicting depression and suicide ideation of divorcees. *Journal of Applied Sciences, 9*, 3598-3601. doi: 10.3923/jas.2009.3598.3601

Aquino, K., Tripp, T. M., & Bies, R. J. (2006). Getting even or moving on? Power, procedural justice, and types of offense as predictors of revenge, forgiveness, reconciliation, and avoidance in organizations. *Journal of Applied Psychology, 91*, 653-668. doi: 10.1037/0021-9010.91.3.653

Arcuri, L., Castelli, L., Galdi, S., Zogmaister, C., & Amadori, A. (2008). Predicting the vote: Implicit attitudes as predictors of the future behavior of decided and undecided voters. *Political Psychology, 29*, 369-387. doi: 10.1111/j.1467-9221.2008.00635.x

Asendorpf, J. B., Banse, R., & Mucke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology, 83*, 380-393. doi: 10.1037/0022-3514.83.2.380

Axelrod, R. (1980a). Effective choice in the Prisoner's Dilemma. *The Journal of Conflict Resolution, 24*, 3-25.

Axelrod, R. (1980b). More effective choice in the Prisoner's Dilemma. *The Journal of Conflict Resolution, 24*, 379-403.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Baccus, J. R., Baldwin, M. W., & Packer, D. J. (2004). Increasing implicit self-esteem through classical conditioning. *Psychological Science, 15*, 498-502. doi: 10.1111/j.0956-7976.2004.00708.x

Baldwin, M. W. (1992). Relational schemas and the processing of social information. *Psychological Bulletin, 112*, 461-484. doi: 10.1037/0033-2909.112.3.461

Ballard, R. (1992). Short forms of the Marlowe–Crowne Social Desirability scale. *Psychological Reports, 71*, 1155-1160. doi: 10.2466/PR0.71.8.1155-1160

Ballester, S., Sastre, M. T. M., & Mullet, E. (2009). Forgivingness and lay conceptualizations of forgiveness. *Personality and Individual Differences, 47*, 605-609. doi: 10.1016/j.paid.2009.05.016

Balliet, D. (2010). Conscientiousness and forgivingness: A meta-analysis. *Personality and Individual Differences, 48*, 259-263. doi: 10.1016/j.paid.2009.10.021

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger III, J. S. Nairne, I. Neath & A. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117-150). Washington, DC: American Pscyhological Association.

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

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3-51). New York: Guilford.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1-40). Hillsdale, NJ: Erlbaum.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs, 80*, 1-46. doi: 10.1037/h0027577

Baumeister, R. F., Exline, J. J., & Sommer, K. L. (1998). The victim role, grudge theory, and two dimensions of forgiveness. In E. L. Worthington Jr (Ed.), *Dimensions of forgiveness: Psychological research and theological perspectives* (pp. 79-104). Radnor, PA: Templeton.

Bellezza, F. S., Greenwald, A. G., & Banaji, M. R. (1986). Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers, 18*, 299-303.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74*, 183-200. doi: 10.1037/h0024835

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behavior, 10*, 122-142.

Berry, J. W., Worthington Jr, E. L., O'Connor, L. E., Parrott III, L., & Wade, N. G. (2005). Forgiveness, vengeful rumination, and affective traits. *Journal of Personality, 73*, 183-225. doi: 10.1111/j.1467-6494.2004.00308.x

Berry, J. W., Worthington Jr, E. L., Parrott III, L., O'Connor, L. E., & Wade, N. G. (2001). Dispositional forgivingness: Development and construct validity of the Transgression Narrative Test of Forgivingness (TNTF). *Personality and Social Psychology Bulletin, 27,* 1277-1290. doi: 10.1177/01461672012710004

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*, 242-261. doi: 10.1207/S15327957PSPR0603_8

Blake, B. F., Valdiserri, J., & Neuendorf, K. A. (2006). Validity of the SDS-17 measure of social desirability in the American context. *Personality and Individual Differences, 40*, 1625-1636. doi: 10.1016/j.paid.2005.12.007

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009a). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology, 94*, 567–582. doi: 10.1037/a0014665

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009b). Transparency should trump trust: Rejoinder to McConnell and Leibold (2009) and Ziegert and Hanges (2009). *Journal of Applied Psychology, 94*, 598-603. doi: 10.1037/a0014666

Bluemke, M., & Fiedler, K. (2009). Base rate effects on the IAT. *Consciousness and Cognition: An International Journal, 18*, 1029-1038. doi: 10.1016/j.concog.2009.07.010

Bluemke, M., Friedrich, M., & Zumbach, J. (2010). The influence of violent and nonviolent computer games on implicit measures of aggressiveness. *Aggressive Behavior, 36*, 1-13. doi: 10.1002/ab.20329

Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology, 42*, 163-176. doi: 10.1016/j.jesp.2005.03.004

Bluemke, M., & Friese, M. (2008). Reliability and validity of the single-target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology, 38*, 21. doi: 10.1002/ejsp.487

Bohner, G., Siebler, F., Gonzalez, R., Haye, A., & Schmidt, E. A. (2008). Situational flexibility of in-group-related attitudes: A single category IAT study of people with dual

National identity. *Group Processes and Intergroup Relations, 11*, 301-317. doi: 10.1177/1368430208090644

Bono, G., McCullough, M. E., & Root, L. M. (2008). Forgiveness, feeling connected to others, and well-being: Two longitudinal studies. *Personality and Social Psychology Bulletin, 34*, 182-195. doi: 10.1177/0146167207310025

Bosson, J. K., Brown, R. P., Zeigler-Hill, V., & Swann Jr, W. B. (2003). Self-enhancement tendencies among people with high explicit self-esteem: The moderating role of implicit self-esteem. *Self and Identity, 2*, 169-187. doi: 10.1080/15298860309029

Bosson, J. K., Swann Jr, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*, 631-643. doi: 10.1037/0022-3514.79.4.631

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology, 81*, 760-773. doi: 10.1037/0022-3514.81.5.760

Brose, L. A., Rye, M. S., & Lutz-Zois, C. (2005). Forgiveness and personality traits. *Personality and Individual Differences, 39*, 35-46. doi: 10.1016/j.paid.2004.11.001

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology, 84*, 822-848.

Brown, R. P. (2003). Measuring individual differences in the tendency to forgive: Construct validity and links with depression. *Personality and Social Psychology Bulletin, 29*, 759-771. doi: 10.1177/0146167203029006008

Brown, R. P. (2004). Vengeance is mine: Narcissism, vengeance, and the tendency to forgive. *Journal of Research in Personality, 38*, 576-584. doi: 10.1016/j.jrp.2003.10.003

Brown, R. P., Barnes, C. D., & Campbell, N. J. (2007). Fundamentalism and forgiveness. *Personality and Individual Differences, 43*, 1437-1447. doi: 10.1016/j.paid.2007.04.025

Brown, R. P., & Phillips, A. (2005). Letting bygones be bygones: Further evidence for the validity of the Tendency to Forgive scale. *Personality and Individual Differences, 38*, 627-638. doi: 10.1016/j.paid.2004.05.017

Buhrmester, M. D., Blanton, H., & Swann Jr, W. B. (in press). Implicit self-esteem: Nature, measurement, and a new way forward. *Journal of Personality and Social Psychology*. doi: 10.1037/a0021341

Burnette, J. L., & Franiuk, R. (2009). Individual differences in implicit theories of relationships and partner fit: Predicting forgiveness in developing relationships. *Personality and Individual Differences, 48,* 144-148. doi: 10.1016/j.paid.2009.09.011

Burnette, J. L., Taylor, K. W., Worthington Jr, E. L., & Forsyth, D. R. (2007). Attachment and trait forgivingness: The mediating role of angry rumination. *Personality and Individual Differences, 42*, 1585-1596. doi: 10.1016/j.paid.2006.10.033

Caruso, E. M. (2010). When the future feels worse than the past: A temporal inconsistency in moral judgment. *Journal of Experimental Psychology: General, 139*, 610-624. doi: 10.1037/a0020757

Casper, C., Rothermund, K., & Wentura, D. (2010). Automatic stereotype activation is context dependent. *Social Psychology, 41*, 131-136. doi: 10.1027/1864-9335/a000019

Cate, K. L., Bassett, J. M., & Dabbs Jr, J. M. (2003). Fear primes may not affect women's implicit and explicit mate preferences. *The Journal of Articles in Support of the Null Hypothesis, 1*, 49-56.

Ceschi, G., Banse, R., & Van der Linden, M. (2009). Implicit but stable: Mental imagery changes explicit but not implicit anxiety. *Swiss Journal of Psychology, 68*, 213-220. doi: 10.1024/1421-0185.68.4.213

Chang, B. P. I., & Mitchell, C. J. (2009). Processing fluency as a predictor of salience asymmetries in the Implicit Association Test. *The Quarterly Journal of Experimental Psychology, 62*, 2030-2054. doi: 10.1080/17470210802651737

Choma, B. L., & Hafer, C. L. (2009). Understanding the relation between explicitly and implicitly measured political orientation: The moderating role of political sophistication. *Personality and Individual Differences, 47*, 964-967. doi: 10.1016/j.paid.2009.07.024

Cohen, A. S., Beck, M. R., Brown, L. A., & Najolia, G. M. (2009). Decoupling implicit measures of pleasant and unpleasant social attitudes. *Journal of Behavior Therapy and Experimental Psychiatry, 41,* 24-30. doi: 10.1016/j.jbtep.2009.08.007

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collier, S. A., Ryckman, R. M., Thornton, B., & Gold, J. A. (2010). Competitive personality attitudes and forgiveness of others. *Journal of Psychology: Interdisciplinary and Applied, 144*, 535-543.

Conrey, F., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology, 89*, 469-487. doi: 10.1037/0022-3514.89.4.469

Costa, P. T., & McRae, R. R. (1992). *Revised NEO personality inventory (NEO–PI-R) and NEO five-factor inventory (NEO-FFI) professional manual*. Odessa, Florida: Psychological Assessment Resources, Inc.

Coyle, C. T., & Enright, R. D. (1997). Forgiveness intervention with postabortion men. *Journal of Consulting and Clinical Psychology, 65*, 1042-1046. doi: 10.1037/0022-006X.65.6.1042

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354.

Cumming, G. (2008). Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*, 286-300. doi: 10.1111/j.1745-6924.2008.00079.x

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12*, 163-170. doi: 10.1111/1467-9280.00328

Czellar, S., & Luna, D. (2010). The effect of expertise on the relation between implicit and explicit attitude measures: An information availability/accessibility perspective. *Journal of Consumer Psychology, 20*, 259. doi: 10.1016/j.jcps.2010.06.014

Darby, B. W., & Schlenker, B. R. (1982). Children's reactions to apologies. *Journal of Personality and Social Psychology, 43*, 742-753. doi: 10.1037/0022-3514.43.4.742

Darley, J. (2002). Just punishments: Research on retributional justice. In M. Ross & D. T. Miller (Eds.), *The justice motive in everyday life*. Cambridge, UK: Cambridge University Press.

Dasgupta, N. (2010). Implicit measures of social cognition: Common themes and unresolved questions. *Journal of Psychology, 218*, 54-57. doi: 10.1027/0044-3409/a000009

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81*, 800-814. doi: 10.1037/0022-3514.81.5.800

Davis, J. R., & Gold, G. J. (2011). An examination of emotional empathy, attributions of stability, and the link between perceived remorse and forgiveness. *Personality and Individual Differences, 50*, 392-397. doi: 10.1016/j.paid.2010.10.031

Day, L., & Maltby, J. (2005). Forgiveness and social loneliness. *Journal of Psychology: Interdisciplinary and Applied, 139*, 553-555. doi: 10.3200/JRLP.139.6.553

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology, 50*, 77-85. doi: 10.1026//1618-3169.50.2.77

De Houwer, J., Geldof, T., & De Bruycker, E. (2005). The Implicit Association Test as a general measure of similarity. *Canadian Journal of Experimental Psychology, 59*, 228-239. doi: 10.1037/h0087478

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*, 347-368. doi: 10.1037/a0014211

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science, 305*, 1254-1258. doi: 10.1126/science.1100735

de Waal, F. B. M., & Pokorny, J. J. (2005). Primate conflict and its relation to human forgiveness. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 17-32). New York: Brunner-Routledge.

Dekel, R. (2010). Couple forgiveness, self-differentiation and secondary traumatization among wives of former POWs. *Journal of Social and Personal Relationships, 27*, 924-937. doi: 10.1177/0265407510377216

Denham, S. A., Neal, K., Wilson, B. J., Pickering, S., & Boyatzis, C. J. (2005). Emotional development and forgiveness in children: Emerging evidence. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 127-142). New York: Brunner-Routledge.

Dentale, F., San Martini, P., De Coro, A., & Di Pomponio, I. (2010). Alexithymia increases the discordance between implicit and explicit self-esteem. *Personality and Individual Differences, 49*, 762-767. doi: 10.1016/j.paid.2010.06.022

DeShea, L. (2003). A scenario-based scale of willingness to forgive. *Individual Differences Research, 1*, 201-217.

DeSteno, D., Dasgupta, N., Bartlett, M. Y., & Cajdric, A. (2004). Prejudice from thin air: The effect of emotion on automatic intergroup attitudes. *Psychological Science, 15*, 319-324. doi: 10.1111/j.0956-7976.2004.00676.x

Devos, T., & Banaji, M. R. (2005). American = white? *Journal of Personality and Social Psychology, 88*, 447-466. doi: 10.1037/0022-3514.88.3.447

Dislich, F. X. R., Zinkernagel, A., Ortner, T. M., & Schmitt, M. (2009). Convergence of direct, indirect, and objective risk-taking measures in gambling: The moderating role of impulsiveness and self-control. *Journal of Psychology, 218*, 20-27. doi: 10.1027/0044-3409/a000004

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*, 62-68. doi: 10.1037/0022-3514.82.1.62

Eaton, J., Struthers, C. W., & Santelli, A. G. (2006). The mediating role of perceptual validation in the repentance–forgiveness process. *Personality and Social Psychology Bulletin, 32*, 1389-1401. doi: 10.1177/0146167206291005

Effron, D. A., & Monin, B. (2010). Letting people off the hook: When do good deeds excuse transgressions? *Personality and Social Psychology Bulletin, 36*, 1618-1634. doi: 10.1177/0146167210385922

Egan, L. A., & Todorov, N. (2009). Forgiveness as a coping strategy to allow school students to deal with the effects of being bullied: Theoretical and empirical discussion. *Journal of Social and Clinical Psychology, 28*, 198-222. doi: 10.1521/jscp.2009.28.2.198

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

Egloff, B., & Schmukle, S. C. (2003). Does social desirability moderate the relationship between implicit and explicit anxiety measures? *Personality and Individual Differences, 35*, 1697-1706. doi: 10.1016/S0191-8869(02)00391-4

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage and coding. *Semiotica, 1*, 49-98.

Emmons, R. A. (2000). Personality and forgiveness. In M. E. McCullough, K. I. Pargament & C. E. Thoreson (Eds.), *Forgiveness: Theory, research, and practice* (pp. 156-175). New York: The Guilford Press.

Enright, R. D., & Coyle, C. T. (1998). Researching the process model of forgiveness within psychological interventions. In E. L. Worthington Jr (Ed.), *Dimensions of forgiveness: Psychological research & theological forgiveness* (pp. 139-161). Philadelphia: Templeton Foundation Press.

Enright, R. D., & Fitzgibbons, R. P. (2000). *Helping clients forgive*. Washington: American Psychological Association.

Enright, R. D., Freedman, S., & Rique, J. (1998). The psychology of interpersonal forgiveness. In R. D. Enright & J. North (Eds.), *Exploring forgiveness* (pp. 46-62). Madison, US: University of Wisconsin Press.

Enright, R. D., & Gassin, E. A. (1992). Forgiveness: A developmental view. *Journal of Moral Education, 21*, 99-114.

Enright, R. D., & Group, T. H. D. S. (1991). The moral development of forgiveness. In W. Kurtines & J. Gerwirtz (Eds.), *Handbook of moral behavior and development* (Vol. 1). Hillsdale, NJ: Lawrence Erlbaum.

Enright, R. D., & Group, T. H. D. S. (1994). Piaget on the moral development of forgiveness: Identity or reciprocity? *Human Development, 37*, 63-80. doi: 10.1159/000278239

Enright, R. D., & North, J. (1998a). Introducing forgiveness. In R. D. Enright & J. North (Eds.), *Exploring forgiveness*. Madison, Wisconsin: The University of Wisconsin Press.

Enright, R. D., & North, J. (Eds.). (1998b). *Exploring forgiveness*. Madison, US: The University of Wisconsin Press.

Enright, R. D., & Zell, R. L. (1989). Problems encountered when we forgive one another. *Journal of Psychology and Christianity, 8*, 52-60.

Exline, J. J., & Baumeister, R. F. (2000). Expressing forgiveness and repentance: benefits and barriers. In M. E. McCullough, K. I. Pargament & C. E. Thoreson (Eds.), *Forgiveness: Theory, research and practice* (pp. 133-155). New York: The Guilford Press.

Exline, J. J., Baumeister, R. F., Bushman, B. J., Campbell, W. K., & Finkel, E. J. (2004). Too proud to let go: Narcissistic entitlement as a barrier to forgiveness. *Journal of Personality and Social Psychology, 87*, 894-912. doi: 10.1037/0022-3514.87.6.894

Exline, J. J., Deshea, L., & Holeman, V. T. (2007). Is apology worth the risk? Predictors, outcomes, and ways to avoid regret. *Journal of Social and Clinical Psychology, 26*, 479-504. doi: 10.1521/jscp.2007.26.4.479

Exline, J. J., Worthington Jr, E. L., Hill, P., & McCullough, M. E. (2003). Forgiveness and justice: A research agenda for social and personality psychology. *Personality and Social Psychology Review, 7*, 337-348. doi: 10.1207/S15327957PSPR0704_06

Farrow, T. F. D., & Woodruff, P. W. R. (2005). Neuroimaging of forgivability. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 259-272). New York: Brunner-Routledge.

Farrow, T. F. D., Zheng, Y., Wilkinson, I. D., Spence, S. A., Deakin, J. F. W., Tarrier, N., et al. (2001). Investigating the functional anatomy of empathy and forgiveness. *NeuroReport, 12*, 2433-2438. doi: 10.1097/00001756-200108080-00029

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology, 23*, 75-109.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*, 1013-1027. doi: 10.1037/0022-3514.69.6.1013

Fazio, R. H., & Olsen, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297-327. doi: 10.1146/annurev.psych.54.101601.145225

Feather, N. T. (2002). Deservingness, entitlement, and reactions to outcomes. In M. Ross & D. T. Miller (Eds.), *The justice motive in everyday life* (pp. 334-349). Cambridge, UK: Cambridge University Press.

Fehr, R., & Gelfand, M. J. (2010). When apologies work: How matching apology components to victims' self-construals facilitates forgiveness. *Organizational Behavior and Human Decision Processes, 113*, 37-50. doi: 10.1016/j.obhdp.2010.04.002

Fehr, R., Gelfand, M. J., & Nag, M. (2010). The road to forgiveness: A meta-analytic synthesis of its situational and dispositional correlates. *Psychological Bulletin, 136*, 894-914. doi: 10.1037/a0019993

Festinger, L. (1954). A theory of social comparison process. *Human Relations, 7*, 117-140. doi: 10.1177/001872675400700202

Fincham, F. D. (2000). The kiss of the porcupines: From attributing responsibility to forgiving. *Personal Relationships, 7*, 1-23. doi: 10.1111/j.1475-6811.2000.tb00001.x

Fincham, F. D. (2009). *Forgiveness: Integral to a science of close relationships?* Paper presented at the Inaugural Herzliya Symposium on Personality and Social Psychology, Herzliya, Israel.

Fincham, F. D., Beach, S. R. H., & Davila, J. (2004). Forgiveness and conflict resolution in marriage. *Journal of Family Psychology, 18*, 72-81. doi: 10.1037/0893-3200.18.1.72

Fincham, F. D., Jackson, H., & Beach, S. R. H. (2005). Transgression severity and forgiveness: Different moderators for objective and subjective severity. *Journal of Social and Clinical Psychology, 24*, 860-875. doi: 10.1521/jscp.2005.24.6.860

Finkel, E. J., Rusbult, C. E., Kumashiro, M., & Hannon, P. A. (2002). Dealing with betrayal in close relationships: Does commitment promote forgiveness? *Journal of Personality and Social Psychology, 82*, 956-974. doi: 10.1037/0022-3514.82.6.956

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20*, 303-315. doi: 10.1234/12345678

Fisher, R. J., & Katz, J. E. (2000). Social-desirability bias and the validity of self-reported values. *Psychology and Marketing, 17*, 105-120.

Fitness, J., & Peterson, J. (2008). Punishment and forgiveness in close relationships: An evolutionary, social-psychological perspective. In J. P. Forgas & J. Fitness (Eds.), *Social relationships: Cognitive, affective, and motivational processes* (pp. 255-269). New York, NY, US: Psychology Press.

Flanigan, B. (1998). Forgivers and the unforgivable. In R. D. Enright & J. North (Eds.), *Exploring forgiveness*. Madison, Wisconsin: The University of Wisconsin Press.

Fox, E., Lester, V., Russo, R., Bowles, R. J., Pichler, A., & Dutton, K. (2000). Facial expressions of emotion: Are angry faces detected more efficiently? *Cognition and Emotion, 14*, 61-92. doi: 10.1080/026999300378996

Frantz, C. M., & Benningson, C. (2005). Better late than early: The influence of timing on apology effectiveness. *Journal of Experimental Social Psychology, 41*, 201-207. doi: 10.1016/j.jesp.2004.07.007

Freedman, S., Enright, R. D., & Knutson, J. (2005). A progress report on the process model of forgiveness. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 393-406). New York: Brunner-Routledge.

Frey-Rohn, L. (1974). *From Freud to Jung: A comparative study of the psychology of the unconscious* (F. E. Engreen & E. K. Engreen, Trans.). New York: G. P. Putnam's sons for the C. G. Jung Foundation.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*, 379-390. doi: 10.1037/1082-989X.1.4.379

Friedberg, J. P., Suchday, S., & Srinivas, V. S. (2009). Relationship between forgiveness and psychological and physiological indices in cardiac patients. *International Journal of Behavioral Medicine*, 1-7. doi: 10.1007/s12529-008-9016-2

Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology, 19*, 285-338. doi: 10.1080/10463280802556958

Friese, M., Hofmann, W., & Wänke, M. (2008). When impulses take over: Moderated predictive validity of implicit and explicit attitude measures in predicting food choice and consumption behavior. *British Journal of Social Psychology, 47*, 397-419. doi: 10.1348/014466607X241540

Friese, M., Hofmann, W., & Wänke, M. (2009). The impulsive consumer: Predicting consumer behavior with implicit reaction time measures. In M. Wänke (Ed.), *Frontiers of social psychology: The social psychology of consumer behavior* (pp. 335-364). New York: Psychology Press.

Friese, M., Wänke, M., & Plessner, H. (2006). Implicit consumer preferences and their influence on product choice. *Psychology and Marketing, 23*, 727-740. doi: 10.1002/mar.20126

Frise, N. R., & McMinn, M. R. (2010). Forgiveness and reconciliation: The differing perspectives of psychologists and Christian theologians. *Journal of Psychology and Theology, 38*, 83-90.

Friesen, M. D., & Fletcher, G. J. O. (2007). Exploring the lay representation of forgiveness: Convergent and discriminant validity. *Personal Relationships, 14*, 209-223. doi: 10.1111/j.1475-6811.2007.00151.x

Frost, R., & Untermeyer, L. (1963). *The Letters of Robert Frost to Louis Untermeyer*. New York: Holt, Rinehart and Winston.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *British Medical Journal, 292*, 746-750.

Gassin, E. A., Enright, R. D., & Knutson, J. (2005). Bringing peace to the central city: Forgiveness education in Milwaukee. *Theory Into Practice, 44*, 319-328. doi: 10.1207/s15430421tip4404_5

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692-731. doi: 10.1037/0033-2909.132.5.692

Gawronski, B., & Bodenhausen, G. V. (2007). Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE model. *Social Cognition, 25*, 687-717. doi: 10.1521/soco.2007.25.5.687

Gawronski, B., Peters, K. R., & LeBel, E. P. (2008). What makes mental associations personal or extra-personal? Conceptual issues in the methodological debate about implicit attitude measures. *Social and Personality Psychology Compass, 2*, 1002-1023. doi: 10.1111/j.1751-9004.2008.00085.x

Gawronski, B., Strack, F., & Bodenhausen, G. V. (2008). Attitudes and cognitive consistency: The role of associative and prepositional processes. In R. E. Petty, R. H. Fazio & P. Briñol (Eds.) *Attitudes: Insights from the new implicit measures* (pp. 85-117). New York: Psychology Press.

Gonzales, M. H., Manning, D. J., & Haugen, J. A. (1992). Explaining our sins: Factors influencing offender accounts and anticipated victim responses. *Journal of Personality and Social Psychology, 62*, 958-971. doi: 10.1037/0022-3514.62.6.958

Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics, 32*, 148-170.

Gorsuch, R. L., & Hao, J. Y. (1993). Forgiveness: An exploratory factor analysis and its relationships to religious variables. *Review of Religious Research, 34*, 333-348.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*, 93-104. doi: 10.1037/0003-066X.59.2.93

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology, 40*, 357-365. doi: 10.1016/j.jesp.2003.07.002

Gray, N. S., MacCulloch, M. J., Smith, J., Morris, M., & Snowden, R. J. (2003). Forensic psychology: Violence viewed by psychopathic murderers. *Nature, 423*, 497-498.

Green, J. D., Burnette, J. L., & Davis, J. L. (2008). Third-party forgiveness: (Not) forgiving your close other's betrayer. *Personality and Social Psychology Bulletin, 34*, 407-418. doi: 10.1177/0146167207311534

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20.

Greenwald, A. G. (1990). What cognitive representations underlie social attitudes? *Bulletin of the Psychonomic Society, 28*, 254-260.

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

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*, 3-25. doi: 10.1037/0033-295X.109.1.3

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology, 79*, 1022-1038. doi: 10.1037/0022-3514.79.6.1022

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

Greenwald, A. G., & Nosek, B. A. (2008). Attitudinal dissociation: What does it mean? In R. E. Petty, R. H. Fazio & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 65-82). New York: Psychology Press.

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*, 197-216. doi: 10.1037/0022-3514.85.2.197

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

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

Güth, W., Ockenfels, P., & Wendel, M. (1997). Cooperation based on trust: An experimental investigation. *Journal of Economic Psychology, 18*, 15-43. doi: 10.1016/S0167-4870(96)00045-1

Han, H. A., Czellar, S., Olsen, M. A., & Fazio, R. H. (2010). Malleability of attitudes or malleability of the IAT? *Journal of Experimental Social Psychology, 46*, 286-298. doi: 10.1016/j.jesp.2009.11.011

Hannon, P. A., Rusbult, C. E., Finkel, E. J., & Kamashiro, M. (2010). In the wake of betrayal: Amends, forgiveness, and the resolution of betrayal. *Personal Relationships, 17,* 253-278. doi: 10.1111/j.1475-6811.2010.01275.x

Harris, A. H. S., & Thoreson, C. E. (2005). Forgiveness, unforgiveness, health, and disease. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 321-333). New York: Brunner-Routledge.

Haselhuhn, M. P., Schweitzer, M. E., & Wood, A. M. (2010). How implicit beliefs influence trust recovery. *Psychological Science, 21*, 645-648.

Hayashi, A., Abe, N., Ueno, A., Shigemune, Y., Mori, E., Tashiro, M., et al. (2010). Neural correlates of forgiveness for moral transgressions involving deception. *Brain Research, 1332*, 90-99. doi: 10.1016/j.brainres.2010.03.045

Hebl, J. H., & Enright, R. D. (1993). Forgiveness as a psychotherapeutic goal with elderly females. *Psychotherapy: Theory, Research, Practice, Training, 30*, 658-667. doi: 10.1037/0033-3204.30.4.658

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

Hofmann, W., Gschwendner, T., Nosek, B. A., & Schmitt, M. (2005). What moderates implicit–explicit consistency? *European Journal of Social Psychology, 16*, 335-390. doi: 10.1080/10463280500443228

Hofmann, W., Gschwendner, T., & Schmitt, M. (2005). On implicit-explicit consistency: The moderating role of individual differences in awareness and adjustment. *European Journal of Personality, 19*, 25-49. doi: 10.1002/per.537

Hofmann, W., & Schmitt, M. (2008). Advances and challenges in the indirect measurement of individual differences at age 10 of the Implicit Association Test. *European Journal of Psychological Assessment, 24*, 207-209. doi: 10.1027/1015-5759.24.4.207

Holden, R. R. (2008). Underestimating the effects of faking on the validity of self-report personality scales. *Personality and Individual Differences, 44*, 311-321. doi: 10.1016/j.paid.2007.08.012

Holmes, J. G. (2002). Interpersonal expectations as the building blocks of social cognition: An Interdependence Theory perspective. *Personal Relationships, 9*, 1-26. doi: 10.1111/1475-6811.00001

Holtgraves, T., Srull, T. K., & Socall, D. (1989). Conversation memory: The effects of speaker status on memory for the assertiveness of conversation remarks. *Journal of Personality and Social Psychology, 56*, 149-160. doi: 10.1037/0022-3514.56.2.149

Houben, K. (2007). *Decoding the Alcohol-IAT: The Implicit Association Test as a measure of individual differences in implicit preferences for alcohol* (Unpublished doctoral dissertation). Maastricht University, Maastricht, Germany.

Houben, K., Nosek, B. A., & Wiers, R. W. (2010). Seeing the forest through the trees: A comparison of different IAT variants measuring implicit alcohol associations. *Drug and Alcohol Dependence, 106*, 204-211. doi: 10.1016/j.drugalcdep.2009.08.016

Houben, K., & Wiers, R. W. (2008). Measuring implicit alcohol associations via the Internet: Validation of web-based implicit association tests. *Behavior Research methods, 40*, 1134-1143. doi: 10.3758/BRM.40.4.1134

Hoyt, W. T., Fincham, F. D., McCullough, M. E., Maio, G., & Davila, J. (2005). Responses to interpersonal transgressions in families: Forgivingness, forgivability, and relationship-specific effects. *Journal of Personality and Social Psychology, 89*, 375-394. doi: 10.1037/0022-3514.89.3.375

Hoyt, W. T., & McCullough, M. E. (2005). Issues in the multimodal measurement of forgiveness. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 109-124). New York: Brunner-Routledge.

Hu, S.-m., Zhang, A.-q., & Ja, Y.-j. (2005). A study on interpersonal forgive and revenge of undergraduates. *Chinese Journal of Clinical Psychology, 13*, 55-57.

Hui, E. K. P., & Chau, T. S. (2009). The impact of a forgiveness intervention with Hong Kong Chinese children hurt in interpersonal relationships. *British Journal of Guidance and Counselling, 37*, 141-156. doi: 10.1080/03069880902728572

Humphreys, M. S., Tangen, J. M., Cornwell, B., Quinn, E., & Murray, K. L. (2010). Unintended effects of memory on decision making: A breakdown in access control. *Journal of Memory and Language, 63* 400-415. doi: 10.1016/j.jml.2010.06.006

Huntsinger, J. R., & Smith, C. T. (2009). First thought, best thought: Positive mood maintains and negative mood degrades implicit-explicit attitude correspondence. *Personality and Social Psychology Bulletin, 35*, 11. doi: 10.1177/0146167208327000

Hyde, A. L., Doerksen, S. E., Ribeiro, N. F., & Conroy, D. E. The independence of implicit and explicit attitudes toward physical activity: Introspective access and attitudinal concordance. *Psychology of Sport and Exercise, 11*, 387-393. doi: 10.1016/j.psychsport.2010.04.008

Ibáñez, A., Gleichgerrcht, E., Hurtado, E., González, R., Haye, A., & Manes, F. F. (2010). Early neural markers of implicit attitudes: N170 modulated by intergroup and evaluative contexts in IAT. *Frontiers in Human Neuroscience*, *4,* 1-14. doi: 10.3389/fnhum.2010.00188

Johnson, J. L., Kim, L. M., Giovannelli, T. S., & Cagle, T. (2010). Reinforcement sensitivity theory, vengeance, and forgiveness. *Personality and Individual Differences, 48*, 612-616. doi: 10.1016/j.paid.2009.12.018

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers, 31*, 433-438

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76*, 349-364. doi: 10.1037/h0031617

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality and Social Psychology, 85*, 969-978. doi: 10.1037/0022-3514.85.5.969

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., et al. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of 10 studies that no manager should ignore. *Research in Organizational Behavior, 29,* 39-69.

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology, 41*, 137-146. doi: 10.1027/1864-9335/a000020

Jung, K. H., & Lee, J.-H. (2009). Implicit and explicit attitude dissociation in spontaneous deceptive behavior. *Acta Psychologica, 132*, 62-67. doi: 10.1016/j.actpsy.2009.06.004

Kadima, K. J., Gauché, M., Vinsonneau, G., & Mullet, E. (2007). Conceptualizations of forgiveness: Collectivist-Congolese versus individualist-French viewpoints. *Journal of Cross-Cultural Psychology, 38*, 432-437. doi: 0.1177/0022022107302312

Kanz, J. E. (2000). How do people conceptualize and use forgiveness? The Forgiveness Attitudes Questionnaire. *Counseling and Values, 44*, 174-188

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81*, 774-788. doi: 10.1037/0022-3514.81.5.774

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*, 16-32.doi: 10.1037/0022-3514.91.1.16

Karremans, J. C., & Aarts, H. (2007). The role of automaticity in determining the inclination to forgive close others. *Journal of Experimental Social Psychology, 43*, 902-917. doi: 10.1016/j.jesp.2006.10.012

Karremans, J. C., & Van Lange, P. A. M. (2005). Does activating justice help or hurt in promoting forgiveness? *Journal of Experimental Social Psychology, 41*, 290-297. doi: 10.1016/j.jesp.2004.06.005

Karremans, J. C., & Van Lange, P. A. M. (2008). Forgiveness in personal relationships: Its malleability and powerful consequences. *European Review of Social Psychology, 19*, 202-241. doi: 10.1080/10463280802402609

Karremans, J. C., Van Lange, P. A. M., & Holland, R. W. (2005). Forgiveness and its associations with prosocial thinking, feeling, and doing beyond the relationship with the offender. *Personality and Social Psychology Bulletin, 31*, 1315-1326. doi: 10.1177/0146167205274892

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions

through approach behaviors. *Journal of Personality and Social Psychology, 92*, 957-971. doi: 10.1037/0022-3514.92.6.957

Kearns, J. N., & Fincham, F. D. (2004). A prototype analysis of forgiveness. *Personality and Social Psychology Bulletin, 30*, 838-855. doi: 10.1177/0146167204264237

Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations: a theory of interdependence*. New York: John Wiley & Sons, Inc.

Kelln, B. R. C., & Ellard, J. H. (1999). An equity theory analysis of the impact of forgiveness and retribution on transgressor compliance. *Personality and Social Psychology Bulletin, 25*, 864-872. doi: 10.1177/0146167299025007008

Kim, D.-Y. (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly, 66*, 83-96. doi: 10.2307/3090143

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology, 89*, 104-118. doi: 10.1037/0021-9010.89.1.104

Kinoshita, S., & Peek-O'leary, M. (2006). Two bases of the compatibility effect in the Implicit Association Test (IAT). *The Quarterly Journal of Experimental Psychology, 59*, 2102-2120. doi: 0.1080/17470210500451141

Klauer, K. C., & Mierke, J. (2005). Task-set inertia, attitude accessibility, and compatibility-order effects: New evidence for a task-set switching account of the Implicit Association Test effect. *Personality and Social Psychology Bulletin, 31*, 208-210. doi: 10.1177/0146167204271416

Korsgaard, M. A., Brodt, S. E., & Whitener, E. M. (2002). Trust in the face of conflict: Role of managerial trustworthy behavior and organizational context. *Journal of Applied Psychology, 87*, 312-319. doi: 10.1037/0021-9010.87.2.312

Krause, S., Back, M. D., Egloff, B., & Schmukle, S. C. (2010). Reliability of implicit self-esteem measures revisited. *European Journal of Personality*. doi: 10.1002/per.792

Lamb, S. (2006). Forgiveness, women, and responsibility to the group. *Journal of Human Rights, 5*, 45-60.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the Implicit Association Test IV: What we know (so far) about the method. In B. Wittenbrink & N. S. Schwarz (Eds.), *Implicit measures of attitudes: Procedures and controversies* (pp. 59-102). New York: Guilford Press.

Lawler-Row, K. A., Karremans, J. C., Scott, C., Edlis-Matityahou, M., & Edwards, L. (2008). Forgiveness, physiological reactivity and health: The role of anger. *International Journal of Psychophysiology, 68*, 51-58.

Leach, M. M., Greer, T., & Gaughf, J. (2009). Linguistic analysis of interpersonal forgiveness: Process trajectories. *Personality and Individual Differences*, *48,* 117-122. doi:10.1016/j.paid.2009.09.005

Lerner, M. (1980). *The belief in a just world: A fundamental delusion*. New York: Plenum Press.

Lerner, M. (2002). Pursuing the justice motive. In M. Ross & D. T. Miller (Eds.), *The justice motive in everyday life* (pp. 10-40). Cambridge, UK: Cambridge University Press.

Loo, R., & Loewen, P. (2004). Confirmatory factor analyses of scores from full and short versions of the Marlowe–Crowne social desirability scale. *Journal of Applied Social Psychology, 34*, 2343-2352. doi: 10.1111/j.1559-1816.2004.tb01980.x

Lount, R. B., Zhong, C.-B., Sivanathan, N., & Murninghan, J. K. (2008). Getting off on the wrong foot: The timing of a breach and the restoration of trust. *Personality and Social Psychology Bulletin, 34*, 1601-1612. doi: 10.1177/0146167208324512

Lucas, T., Young, J. D., Zhdanova, L., & Alexander, S. (2010). Self and other justice beliefs, impulsivity, rumination, and forgiveness: Justice beliefs can both prevent and promote forgiveness. *Personality and Individual Differences, 49*, 851-856. doi: 10.1016/j.paid.2010.07.014

Luchies, L. B., Finkel, E. J., McNulty, J. K., & Kumashiro, M. (2010). The doormat effect: When forgiving erodes self-respect and self-concept clarity. *Journal of Personality and Social Psychology, 98*, 734-749. doi: 10.1037/a0017838

Macaskill, A. (2007). Exploring religious involvement, forgiveness, trust, and cynicism. *Mental Health, Religion & Culture, 10*, 203-218. doi: 10.1080/13694670600616092

Maison, D., Greenwald, A. G., & Bruin, R. H. (2004). Predictive validity of the Implicit Association Test in studies of brands, consumer attitudes and behavior. *Journal of Consumer Psychology, 14*, 405-415. doi: 10.1207/s15327663jcp1404_9

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37*, 435-442. doi: 10.1006/jesp.2000.1470

McConnell, A. R., & Leibold, J. M. (2009). Weak criticisms and selective evidence: Reply to Blanton et al. *Journal of Applied Psychology, 94*, 583–589. doi: 10.1037/a0014649

McCullough, M. E. (2008). *Beyond revenge: The evolution of the forgiveness instinct* (1st ed.). San Francisco: Jossey-Bass.

McCullough, M. E., Bono, G., & Root, L. M. (2007). Rumination, emotion, and forgiveness: Three longitudinal studies. *Journal of Personality and Social Psychology, 92*, 490-505. doi: 10.1037/0022-3514.92.3.490

McCullough, M. E., & Hoyt, W. T. (2002). Transgression-related motivational dispositions: Personality substrates of forgiveness and their links to the Big Five. *Personality and Social Psychology Bulletin, 28*, 1556-1573. doi: 10.1177/014616702237583

McCullough, M. E., Hoyt, W. T., & Rachal, K. C. (2000). What we know (and need to know) about assessing forgiveness constructs. In M. E. McCullough, K. I. Pargament & C. E. Thoreson (Eds.), *Forgiveness: Theory, research, and practice* (pp. 65-88). New York: The Guilford Press.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2010). Evolved mechanisms for revenge and forgiveness. In P. R. Shaver & M. Mikulincer (Eds.), *Understanding and reducing aggression, violence, and their consequences.* Washington DC: American Psychological Association.

McCullough, M. E., Pargament, K. I., & Thoreson, C. E. (2000a). The psychology of forgiveness: History, conceptual issues, and overview. In M. E. McCullough, K. I. Pargament & C. E. Thoreson (Eds.), *Forgiveness: Theory, research, and practice.* New York: The Guilford Press.

McCullough, M. E., Pargament, K. I., & Thoreson, C. E. (Eds.). (2000b). *Forgiveness: Theory, research, and practice*. New York: The Guilford Press.

McCullough, M. E., Rachal, K. C., Sandage, S. J., Worthington Jr, E. L., Wade Brown, S., & Hight, T. L. (1998). Interpersonal forgiving in close relationships: II. Theoretical elaboration and measurement. *Journal of Personality and Social Psychology, 75*, 1586-1603. doi: 10.1037/0022-3514.75.6.1586

McCullough, M. E., & Root, L. M. (2005). Forgiveness as change. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 91-108). New York: Brunner-Routledge.

McCullough, M. E., Root, L. M., & Cohen, A. D. (2006). Writing about the benefits of an interpersonal transgression facilitates forgiveness. *Journal of Consulting and Clinical Psychology, 74*, 887-897. doi: 10.1037/0022-006X.74.5.887

McCullough, M. E., Root, L. M., Tabak, B. A., & Witvliet, C. (2009). Forgiveness. In S. J. Lopez (Ed.), *Handbook of positive psychology* (2nd ed., pp. 427-435). New York: Oxford.

McCullough, M. E., Root, L. M., Berry, J. W., Tabak, B. A., & Bono, G. (2010). On the form and function of forgiving: Modeling the time-forgiveness relationship and testing the valuable relationships hypothesis. *Emotion, 10*, 358-376. doi: 10.1037/a0019349

McCullough, M. E., & Worthington Jr, E. L. (1994). Encouraging clients to forgive people who have hurt them: Review, critique, and research prospectus. *Journal of Psychology and Theology, 22*, 3-20.

McCullough, M. E., & Worthington Jr, E. L. (1999). Religion and the forgiving personality. *Journal of Personality, 67*, 1141-1164. doi: 10.1111/1467-6494.00085

McCullough, M. E., Worthington Jr, E. L., & Rachal, K. C. (1997). Interpersonal forgiving in close relationships. *Journal of Personality and Social Psychology, 73*, 321-336. doi: 10.1037/0022-3514.73.2.321

McKee, I. R., & Feather, N. T. (2008). Revenge, retribution, and values: Social attitudes and punitive sentencing. *Social Justice Research, 21*, 138-163. doi: 10.1007/s11211-008-0066-z

Meade, A. W. (2009). FreeIAT: An open-source program to administer the implicit association test. *Applied Psychological Measurement*, *33*, 643-643. doi: 10.1177/0146621608327803

Meek, K. R., Albright, J. S., & McMinn, M. R. (1995). Religious orientation, guilt, confession, and forgiveness. *Journal of Psychology and Theology, 23,* 190-197.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology, 85*, 1180-1113. doi: 10.1037/0022-3514.85.6.1180

Mineka, S., & Öhman, A. (2002). Phobias and preparedness: The selective, automatic, and encapsulated nature of fear. *Biological Psychiatry, 52*, 927-937. doi: 10.1016/S0006-3223(02)01669-4

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology, 132*, 455-469. doi: 10.1037/0096-3445.132.3.455

Molesworth, B. R. C., & Chang, B. P. I. (2009). Risk management: An implicit measure designed to predict pilots' risk-taking behaviour through an implicit association test. *Human Factors, 51*, 845-857. doi: 10.1177/0018720809357756

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132*, 297-326. doi: 10.1037/0033-2909.132.2.297

Mullet, E., Neto, F., & Riviere, S. (2005). Personality and its effects on resentment, revenge, forgiveness, and self-forgiveness. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 159-181). New York: Brunner-Routledge.

Murphy, J. G. (2005). Forgiveness, self-respect, and the value of resentment. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 33-40). New York: Brunner-Routledge.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*, 263-280. doi: 10.1002/ejsp.2420150303

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes and Intergroup Relations, 8*, 39-52. doi: 10.1177/1368430205048615

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259. doi: 10.1037/0033-295X.84.3.231

Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. *Psychological Science, 21*, 511-517.

North, J. (1998). The "ideal" of forgiveness: A philosopher's exploration. In R. D. Enright & J. North (Eds.), *Exploring forgiveness*. Madison, Wisconsin: The University of Wisconsin Press.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134*, 565-584. doi: 10.1037/0096-3445.134.4.565

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19*, 625-664. doi: 10.1521/soco.19.6.625.20886

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

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002b). Math = male, me = female, therefore math is not equal to me. *Journal of Personality and Social Psychology, 83*, 44-59. doi: 10.1037/0022-3514.83.1.44

Nosek, B. A., & Greenwald, A. G. (2009). (Part of) the case for a pragmatic approach to validity: Comment on De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009). *Psychological Bulletin, 135*, 373-376. doi: 10.1037/a0015047

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and Using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin., 31*, 166-180. doi: 10.1177/0146167204271418

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at Age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265-292): Psychology Press.

Nosek, B. A., & Hansen, J. (2008a). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion, 22*, 553-594. doi: 10.1080/02699930701438186

Nosek, B. A., & Hansen, J. (2008b). Personalizing the Implicit Association Test increases explicit evaluation of target concepts. *European Journal of Personality Assessment, 24*, 226-236. doi: 10.1027/1015-5759.24.4.226

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test: Implicit and explicit attitudes are related but distinct constructs. *Experimental Psychology, 54*, 14-29. doi: 10.1027/1618-3169.54.1.14

Öhman, A. (2005). The role of the amygdala in human fear: Automatic detection of threat. *Psychoneuroendocrinology, 30*, 953-958. doi: 10.1016/j.psyneuen.2005.03.019

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of experimental Psychology: General, 130*, 466-478. doi: 10.1037/0096-3445.130.3.466

Olsen, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14*, 636-639. doi: 10.1046/j.0956-7976.2003.psci_1477.x

Olsen, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86*, 653-667. doi: 10.1037/0022-3514.86.5.653

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaires*. Berlin: Springer-Verlag.

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94*, 16-31. doi: 10.1037/0022-3514.94.1.16

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277-293. doi: 10.1037/0022-3514.89.3.277

Penke, L., Eichstaedt, J., & Asendorpf, J. B. (2006). Single-attribute implicit association tests (SA-IAT) for the assessment of unipolar constructs. *Experimental Psychology, 53*, 283-291. doi: 10.1027/1618-3169.53.4.283

Pérez, E. (2010). Explicit evidence on the import of implicit attitudes: The IAT and immigration policy judgments. *Political Behavior, 32*, 517-545. doi: 10.1007/s11109-010-9115-z

Perkins, A. W., & Forehand, M. R. (2006). Decomposing the implicit self-concept: The relative influence of semantic meaning and valence on attribute self-association. *Social Cognition, 24*, 387-408. doi: 10.1521/soco.2006.24.4.387

Perugini, M. (2005). Predictive models of implicit and explicit attitudes. *British Journal of Social Psychology, 44*, 29-45. doi: 10.1348/014466604X23491

Perugini, M., & Leone, L. (2009). Implicit self-concept and moral action. *Journal of Research in Personality, 43*, 747-754. doi: 10.1016/j.jrp.2009.03.015

Pinter, B., & Greenwald, A. G. (2005). Clarifying the role of the "other" category in the self-esteem IAT. *Experimental Psychology, 52*, 74-79. doi: 10.1027/1618-3169.52.1.74

Poehlman, T. A., Uhlmann, E. L., Greenwald, A. G., & Banaji, M. R. (2005). *Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity*. Unpublished manuscript.

Powers, C., Nam, R. K., Rowatt, W. C., & Hill, P. (2007). Associations between humility, spiritual transcendence, and forgiveness. *Research in the Social Scientific Study of Religion, 18*, 75-94. doi: 10.1163/ej.9789004158511.i-301.32

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*, 380-391. doi: 10.1037/0022-3514.61.3.380

Price, R. B., Nock, M. K., Charney, D. S., & Mathew, S. J. (2009). Effects of intravenous ketamine on explicit and implicit measures of suicidality in treatment-resistant depression. *Biological Psychiatry, 66*, 522–526. doi: 10.1016/j.biopsych.2009.04.029

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2009). Does unconscious racial bias affect trial judges? *Notre Dame Law Review, 84*, 1195-1246.

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology, 44*, 386-396. doi: 10.1016/j.jesp.2006.12.008

Ratliff, K. A., & Nosek, B. A. (2010). Creating distinct implicit and explicit attitudes with an illusory correlation paradigm. *Journal of Experimental Social Psychology, 46*, 721-728. doi: 10.1016/j.jesp.2010.04.011

Redker, C. M., & Gibson, B. (2009). Music as an unconditioned stimulus: Positive and negative effects of country music on implicit attitudes, explicit attitudes, and brand choice. *Journal of Applied Social Psychology, 39*, 2689-2705. doi: 10.1111/j.1559-1816.2009.00544.x

Reed, A., & Aquino, K. (2003). Moral identity and the expanding circle of moral regard toward out-groups. *Journal of Personality and Social Psychology, 84*, 1270-1286. doi: 10.1037/0022-3514.84.6.1270

Reed, G. L., & Enright, R. D. (2006). The effects of forgiveness therapy on depression, anxiety, and posttraumatic stress for women after spousal emotional abuse. *Journal of Consulting and Clinical Psychology, 74*, 920-929. doi: 10.1037/0022-006X.74.5.920

Richetin, J., Richardson, D. S., & Mason, G. D. (2010). Predictive validity of IAT aggressiveness in the context of provocation. *Social Psychology, 41*, 27-34. doi: 10.1027/1864-9335/a000005

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*, 201-212. doi: 10.1207/s15324834basp2703_2

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17*, 523-534. doi: 10.1016/j.labeco.2009.04.005

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192-233. doi: 10.1037/0096-3445.104.3.192

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hillsdale: Lawrence Erlbaum Associates.

Rothermund, K., & Wentura, D. (2001). Figure–ground asymmetries in the Implicit Association Test (IAT). *Zeitschrift fur Experimentelle Psychologie, 48*, 94-106. doi: 10.1026//0949-3946.48.2.94

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General, 133*, 139-165. doi: 10.1037/0096-3445.133.2.139

Rothermund, K., Wentura, D., & De Houwer, J. (2005). Validity of the salience asymmetry account of the Implicit Association Test: Reply to Greenwald, Nosek, Banaji, and Klauer (2005). *Journal of Experimental Psychology: General, 134*, 426-430. doi: 10.1037/0096-3445.134.3.426

Rudolph, A., Schröder-Abé, M., Riketta, M., & Schütz, A. (2010). Easier when done than said! Implicit self-esteem predicts observed or spontaneous behavior, but not self-reported or controlled behavior. *Zeitschrift fur Psychologie, 218*, 12-19. doi: 10.1027/0044-3409/a000003

Rudolph, A., Schröder-Abé, M., Schütz, A., Gregg, A. P., & Sedikides, C. (2008). Through a glass, less darkly? Reassessing convergent and discriminant validity in measures of implicit self-esteem. *European Journal of Psychological Assessment, 24*, 273-281. doi: 10.1027/1015-5759.24.4.273

Rusbult, C. E., Hannon, P. A., Stocker, S. L., & Finkel, E. J. (2005). Forgiveness and relational repair. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 185-205). New York: Brunner-Routledge.

Rydell, R. J., & Gawronski, B. (2009). I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. *Cognition and Emotion, 23*, 1118-1152. doi: 10.1080/02699930802355255

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*, 995-1008. doi: 10.1037/0022-3514.91.6.995

Rye, M. S., Loiacono, D. M., Folck, C. D., Olszewski, B. T., Heim, T. A., & Madia, B. P. (2001). Evaluation of the psychometric properties of two forgiveness scales. *Current Psychology, 2*, 260-277. doi: 0.1007/s12144-001-1011-6

Rye, M. S., Pargament, K. I., Ali, M. A., Beck, G. L., Dorff, E. N., Hallisey, C., et al. (2000). Religious perspective on forgiveness. In M. E. McCullough, K. I. Pargament & C. E. Thoreson (Eds.), *Forgiveness: Theory, research, and practice* (pp. 17-40). New York: The Guilford Press.

Saling, L. L., & Phillips, J. G. (2007). Automatic behaviour: Efficient not mindless. *Brain Research Bulletin, 73*, 1-20. doi: 10.1016/j.brainresbull.2007.02.009

Sandage, S. J., & Williamson, I. (2005). Forgiveness in cultural context. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 41-56). New York: Brunner-Routledge.

Scarabis, M., Florack, A., & Gosejohann, S. (2006). When consumers follow their feelings: the impact of affective or cognitive focus on the basis of consumers' choice. *Psychology & Marketing, 23*, 1005-1036. doi: 0.1002/mar.20144

Schmukle, S. C., & Egloff, B. (2005). A latent state-trait analysis of implicit and explicit personality measures. *European Journal of Psychological Assessment, 21*, 100-107. doi: 10.1027/1015-5759.21.2.100

Schnabel, K., Asendorpf, J. B., & Greenwald, A. G. (2008a). Assessment of individual differences in implicit cognition: A review of IAT measures. *European Journal of Psychological Assessment, 24*, 210-217. doi: 10.1027/1015-5759.24.4.210

Schnabel, K., Asendorpf, J. B., & Greenwald, A. G. (2008b). Implicit Association Tests: A landmark for the assessment of implicit personality self-concept. In G. J. Boyle, G. Matthews & H. Saklofske (Eds.), *Handbook of personality theory and testing* (pp. 508-528). London: Sage.

Schnabel, K., Asendorpf, J. B., & Greenwald, A. G. (2008c). Understanding and using the Implicit Association Test: V. Measuring semantic aspects of trait self-concepts. *European Journal of Personality, 22*, 695-706. doi: 10.1002/per.697

Schnabel, K., Banse, R., & Asendorpf, J. B. (2006a). Assessment of implicit personality self-concept using the implicit association test (IAT): Concurrent assessment of anxiousness and angriness. *British Journal of Social Psychology, 45*, 373-396. doi: 10.1348/014466605X49159

Schnabel, K., Banse, R., & Asendorpf, J. B. (2006b). Employing automatic approach and avoidance tendencies for the assessment of implicit personality self-concept: The Implicit Association Procedure (IAP). *Experimental Psychology, 53*, 69-76. doi: 10.1027/1618-3169.53.1.69

Schröder-Abé, M., Rudolph, A., & Schütz, A. (2007). High implicit self-esteem is not necessarily advantageous: Discrepancies between explicit and implicit self-esteem

and their relationship with anger expression and psychological health. *European Journal of Personality, 21*, 319-339. doi: 10.1002/per.626

Schröder-Abé, M., Rudolph, A., Wiesner, A., & Schütz, A. (2007). Self-esteem discrepancies and defensive reactions to social feedback. *International Journal of Psychology, 42*, 174-183. doi: 10.1080/00207590601068134

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes, 101*, 1-19. doi: 10.1016/j.obhdp.2006.05.005

Scobie, E. D., & Scobie, G. E. W. (1998). Damaging events: The perceived need for forgiveness. *Journal for the Theory of Social Behaviour, 28*, 373-401. doi: 10.1111/1468-5914.00081

Seibt, B., Hafner, M., & Deutsch, R. (2007). Prepared to eat: How immediate affective and motivational responses to food cues are influenced by food deprivation. *European Journal of Social Psychology, 37*, 359-379. doi: 10.1002/ejsp.365

Siassi, S. (2007). Forgiveness, acceptance and the matter of expectation. *International Journal of Psychoanalysis, 88*, 1423-1440. doi: 10.1516/4W17-3154-1T35-T460

Siegel, E. F. (2006). *Questioning the validity of the IAT: Knowledge or attitude?* (Unpublished master's thesis). University of Maryland.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*, 108-131. doi: 10.1207/S15327957PSPR0402_01

Smith, K. A. (200?). *Measuring forgiveness attitudes with the implicit association test: An exploratory study* (Unpublished honour's thesis). Charles Sturt University.

Spalding, L. R., & Hardin, C. D. (1999). Unconscious unease and self-handicapping: Behavioral consequences of individual differences in implicit and explicit self-esteem. *Psychological Science, 10*, 535-539. doi: 10.1111/1467-9280.00202

Sriram, N., & Greenwald, A. G. (2009). The brief Implicit Association Test. *Experimental Psychology, 56*, 283-294. doi: 10.1111/1467-9280.00202

Steffens, M. C. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology, 51*, 165-179. doi: 10.1027/1618-3169.51.3.165

Steffens, M. C., & Buchner, A. (2003). Implicit Association Test: Separating transsituationally stable and variable components of attitudes toward gay men. *Experimental Psychology, 50*, 33-48. doi: 10.1026//1618-3169.50.1.33

Steffens, M. C., Yundina, E., & Panning, M. (2008). Automatic associations with "erotic" in child sexual offenders: Identifying those in danger of reoffence. *Sexual Offender Treatment, 3*, 1-9.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251. doi: 10.1037/0033-2909.87.2.245

Stoeber, J. (2001). The Social Desirability Scale-17 (SDS-17) convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment, 17*, 222-232. doi: 10.1027//1015-5759.17.3.222

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*, 220-247. doi: 10.1207/s15327957pspr0803_1

Strelan, P. (2007). The prosocial, adaptive qualities of just world beliefs: Implications for the relationship between justice and forgiveness. *Personality and Individual Differences, 43,* 881-890. doi: 10.1016/j.paid.2007.02.015

Strelan, P., & Covic, T. (2006). A review of forgiveness process models and a coping framework to guide future research. *Journal of Social and Clinical Psychology, 25*, 1059-1085. doi: 10.1521/jscp.2006.25.10.1059

Strelan, P., Feather, N. T., & McKee, I. (2008). Justice and forgiveness: Experimental evidence for compatibility. *Journal of Experimental Social Psychology, 44*, 1538-1544. doi: 10.1016/j.jesp.2008.07.014

Strelan, P., & McKee, I. (under review). For whom do we forgive? A functional analysis of motives for forgiveness. *Personal Relationships*.

Strelan, P., & Sutton, R. M. (2011). When just-world beliefs promote and when they inhibit forgiveness. *Personality and Individual Differences, 50*, 163-168. doi: 10.1016/j.paid.2010.09.019

Struthers, C. W., Dupuis, R., & Eaton, J. (2005). Promoting forgiveness among co-workers following a workplace transgression: The effects of social motivation training. *Canadian Journal of Behavioural Science, 37*, 299-308. doi: 10.1037/h0087264

Struthers, C. W., Eaton, J., Santelli, A. G., Uchiyama, M., & Shirvani, N. (2008). The effects of attributions of intent and apology on forgiveness: When saying sorry may not help the story. *Journal of Experimental Social Psychology, 44*, 983-992. doi: 10.1016/j.jesp.2008.02.006

Struthers, C. W., Eaton, J., Shirvani, N., Georghiou, M., & Edell, E. (2008). The effect of preemptive forgiveness and a transgressor's responsibility on shame, motivation

to reconcile, and repentance. *Basic and Applied Social Psychology, 30*, 130-141. doi: 10.1080/01973530802209178

Stuckless, N., & Goranson, R. (1992). The Vengeance scale: Development of a measure of attitudes toward revenge. *Journal of Social Behavior and Personality, 7,* 25-42.

Subkoviak, M. J., Enright, R. D., Wu, C.-R., Gassin, E. A., Freedman, S., Olson, L. M., et al. (1995). Measuring interpersonal forgiveness in late adolescence and middle adulthood. *Journal of Adolescence, 18*, 641-655. doi: 10.1006/jado.1995.1045

Suwartono, C., Prawasti, C. Y., & Mullet, E. (2007). Effect of culture on forgivingness: A Southern Asia-Western Europe comparison. *Personality and Individual Differences, 42*, 515-523. doi: 10.1016/j.paid.2006.07.027

Swanson, J. E., Rudman, L. A., & Greenwald, A. G. (2001). Using the Implicit Association Test to investigate attitude-behavior consistency for stigmatized behavior. *Cognition and Emotion, 15*, 207-230. doi: 10.1080/0269993004200060

Tangney, J. P., Boone, A. L., & Dearing, R. (2005). Forgiving the self: Conceptual issues and empirical findings. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 143-158). New York: Brunner-Routledge.

Terzino, K., Fincham, F. D., & Cross, S. E. (2007). *But do you really forgive me? An implicit test of forgiveness.* Unpublished manuscript.

Thompson, L. Y., & Snyder, C. R. (2003). Measuring forgiveness. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook of models and measures* (pp. 301-312). Washington: American Psychological Association.

Thompson, L. Y., Snyder, C. R., Hoffman, L., Michael, S. T., Rasmussen, H. T., Billings, L. S., et al. (2005). Dispositional forgiveness of self, others, and situations. *Journal of Personality, 73*, 313-360. doi: 10.1111/j.1467-6494.2005.00311.x

Toussaint, L., & Webb, J. R. (2005a). Gender differences in the relationship between empathy and forgiveness. *The Journal of Social Psychology, 145*, 673-685. doi: 10.3200/SOCP.145.6.673-686

Toussaint, L., & Webb, J. R. (2005b). Theoretical and empirical connections between forgiveness, mental health, and well-being. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 349-362). New York: Brunner-Routledge.

Tripp, T. M., Bies, R. J., & Aquino, K. (2007). A vigilante model of justice: Revenge, reconciliation, forgiveness, and avoidance. *Social Justice Research, 20*, 10-34.

Tsang, J.-A., McCullough, M. E., & Fincham, F. D. (2006). The longitudinal association between forgiveness and relationship closeness and commitment. *Journal of Social and Clinical Psychology, 25*, 448-472. doi: 10.1521/jscp.2006.25.4.448

Tsang, J.-A., McCullough, M. E., & Hoyt, W. T. (2005). Psychometric and rationalization accounts of the religion-forgiveness discrepancy. *Journal of Social Issues, 61*, 785-805. doi: 10.1111/j.1540-4560.2005.00432.x

Tsuang, M. T., Eaves, L., Nir, T., Jerskey, B. A., & Lyons, M. J. (2005). Genetic influences on forgiving. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 245-258). New York: Brunner-Routledge.

Uhlmann, E. L., & Swanson, J. E. (2004). Exposure to violent video games increases automatic aggressiveness. *Journal of Adolescence, 27*, 41-52. doi: 0.1016/j.adolescence.2003.10.004

Umbreit, M. S. (1989). Crime victims seeking fairness, not revenge: Toward restorative justice. *Federal Probation, 53*, 52-57.

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self control. *Perspectives on Psychological Science, 5*, 243-262. doi: 10.1177/1745691610369465

van Bockstaele, B., Verschuere, B., Koster, E. H., Tibboel, H., De Houwer, J., & Crombez, G. (in press). Differential predictive power of self report and implicit measures on behavioural and physiological fear responses to spiders. *International Journal of Psychophysiology*. doi: 10.1016/j.ijpsycho.2010.10.003

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal, 47*, 497-527. doi: 10.3102/0002831209353594

van Goethem, A. A. J., Scholte, R. H. J., & Wiers, R. W. (2010). Explicit- and implicit bullying attitudes in relation to bullying behavior. *Journal of Abnormal Child Psychology, 38*, 829-842. doi: 10.1007/s10802-010-9405-2

van Quaquebeke, N., & Schmerling, A. (2010). Wie die bloße abbildung bekannter weiblicher und männlicher führungskräfte unser implizites denken zu führung beeinflusst. [Cognitive equal opportunities: How the mere presentation of renowned female and male leaders affects our implicit thinking on leadership]. *Zeitschrift für Arbeits- und Organisationspsychologie, 54*, 91-104. doi: 10.1026/0932-4089/a000020

Veenstra, G. (1992). Psychological concepts of forgiveness. *Journal of Psychology and Christianity, 11*, 160-169.

Verschuere, B., Prati, V., & De Houwer, J. (2009). Cheating the lie detector: Faking in the autobiographical Implicit Association Test. *Psychological Science, 20,* 410-413. doi: 10.1111/j.1467-9280.2009.02308.x

Vianello, M., Robusto, E., & Anselmi, P. (2010). Implicit conscientiousness predicts academic performance. *Personality and Individual Differences, 48*, 452-457. doi: 10.1016/j.paid.2009.11.019

Vidmar, N. (2000). Retribution and revenge. In J. Sanders & V. L. Hamilton (Eds.), *Handbook of justice research in law* (pp. 31-63). New York: Kluwer/Plenum.

Vidmar, N. (2002). Redistributive justice: Its social context. In M. Ross & D. T. Miller (Eds.), *The justice motive in everyday life* (pp. 291-313). Cambridge, UK: Cambridge University Press.

Wade, N. G., & Worthington Jr, E. L. (2003). Overcoming interpersonal offenses: Is forgiveness the only way to deal with unforgiveness? *Journal of Counseling & Development, 81*, 343-353.

Walker, D. F., & Gorsuch, R. L. (2002). Forgiveness within the Big Five personality model. *Personality and Individual Differences, 32*, 1127-1137. doi: 10.1016/S0191-8869(00)00185-9

Walker, L. J., Pitts, R. C., Hennig, K. H., & Matsuba, M. K. (1995). Reasoning about morality and real-life moral problems. In M. Killen & D. Hart (Eds.), *Morality in everyday life: Developmental perspectives* (pp. 371-407). New York: Cambridge University Press.

Wallace, H. M., Exline, J. J., & Baumeister, R. F. (2008). Interpersonal consequences of forgiveness: Does forgiveness deter or encourage repeat offenses? *Journal of Experimental Social Psychology, 44*, 453-460. doi: 10.1016/j.jesp.2007.02.012

Wallaert, M., Ward, A., & Mann, T. (2010). Explicit control of implicit responses: Simple directives can alter IAT performance. *Social Psychology, 41*, 152-157. doi: 10.1027/1864-9335/a000022

Waltman, M. A., Russell, D. C., Coyle, C. T., Enright, R. D., Holter, A. C., & M. Swoboda, C. (2009). The effects of a forgiveness intervention on patients with coronary artery disease. *Psychology and Health, 24*, 11-27. doi: 10.1080/08870440801975127

Webb, J. R., & Brewer, K. (2010). Forgiveness, health, and problematic drinking among college students in southern Appalachia. *Journal of Health Psychology, 15*, 1257-1266. doi: 10.1177/1359105310365177

Webb, J. R., Toussaint, L., Kalpakjian, C. Z., & Tate, D. G. (2010). Forgiveness and health-related outcomes among people with spinal cord injury. *Disability and Rehabilitation, 32*, 360-366. doi: 10.3109/09638280903166360

Webb, T. L., Sheeran, P., & Pepper, J. (in press). Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials. *British Journal of Social Psychology*.

Wegner, D. M., & Bargh, J. A. (1998). Control and automaticity in social life. In D. T. Gilbert, S. E. Fiske & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 446-496). New York: McGraw-Hill.

Wenzel, M., & Okimoto, T. G. (2010). How acts of forgiveness restore a sense of justice: Addressing status/power and value concerns raised by transgressions. *European Journal of Social Psychology, 40*, 401-417.

Whited, M. C., Wheat, A. L., & Larkin, K. T. (2010). The influence of forgiveness and apology on cardiovascular reactivity and recovery in response to mental stress. *Journal of Behavioral Medicine*, *33,* 293-304. doi: 10.1007/s10865-010-9259-7

Wiers, R. W., van Woerden, N., Smulders, F. T. Y., & de Jong, P. J. (2002). Implicit and explicit alcohol-related cognitions in heavy and light drinkers. *Journal of Abnormal Psychology, 111*, 648-658. doi: 10.1037/0021-843X.111.4.648

Wigboldus, D. H. J., Holland, R. W., & van Knippenberg, A. (2004). *Single target implicit associations*. Unpublished manuscript.

Wilkowski, B. M., Robinson, M. D., & Troop-Gordon, W. (2010). How does cognitive control reduce anger and aggression? The role of conflict monitoring and forgiveness processes. *Journal of Personality and Social Psychology, 98*, 830-840. doi: 10.1037/a0018962

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, *55,* 493-518. doi: 10.1146/annurev.psych.55.090902.141954

Wilson, T. D., Lindsey, S., & Schooler, T. (2000). A model of dual attitudes. *Psychological Review, 107*, 101-126. doi: 10.1037/0033-295X.107.1.101

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81*, 815-827. doi: 0.1037/0022-3514.81.5.815

Witvliet, C. v., Ludwig, T. E., & Vander Laan, K. L. (2001). Granting forgiveness or harboring grudges: Implications for emotion, physiology, and health. *Psychological Science, 12*, 117-123. doi: 10.1111/1467-9280.00320

Witvliet, C. v., Worthington Jr, E. L., Root, L. M., Sato, A. F., Ludwig, T. E., & Exline, J. J. (2008). Retributive justice, restorative justice, and forgiveness: An experimental psychophysiology analysis. *Journal of Experimental Social Psychology, 44*, 10-25. doi: 10.1016/j.jesp.2007.01.009

Worthington Jr, E. L. (1998). The pyramid model of forgiveness: Some interdisciplinary speculations about unforgiveness and the promotion of forgiveness. In E. L. Worthington Jr (Ed.), *Dimensions of forgiveness*. Radnor, Pennsylvania: Templeton Foundation Press.

Worthington Jr, E. L. (2000). Is there a place for forgiveness in the justice system? *Fordham Urban Law Journal, 27*, 1721-1734.

Worthington Jr, E. L. (2001). Unforgiveness, forgiveness, and reconciliation and their implications for societal interventions. In R. G. Helmick & R. L. Petersen (Eds.), *Forgiveness and reconciliation* (pp. 171-192). Philadelphia: Templeton Foundation Press.

Worthington Jr, E. L. (2005a). Initial questions about the art and science of forgiving. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 1-14). New York: Brunner-Routledge.

Worthington Jr, E. L. (2005b). More questions about forgiveness: Research agenda for 2005-2015. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness* (pp. 557-573). New York: Brunner-Routledge.

Worthington Jr, E. L. (2005c). The social psychology of justice. In E. L. Worthington Jr (Ed.), *Handbook of forgiveness*. New York: Brunner-Routledge.

Worthington Jr, E. L. (2008). *Forgiveness and justice*. Retrieved from http://www.thepowerofforgiveness.com/outreach/index.html

Worthington Jr, E. L. (2009). *A just forgiveness: Responsible healing without excusing injustice*. Downers Grove, IL: InterVarsity Press.

Worthington Jr, E. L., & Scherer, M. (2004). Forgiveness is an emotion-focused coping strategy that can reduce health risks and promote health resilience: Theory, review, and hypotheses. *Psychology and Health, 19*, 385–405. doi: 10.1080/0887044042000196674

Wu, J., & Axelrod, R. (1995). How to cope with noise in the iterated prisoner's dilemma. *The Journal of Conflict Resolution, 39*, 183-189. doi: 10.1177/0022002795039001008

Wygant, S. A. (1997). Moral reasoning about real-life moral dilemmas: Paradox in research using the Defining Issues Test. *Personality and Social Psychology Bulletin, 23*, 1022-1033. doi: 10.1177/01461672972310003

Yandell, K. E. (1998). The metaphysics and morality of forgiveness. In R. D. Enright & J. North (Eds.), *Exploring forgiveness* (pp. 35-45). Wisconsin: The University of Wisconsin Press.

Young, L., & Saxe, R. (2009). Innocent Intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47*, 2065-2072. doi: 10.1016/j.neuropsychologia.2009.03.020

Younger, J. W., Piferi, R. L., Jobe, R., & Lawler, K. A. (2004). Dimensions of forgiveness: The views of laypersons. *Journal of Social and Personal Relationships, 21*, 837-855. doi: 10.1177/0265407504047843

Ysseldyk, R., Matheson, K., & Anisman, H. (2009). Forgiveness and the appraisal-coping process in response to relationship conflicts: Implications for depressive symptoms. *Stress, 12*, 152-166. doi: 0.1080/10253890802228178

Zechmeister, J. S., Garcia, S., Romero, C., & Vas, S. N. (2004). Don't apologize unless you mean it: A laboratory investigation of forgiveness and retaliation. *Journal of Social and Clinical Psychology, 23*, 532-564. doi: 10.1521/jscp.23.4.532.40309

Zechmeister, J. S., & Romero, C. (2002). Victim and offender accounts of interpersonal conflict: Autobiographical narratives of forgiveness and unforgiveness. *Journal of Personality and Social Psychology, 82*, 675-686. doi: 10.1037/0022-3514.82.4.675