Redemption Through Suffering:
How Self-Punishment Restores Moral Identity

by

Melissa de Vel-Palumbo

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy

College of Education, Psychology and Social Work
January 2018

Table of Contents

Summary

Individuals sometimes respond to their misdeeds by punishing themselves. Though such behaviours might be thought of as dysfunctional, in this thesis I argue that self-punishment is a strategy transgressors may use in an attempt to restore their sense of moral identity. In particular, my research describes several mechanisms through which self-punishment achieves moral redemption.

The primary focus of this thesis is on the self-punisher's experience. I propose that self-punishment can be utilised in two distinct ways to resolve the threat to moral identity triggered by one's wrongdoing. One process is in line with experimental research arguing that self-punishment "cleanses" a guilty conscience (Bastian, Jetten, & Fasoli, 2011; Inbar, Pizarro, Gilovich, & Ariely, 2013), thereby protecting one's moral identity by avoiding the implications of the wrongdoing. Yet, clinical research suggests that psychological self-punishment might exacerbate distress (Dyer et al., 2017; Whelton & Greenberg, 2005)—a finding that contradicts the predominant view. Thus, I delineate a second function of self-punishment, one of moral repair, whereby self-punishment acts as an exploration (rather than an evasion) of one's guilt. My model brings together both functions by arguing that self-punishment can be both defensive and reparative. I find support for this model using various methodologies including qualitative exploration and quantitative experimental paradigms. Findings indicated that to the extent that transgressors acknowledged the threat to their moral identity, they were more likely to use self-punishment as an avenue for critical self-examination and moral repair.

I also explore the phenomenon of self-punishment from the perspective of third parties. I propose that self-punishment can also exonerate self-punishers in the eyes of *others* by restoring a sense of symbolic justice, thereby securing others' forgiveness. Yet, my results suggested that whether self-punishment restored justice (and transgressors' moral image)

depended somewhat on third parties' interpretation of the self-punishment. Third parties were more forgiving when they perceived that self-punishers were truly revising their moral values, in line with a process of moral repair. Taken together, the findings suggest that while self-punishers can redeem their moral identity through either excusing or confronting their wrongdoing, these two functions have profoundly different implications for intrapersonal and interpersonal repair.

Declaration

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

……..

Melissa de Vel-Palumbo

Acknowledgments

It is good to have an end to journey toward; but it is the journey that matters, in the end.

—Ursula Le Guin, *The Left Hand of Darkness*

For the research contained within the following pages, I am deeply indebted to my two brilliant supervisors. I was fortunate to have both of you guide me through this challenging but rewarding journey. Michael Wenzel: Thank you for your seemingly boundless knowledge (and patience!), which you so generously shared with me in this pursuit. I am grateful for all the intellectual insights and advice you offered me, while inspiring and respecting my intellectual creativity. Lydia Woodyatt: Thank you for always being able to see the wood for the trees, in many senses. Your ideas, perceptiveness, and energy have been a large part of making this both a meaningful and enjoyable ride.

Thanks also to all my peers and colleagues at Flinders University. Cheers to Simon Bury for keeping me sane: for the laughs and the honest musings about life over beers. To Farid Anvari: for the all the research and philosophy discussions over coffees. Thanks to Mikaela Cibich who provided advice on a chapter, and to all in the PsychJEM lab, who have been a source of intellectual and social support through the years. Thanks to colleagues who helped directly with the project: Paul Williamson for statistics help, and the folk in Engineering and Audiology (especially Sarosh Kapadia) for helping me develop my auditory task. And to all the unsuspecting undergraduate students who so kindly agreed to participate in my experimental lab studies, only to be subjected to loud noise bursts and ice buckets!

In addition, I am grateful to my family and friends, without whom I would not have been able to persevere in this task. To George especially, for being my number one fan—I always feel like I've succeeded in your eyes.

Statement of Co-Authorship

Chapter 2

de Vel-Palumbo, M., Woodyatt, L., & Wenzel, M. (in press). Why do we self-punish? Perceptions of the motives and impact of self-punishment outside the laboratory. *European Journal of Social Psychology*.

Chapter 4

de Vel-Palumbo, M., Wenzel, M., & Woodyatt, L. (2017). *Expression or evasion of guilt? Two ways self-punishment resolves the threat to moral identity.* Manuscript in preparation.

Chapter 5

de Vel-Palumbo, M., Wenzel, M., & Woodyatt, L. (2017). *Suffering for justice: Self-punishment and third party forgiveness*. Manuscript in preparation.

CHAPTER 1: Introduction

> I accept the torture of accusation, and my public shame; I want to suffer and by suffering I shall be purified.
>
> —Dostoevsky, *The Brothers Karamazov*

There exists a strange class of behaviours, one that seemingly defies the human drive to seek pleasure and avoid pain: self-punishment. Hungry for our own suffering, sometimes we punish ourselves for our misdeeds, in what can be shocking and brutal ways. History is replete with accounts of self-punishment; indeed, it is a recurring motif in the mythology of many religions. For hundreds of years Muslim Shi'ites have been cutting themselves with knives and chains on the day of Ashura, in repentance for their ancestors' moral failings (Pierre, Hutchinson, & Abdulrazak, 2007). Catholic flagellants of the 13th and 14th centuries marched through cities lashing themselves with cattail whips to demonstrate penance and piety (E. Beard, 2013). Delving even further into the past, Roman Galli priests severed their own genitals on *Dies Sanguinis* [Day of Blood] as purification for the betrayal of their goddess (M. Beard, 1994). Formalising this pervasive psychological desire for our own punishment, Freud (1916/1957) termed it "moral masochism."

But by no means is self-punishment an antiquated response to wrongdoing that can be relegated to specific historical or religious contexts. Attacking the self is a relatively common response to how we deal with shameful events (Nyström & Mikkelsen, 2013; Yelsma, Brown, & Elison, 2002). Take for instance those who find solace in self-harm because they feel they deserve to be punished for their immorality (Klonsky, 2011; Nock, 2009). Consider the following example from research reported in the current thesis, of an individual who physically harmed himself after his wrongdoing: "I *deserved* to get punished for doing such a bad crime against my dear brother… It made me feel as if I needed to *redeem* myself… I think people self-punish because … *it feels as if* you are making all the guilt go by punishing

yourself" (P37, Chapter 2, emphases added). Of course, physical harm is a rather extreme form of punishment, but self-punishment may manifest itself in more subtle, everyday forms: reprimanding and cursing at ourselves in anger; isolating ourselves from family and friends; going out of our way to make amends to those we have hurt.

The motivations behind self-punishment are unclear. Our lay understanding would suggest that we are driven to engage in such acts in the pursuit of moral "purification"—a renewed sense of self as a good person. For the most part, the limited experimental research on self-punishment appears consistent with this notion; two recent studies found that self-punishment can alleviate guilt (Bastian, Jetten & Fasoli, 2011; Inbar, Pizarro, Gilovich, & Ariely, 2013). However, these findings are at odds with suggestions in the clinical literature that attacking or criticising the self can perpetuate and exacerbate psychological distress, rather than quell it (Dyer et al., 2017; Whelton & Greenberg, 2005). Yet nobody has attempted to consolidate these two conflicting lines of research.

Moreover, there are no accounts of the mechanism through which self-punishment might reduce guilt, if it does so. It is not obvious how the negative self-appraisals that are implicitly involved in an act of self-punishment would lead to less guilt. On the face of it, this proposal is counterintuitive: How can condemning oneself as *immoral* make one feel *moral*? Nor does the literature shed any light on the effect of self-punishment on cognitive processes and interpersonal outcomes pertinent to responses to wrongdoing, such as moral engagement and interpersonal reconciliation. For example, if self-punishment reduces guilt, then does it consequently inhibit the motivation to engage in additional reparative behaviours, such as compensating the victim directly? If so, how then might observers of injustice respond to this self-oriented approach: with forgiveness or with frustration? Does moral purification in one's own eyes lead to moral purification in others' eyes?

In broad terms, the objective of this thesis is to understand the intrapersonal and interpersonal functions of self-punishment; that is, why people engage in self-punishment, considering its effects on both their own experiences and also on others' perceptions. Consolidating discrepancies in the literature, I propose that self-punishment may actually reflect not one but two distinct processes that attempt to address one's sense of immorality: *moral cleansing* and *moral repair*. This model provides a framework for predicting the various cognitive, emotional, and interpersonal outcomes of self-punishment. By testing this model across various methodologies, I hope to ascertain whether (and how) self-punishment can redeem one's morality.

Understanding How a Wrongdoing Creates a Threat to Moral Identity and How Self-Punishment May Attempt to Resolve that Threat

Having behaved in a way that is unacceptable, transgressors experience a threat to their moral identity—their sense that they are a good person who is accepted by others as such (SimanTov-Nachlieli, Shnabel, & Nadler, 2013). From a functional perspective, threats to moral identity are part of an adaptive self-regulatory process that moderates social behaviour and motivates us to repair the harm we have done to others, allowing us to maintain positive interpersonal relations (Leary & Baumeister, 2000; Sachdeva, Iliev, & Medin, 2009; J. L. Tracy & Robins, 2004). Thus, a threat to moral identity alerts transgressors to the possibility of social exclusion and the need to defuse the threat somehow.

A threatened moral identity, experienced with moral emotions such as guilt or shame, spurs perpetrators into actions that aim to restore moral self-image (J. M. Carlsmith & Gross, 1969; Gino, Kouchaki, & Galinsky, 2015). Self-punishment may be one such attempt to resolve the threat to moral identity—a claim that can be made on both empirical and theoretical grounds. First, as already highlighted, the literature and lay discourse about self-punishment is replete with the narrative of "moral purification," suggesting the act is aimed at

one's feelings of immorality. Second, there is no sense to the idea of deserving a "punishment" without first believing we have behaved immorally. This feature distinguishes self-punishment from other types of self-inflicted suffering, such as: suffering for the sake of spiritual enlightenment; seeking altered states of consciousness; or asceticism in order to eliminate distractions from religious practice. Indeed, though the notion of self-harm was invoked earlier, many cases of self-harm are not driven by a desire to punish oneself (Klonsky, 2011; Nock, 2009; Nock & Prinstein, 2004). Thus self-harm in general does not equate to self-punishment, but rather, *some* forms of self-harm may be motivated by self-punishment. Leaving aside these other self-inflicted behaviours, self-punishment is theoretically bound to the perception that one has behaved immorally. Accordingly, self-punishment appears to be directed towards restoring one's moral identity. But how precisely might it do so?

I propose that attempts to restore moral identity may be potentially placed into one of two categories, in keeping with "repair" versus "defend" oriented responses to guilt and shame (e.g., Gausel, Leach, Vignoles, & Brown, 2012; Schmader & Lickel, 2006). For instance, one may seek to repair the harm done through apologies or restitution—actions that demonstrate to others that one is a worthy group member who deserves to be brought back into the fold. These types of behaviours can be categorised as *moral repair*: actions that actively confront a wrongdoing and directly repair the material or social harm caused. Alternatively, transgressors may employ defensive processes that downplay or "shake off" their wrongdoing in some way. These sorts of acts traditionally involve a reappraisal of the initial act in order to regain positive self-regard, e.g., minimising the harm done or shifting blame (Bandura, 1991, 1999; Sykes & Matza, 1957). These actions can be categorised as *moral cleansing*: an attempt to ameliorate self-worth by ridding oneself of moral culpability. But can self-punishment be best understood as a process of moral repair, or moral cleansing?

I propose that self-punishment may potentially function as either confronting and repairing one's immorality (moral repair), or avoiding and ridding oneself of doubts about one's morality (moral cleansing). Though both actions may be geared towards the same general goal—to defuse the threat to moral identity—the mechanisms for arriving at that point are different, and as a result they have different outcomes.

Self-Punishment as Moral Cleansing

Moral cleansing is presently defined as a process through which one is "cleansed" of one's past immoral behaviour, providing release from the negative emotional and cognitive features that characterise threatened moral identity (Sachdeva et al., 2009). Moral cleansing has been described broadly as behaviour to recover self-worth, or the act of doing something good after doing something bad (Brañas-Garza, Bucheli, Espinosa, & García-Muñoz, 2013). In the self-punishment context I employ the term in a narrower sense: Moral cleansing is a self-enhancing process that does not focus on the wrongdoing itself (or the values violated in its commission), but rather acts as a "surrogate" for moral repair (Zhong & Liljenquist, 2006, p. 1451). That is, any reduction of distress and restoration of positive moral identity is achieved by bypassing or prematurely casting off self-examination. In this sense, self-punishers are simply purging themselves of guilt or other aversive emotional states and cognitions, rather than unpacking why they are feeling this way.

Consistent with moral cleansing, two empirical studies have found that self-punishment can reduce feelings of guilt (Bastian et al., 2011; Inbar et al., 2013). In these novel paradigms, participants were induced to feel guilty by recalling a past interpersonal transgression, then engaged in self-punishment—either by giving themselves electric shocks (Inbar et al., 2013) or by holding their hand in a bucket of ice water (Bastian et al., 2011). Self-punishers felt less guilty after the ice bucket task relative to a control condition that did not experience pain (Bastian et al., 2011), and the decline in guilt was correlated with the

intensity of the self-punishment (Inbar et al., 2013). These two studies provide some evidence that self-punishment can cleanse transgressors of their immoral behaviour and restore their moral identity; however, it is unclear through what mechanism it might do so.

If self-punishment reduces guilt, a possible mechanism for this might be self-affirmation (Steele, 1988). Self-affirmation provides offenders with the opportunity to maintain self-integrity by viewing identity as global rather than specific, allowing weaknesses in one area to be compensated for by strengths in another. Accordingly, self-punishment may restore the offender's global identity through *unrelated* affirmations (i.e., not the values violated in the transgression). For example, self-punishers might be affirming their sense of self-agency or control—a central feature of self-destructive behaviours such as eating disorders (Fairburn, Shafran, & Cooper, 1999; Slade, 1982), self-harm (Chapman, Gratz, & Brown, 2006; Favazza, 1996) and substance use (Wurmser, 1974; Young, Boyd, & Hubbell, 2000). It may feel comforting to be in control of one's pain, compensating for the guilt and distress arising from the transgression.

Alternatively, affirmations may be within the moral domain; for example, Rothschild and colleagues (Rothschild, Landau, Keefer, & Sullivan, 2015) suggest that self-punishment restores positive self-regard through administering the punishment, rather than receiving it. This argument draws from earlier research showing that punishing moral transgressors can make people feel "morally just" (Adams, 2011). Similarly, the act of punishing oneself might bolster a punisher's identity via an affirmation of one's moral superiority as a righteous administrator of justice. It is important to note that despite being in the moral domain, such an affirmation does not focus on oneself as the *perpetrator* of the wrongdoing, doing little to unpack the implications of the wrongdoing for the transgressor. Here, self-punishers are not giving much thought to their sense of immorality and the values threatened by the wrongdoing itself; rather, they are focused on their sense of morality and virtuousness

apparent in their *response* to the wrongdoing ("I punished myself, therefore I am morally good"). In this sense, these types of affirmations should be considered unrelated to the transgression, and thus an act of moral cleansing.

Self-affirmation allows individuals to deal with the negative psychological experience and moral identity concerns following a transgression by restoring positive identity through alternate resources. It contributes little, however, to an examination or appreciation of the wrong committed. Self-punishment in these cases therefore acts as an evasion of guilt rather than an expression of it (Carveth, 2006), analogous to physical cleansing behaviours that purify by "wiping the slate clean" (S. W. Lee & Schwarz, 2011). Thus, moral cleansing is a form of moral disengagement. But, unlike traditional mechanisms of moral disengagement, I propose that self-punishment does not involve a reappraisal of one's wrongdoing (indeed, it necessitates an acceptance of moral error to have any sense to it), but rather, self-punishment is a form of disengagement in its defence against the *implications* of the wrongdoing for the self. Under this account, an examination of one's guilt is by-passed, as the wrongdoing is initially acknowledged but swiftly excused and its implications left unaddressed. A transgressor can tell oneself, "through this punishment I have atoned, I am morally virtuous, and I can now move on."

Self-Punishment as Moral Repair

In contrast to the moral cleansing account, self-punishment could be an attempt to directly confront one's moral identity. Research has found that when people are faced with unfavourable feedback, though this may be distressing they do not necessarily choose to abandon or avoid this state. In fact, they show a preference for additional unfavourable feedback that allows them to target their weaknesses and improve themselves (Sedikides & Luke, 2007). In the context of transgressions, deconstructing the threat allows transgressors

to learn from the wrongdoing. In line with this logic, perhaps self-punishment is an act of confrontation rather than an act of avoidance.

Recall that self-affirmation protects a threatened moral identity by affirming *unrelated* personal values (as is proposed in moral cleansing). It could be the case that self-punishment seeks to restore moral identity by affirming the values *related* to the transgression. They might wish to affirm that the immoral behaviour is not representative of their moral character; that they truly are a good person precisely because they see this as a wrongdoing and are willing to be punished for it. Notice here that the act of self-punishment is directly targeting the wrongdoing itself. In this way, self-punishers are addressing the source of the moral threat—the threat to the values of their social group and their self-definition as a good group member.

Consequently, moral repair divorces the transgression from the offender's identity but not in a way that avoids or bypasses examination of one's wrongdoing. On the contrary, self-punishment as moral repair can embrace genuine responsibility (i.e., thinking through what one did, why it was wrong, and what else needs to be done to make amends for the harm) while denouncing the immorality of the act. Affirming transgression-relevant values is an opportunity for transgressors to confirm they believe in the value that has been violated (Woodyatt & Wenzel, 2014). Under this formulation, self-punishment acts as a genuine sign of remorse, a condemnation of one's wrongdoing, and an avowal that the violated value is of importance to one's identity. This process solidifies one's commitment to moral values and the people with whom one shares those values. Tetlock, Kristel, Elson, Green, and Learner (2000) conceptualise the affirmation of shared values as "essential for resolutely reasserting the identification of the self with the collective moral order" (p. 854).

Despite going some way to restore one's moral identity, the types of cognitive processes associated with moral repair are rather psychologically demanding. For example,

moral value affirmation involves the transgressor unpacking the wrongdoing, identifying which values were violated, and exploring why these values are important (Woodyatt & Wenzel, 2014). This process does not necessarily reduce guilt and shame as moral cleansing approaches do. In fact, value affirmation can initially increase shame (Woodyatt & Wenzel, 2014). Correspondingly, if self-punishment were a process of moral repair, we would not always expect an immediate reduction in guilt and/or shame. Supporting this conceptualisation of self-punishment, research has found that self-condemnation (measured by endorsing items such as "I deserve to suffer for this" and "I feel like a bad person") is associated with *less* psychological wellbeing (Fisher & Exline, 2006). Furthermore, self-criticism has been linked to psychological distress (Whelton & Greenberg, 2005). This provides some preliminary evidence that people engaging in self-punishment may not always feel good afterwards—perhaps because they are attempting the tough process of moral repair.

Interpersonal Implications of Self-Punishment: Self-Punishment Can Inhibit or Sustain the Motivation to Help Victims

Understanding the psychological mechanisms of moral cleansing and moral repair allows us to formulate predictions about the effect of these processes on self-punishers' behaviour towards others. Moral emotions motivate prosocial behaviour to amend the hurt caused by a transgression (Baumeister, Stillwell, & Heatherton, 1994; J. M. Carlsmith & Gross, 1969; De Hooge, Breugelmans, & Zeelenberg, 2008; Gino et al., 2015; Tangney, Stuewig, & Mashek, 2007). Changes in guilt and shame are thus likely to influence self-punishers' motivation to help their victim or repair the harm done in other ways.

If self-punishment maintains or increases one's guilt via moral repair, then self-punishment would not necessarily replace or inhibit interpersonal reconciliation. Instead, negative emotions would linger or even intensify until interpersonal reparation has been achieved. Understood in this way, self-punishment as moral repair would be but the first step

on the route to moral redemption; a process that prepares transgressors for subsequent action. Once transgressors affirm their values, they would be interested in reconciliation as an outlet for their renewed sense of guilt and commitment—they would want to act consistently, in line with their values. Self-punishment in this context is an opportunity for offenders to recognise their wrong, re-consider their commitment to their moral values and the collective moral order, and lay the foundations for interpersonal reparatory action if needed.

Providing support for this claim, Woodyatt and Wenzel (2014) found that though value affirmation initially increased shame, it also led to a stronger desire to reconcile with the victim. In addition, Tetlock et al. (2000) found that reaffirming moral values did not replace other reparative behaviours.[1] After being given the chance to affirm non-racist values by indicating willingness to participate in anti-racism activities, participants were no less likely to adjust an insurance pricing policy that they were informed was racially-biased (relative to participants who were not provided the opportunity to affirm non-racist values). Similarly, if self-punishment is a process of moral repair, one would expect interpersonal reconciliation to follow. Self-punishment is a starting point for moral repair: It seems likely that if self-punishment is a well-considered, genuine affirmation of values, it would prompt transgressors to mend their relationship with the victim.

In contrast, moral cleansing might be best understood as part of the psychological immune system, that is, as a defensive response that buffers individuals against negative feedback in order to maintain positive affect and self-esteem (D. T. Gilbert, Pinel, Wilson, Blumberg, & Wheatley, 1998). Moral cleansing behaviours allow us to adjust to aversive circumstances rather than be paralysed by our guilt and shame; in this sense it can be beneficial for one's immediate psychological wellbeing. However, though moral cleansing

[1] Tetlock et al. (2000) use the term "moral cleansing," not to be confused with the formulation of moral cleansing in the current research. Their term is consistent with moral repair: defined as symbolic acts designed to reaffirm values and loyalties in order to strengthen solidarity with one's moral community. In fact the authors explicitly differentiate their notion of moral cleansing from "solely … protecting the self-image" (p. 855).

might reduce initial distress about a wrongdoing, allowing transgressors to carry on with life, affirming unrelated values may be an obstacle to reform and interpersonal reconciliation (Woodyatt & Wenzel, 2014).

The release of a self-punisher's guilt or shame through moral cleansing may undermine the motivational power of such moral emotions to address victim needs (O'Keefe, 2000; Woodyatt & Wenzel, 2014). Indeed, research indicates that self-forgiveness that restores self-regard without cognitive and emotional processing of the transgression is positively related to narcissism, self-centredness (Tangney, Boone, & Dearing, 2005), reduced empathy and willingness to reconcile with the victim (Woodyatt & Wenzel, 2013b). Therefore, insofar as self-punishment reduces guilt, it should replace interpersonal repair and reconciliation, consistent with the idea of moral cleansing as a surrogate or substitute for moral repair. For instance, a moral cleansing behaviour such as—quite literally—washing one's hands after recalling unethical behaviour has been found to inhibit subsequent prosocial behaviour (Zhong & Liljenquist, 2006; but see Fayard, Bassi, Bernstein, & Roberts, 2009, for a failed replication of this study). In the same way, moral cleansing through self-punishment may reduce the motivation for transgressors to do anything else about their transgression, releasing them from their debt to the victim.

Only one study has examined the effect of self-punishment on interpersonal action (van Bunderen & Bastian, 2014). Interestingly, it provides some hints that both processes of self-punishment might be at play. In this study, participants were first induced to commit an interpersonal transgression, punished themselves by immersing their hand in iced water (in the no punishment condition, participants held their hand in lukewarm water). Following the water task, participants were also offered the opportunity to compensate the victim of their transgression (ostensibly, another participant) by awarding the victim points in a game. For some participants, self-punishment inhibited compensation; for others, it did not.

Self-punishment only replaced compensation for individuals high in victim justice sensitivity: a trait-based measure of how readily individuals perceive an incident as constituting a personal injustice against themselves; that is, to perceive themselves as a victim of wrongdoing. To the extent that participants saw themselves as a victim, self-punishment re-balanced the scales of justice, and reduced repair (consistent with moral cleansing). The authors explain their finding thus: Those who are sensitive to injustice gain more "psychological currency" from the experience of pain, boosting the ability for pain to restore justice (van Bunderen & Bastian, 2014). Unexpectedly, however, for those low in victim justice sensitivity, self-punishment did not replace repair. In other words, transgressors who were less likely to perceive themselves as a victim were not simply cleansed of their wrongdoing by punishing themselves. For them, it was not a case of simply atoning through a mere symbolic gesture; instead, they remained motivated to engage in interpersonal repair. Therefore, one could speculate that the findings from this study provide some evidence for both conceptualisations of self-punishment.

Cleansing and Repairing: A Dual-Process Model of Self-Punishment

Taking stock of the literature, it is mired with contradictions. Discrepant findings indicate that self-punishment can sometimes alleviate, and at other times perpetuate feelings of distress. Additionally, self-punishment may inhibit restitution for some, while maintaining it for others. I propose that both cleansing and repair might both be accurate accounts of self-punishment, each arising under different circumstances. The model presented in Figure 1.1 summarises the motivational, cognitive, emotional, and interpersonal factors associated with each process, as well as the proposed mechanisms of action.

Factors Determining Whether Self-Punishment Cleanses or Repairs

If self-punishment is both moral repair and moral cleansing, several factors are possible candidates for moderating the function of self-punishment: characteristics of the

perpetrator, the transgression, and the self-punishment itself. First, effects may be driven by personality characteristics that influence how transgressors respond to their wrongdoing. The effects observed in van Bunderen and Bastian (2014) could be interpreted under the current theoretical lens to suggest that the tendency to see oneself as a victim might differentiate those who use self-punishment as moral cleansing versus those who seek moral repair. In fact, I suggest that victim sensitivity could impact perpetrators' interpretation of the transgression itself (rather than of the pain experience itself). Taken this way, van Bunderen and Bastian's results can be interpreted differently: To the extent that participants interpreted their wrongdoing defensively, playing the victim rather than acknowledging their own role in it, self-punishment was used as moral cleansing.

	Moral Cleansing	Moral Repair
Motivation	Avoid implications of wrongdoing	Confront implications of wrongdoing
Cognitive orientation	Excuse responsibility	Accept responsibility
Identity repair mechanism	Affirming unrelated values	Affirming values related to the transgression
Expected effects	↓ Guilt/shame ↓ Moral engagement ↓ Reconciliation	↑ Guilt/shame ↑ Moral engagement ↑ Reconciliation
Interpersonal focus	○ Self-victimisation ○ Prevent retribution	○ Express genuine remorse ○ Seek forgiveness

Figure 1.1. Model of self-punishment as a moral identity-regulating process

Similarly, Tanaka, Yagi, Komiya, Mifune, and Ohtsubo (2015) have suggested differential functions of self-punishment depending on individual differences. They argue that shame-prone individuals use self-punishment to address more general identity concerns, whereas guilt-prone individuals use self-punishment in an attempt to reconcile with the victim. Shame-proneness is a tendency to feel small, and to hide and escape, while guilt-proneness is the tendency to acknowledge one's mistake and respond constructively with actions to repair the situation (Cohen, Wolf, Panter, & Insko, 2011; Luyten, Fontaine, & Corveleyn, 2002). Therefore, individual differences in how perpetrators interpret and process to their wrongdoing may moderate the function of self-punishment.

Second, the function of a self-punishment may vary according to objective features of the transgression itself. Features of the transgression might activate certain concerns and influence the manner in which one uses self-punishment. For example, I propose that serious transgressions may be more likely to prompt moral repair. This contention would be consistent with Tetlock et al.'s (2000) speculation that when transgressions are less serious, people may be content with a "single-pronged defence" (i.e., when one act of affirmation does away with the need for further affirmation or restitution—moral cleansing), while more substantial threats to moral identity are likely to result in ongoing, multiple forms of affirmation. In other words, for less serious acts, it may be more beneficial to engage in an act of moral cleansing, defusing the threat to moral identity through a single costly act.

Similarly, Gromet (2009) found that as offence severity increases, people prefer procedures that include both restorative and retributive elements. Thus, when the threat to oneself is high, transgressors may be more likely to shift into a moral repair mode that demands more of them. The heightened threat prompts them to closely examine their wrongdoing and do what is needed to make peace with others (e.g., punishing oneself and providing restitution to the victim). Providing some support for this hypothesis, Fisher and

Exline's (2006) finding that self-condemnation was associated with more distress (consistent with moral repair) followed a manipulation that asked participants to recall a "fairly serious" offence; whereas the studies finding a reduction in guilt as a result of self-punishment (consistent with moral cleansing) did not make such a specification (Bastian et al., 2011; Inbar et al., 2013).

Although a serious or substantial threat may prompt moral repair, it is also true that when the possibility of rejection is high or when repair is considered difficult or risky, transgressors tend to react defensively (Cibich, Woodyatt, & Wenzel, 2016; De Hooge, Zeelenberg, & Breugelmans, 2010, 2011; Gausel et al., 2012; Leach & Cidam, 2015). Similarly, although generally transgressors prefer to directly counter a moral threat by amending the behaviour or affirming oneself in the relevant moral domain (Stone, Wiegand, Cooper, & Aronson, 1997), one will affirm oneself in a unrelated domain when a direct route is not available or is too costly (Zhong, Liljenquist, & Cain, 2009). Therefore, particular features of the transgression that make it difficult to confront may cause transgressors to deem it too risky to engage in moral repair, instead side-stepping the threat via moral cleansing.

Following from this, the function of self-punishment might depend on whether other actions are available or not; that is, whether self-punishment is the only way to deal with the transgression or whether there is an opportunity to repair the harm in another way. Nelissen and Zeelenberg (2009) argue that self-punishment is utilised *only* at times when victim restitution is not possible. In such cases, they claim, self-punishment is the only possible outlet for transgressors to express their remorse. However, Nelissen and Zeelenberg's paper does not provide strong evidence for their hypothesis. In Study 1, post-hoc analyses appear to indicate that participants were equally likely to self-punish (hypothetically denying themselves a ski trip with friends) whether there was an opportunity to repair the

transgression or not[2]. Similarly, Watanabe and Ohtsubo (2012; Experiment 2) found that the willingness to self-punish did not increase when the opportunity to apologise was removed. These results suggest that self-punishment is equally likely despite reparability—contrary to Nelissen and Zeelenberg's argument (2009).

Yet, reparability might influence the process underlying self-punishment by moderating whether self-punishment cleanses or repairs. If a transgressor is faced with a situation that is difficult or impossible to fix (e.g., where he or she doesn't have the opportunity to apologise), it is perhaps untenable to go down the route of moral repair. Constantly rethinking one's wrongdoing when there is nothing else to be done about it may lead to feelings of hopelessness and a cycle of self-blame. In these cases, being able to rid oneself of the burden of guilt may be a far more useful strategy. Thus, when repair is not available we might expect self-punishment to be driven by moral cleansing rather than moral repair. In other words, constraints on reparability may shift transgressors' specific motivation for self-punishment, rather than their overall desire to self-punish.

Third, the particular type of self-punishment utilised may influence its observed effects. Consider the fact that in the literature, reductions in guilt have so far occurred as a result of *physical* self-punishment (Bastian et al., 2011; Inbar et al., 2013)—in contrast to findings that *psychological* self-punishment is associated with poorer psychological wellbeing (e.g., when self-punishment is defined as the desire for suffering, self-hate and anger; Fisher & Exline, 2006). Different forms of self-punishment may lend themselves better to different aims. Perhaps it is easy to rid oneself of one's sense of immorality through

[2] In Study 1, those in the guilt without repair condition self-punished more than those in the control with repair condition (p = .02), but the guilt without repair condition and the guilt with repair conditions did not significantly differ from one another. Thus, there is only weak evidence for their proposition. In addition, Nelissen and Zeelenberg (2009) report another study that does not appear to be designed to test their hypothesis that self-punishment is a last resort option. In Study 2, those in the no repair condition were able to self-punish (deduct points from self in a game), but those in the repair condition had no opportunity for self-punishment, only for compensation. Therefore it is impossible to compare the repair and no repair conditions on self-punishment, which is necessary to assess their hypothesis. The findings only indicate that self-punishment does occur when there is no opportunity for repair, but it does not confirm that it *only* occurs, or occurs with more frequency/intensity, when there is no opportunity for repair.

a physical self-punishment that can provide a distraction from one's psychological distress. In contrast, though it is conceptually possible that ruminating on self-deprecating thoughts or feelings of guilt could make one feel better (analogous to the reported cathartic effect of grief, see Cutcliffe, 1998), psychological self-condemnation seems less amenable to a process of cleansing. Thus, in reality, individuals might choose their mode of self-punishment depending on their motivation to self-punish, but in experimental designs they are all forced to self-punish using the same mode (e.g., physical self-punishment), which may result in effects specific to that mode. Experimental findings are yet to be generalised to a wide range of self-punishment behaviours.

Overall, self-punishment may vary in its effects (whether it cleanses or repairs) depending on characteristics of the self-punisher, the transgression, and the self-punishment. Importantly, the way that perpetrators interpret their wrongdoing may influence their motivation for self-punishment—and the way in which it can restore their moral identity.

The Social Value of Self-Punishment

Although I propose that from the transgressor's perspective, the experience of self-punishment is one of restoring one's personal moral identity, self-punishment may also have an interpersonal function that we should not overlook. Indeed, in line with a social functional view of emotions moral and identity (Baumeister & Leary, 1995; Brewer & Caporael, 2006; Keltner & Haidt, 1999; Leary, 2004; Sachdeva et al., 2009; J. L. Tracy & Robins, 2004), the intrapersonal processes detailed in the previous sections would have evolved to solve a social problem. That is, the suggestion that self-punishment regulates moral identity is only a partial explanation for self-punishment; one must also consider why moral cleansing or moral repair would have been selected for in the first place. In other words, the desire to self-punish, and the signal to do so triggered by moral identity threat, is likely to have been underwritten by a tangible social benefit. To gain a comprehensive understanding of why people engage in self-

punishment we must also consider how other people respond to witnessing transgressors punishing themselves—and thus how self-punishment addresses the threat to one's social standing.

First it is worth reviewing the studies that demonstrate that self-punishment has an interpersonal dimension, for clues about how the how its value may be conferred. In one of the earliest demonstrations of self-punishment in the laboratory, Wallace and Sadalla (1966) found that transgressors (participants who were induced to ostensibly break an experimenter's apparatus) were most likely to self-punish when the victim knew about the transgression (i.e., the transgressor was "caught"; relative to a "non-caught" transgressor). Interestingly, this is in direct opposition to more recent research finding that participants were more likely to self-punish as a result of secret transgressions than known transgressions (Slepian & Bastian, 2017). These studies, however, speak to the types of transgressions that may elicit self-punishment (i.e., those that have interpersonal consequences, or those that have not been resolved through other means)—implying that one might be motivated to self-punish in order to avoid interpersonal rejection or remedy a social threat. These studies, however, do not definitely identify who the audience of the self-punishment may be, or how self-punishment can address that threat.

Other research indicates that self-punishment is aimed at victims (Nelissen, 2012). In this study, participants first played a performance task, in which they were led to believe they had let another player down. Participants were then instructed to select a shock level that would be administered to them in the next part of the experiment, serving as the measure of self-punishment. Participants selected the shock level either while alone, or in the presence of their partner from the previous task (the ostensible victim), or in the presence of an unknown person. The intensity of self-punishment was highest when the victim of the transgression

was present, underlining that the presence of a victim audience can boost self-punishment's appeal.

Equally, some scholars have argued that uninvolved third parties can be target audiences for self-punishment, so long as they are aware of the transgression (Zhu et al., 2017). Zhu et al. (2017) point out that Nelissen's (2012) victim audience knew about the misdeed, and that this might explain why this audience was favoured over the unknown person audience (who was unaware of the misdeed). The researchers tested their proposition in a set of studies. Participants were more likely to deduct points from themselves in a cooperative game when the other "player" was expected to play with them in a future round of the game (demonstrating that self-punishment has interpersonal value). Importantly, this occurred regardless of whether the other player was believed to be a person they had previously spurned (Study 2) or an impartial witness to the transgression (Study 3). Moreover, participants explicitly believed that punishing themselves would repair their reputation in the other player's eyes (Zhu et al., 2017). In sum, these studies indicate that self-punishment has a perceived utility in maintaining positive relationships with both victims and third parties.

Broader research on the expression of pain provides further evidence that self-punishment may have social utility. Physical pain has the ability to arouse empathy and social support from others (Craig, 2009; see also Bastian, Jetten, Horney, & Leknes, 2014, for a review of interpersonal responses to pain). Additionally, transgressors who express their pain are perceived as less blameworthy than those who respond stoically to their pain (Gray & Wegner, 2010). Even the tendency to self-harm may have been selected for as an adaptive strategy for leveraging help from social partners (Hagen, Watson, & Hammerstein, 2008; Nock, 2009; P. J. Watson & Andrews, 2002). Moreover, it has been suggested that self-criticism is a strategy to deal with social insecurity (Zuroff, Moskowitz, & Côté, 1999);

observers consider self-critics to be poorer functioning individuals, yet also more desirable for future interaction (Powers & Zuroff, 1988). Self-punishment may similarly elicit forgiveness and social acceptance.

As already detailed, punishing oneself could be a communication directed at others (Nelissen, 2012; Zhu et al., 2017). However, this does not shed much further light on whether self-punishment is a process of moral repair or cleansing at the intrapersonal level—both processes might benefit from an audience. In moral repair, whereby the transgressor is acknowledging responsibility for violating a shared value, audiences that are aware of the transgressor's wrongdoing and likely to interact with them—that is, those holding the transgressor accountable for their misdeed—are best placed to offer the offender the sense that he/she will be forgiven or welcomed back into the moral community. Equally, in trying to move on from the transgression, moral cleansers' best strategy is to appeal to any transgression-aware and potentially vengeful audiences in order to save themselves from retribution. This is in line with the characterisation of self-critics as manipulators who exploit others to be spared attack (Goffman, 1955; Jones & Pittman, 1982). Thus, regardless of whether self-punishment is motivated by moral repair or moral cleansing, moving the behaviour to the public sphere—when the audience is in a position to reinstate the offender's moral image (or to issue sanctions for the misbehaviour)—may boost its value beyond what one might expect in private.

The notion that self-punishment has social value does not imply that it requires an audience to fulfil its function: The identity-regulating function can still occur at the intrapersonal level if it has been internalised. Indeed, there is evidence that self-punishment occurs even when it is anonymous and cannot possibly serve any immediate interpersonal goal (Tanaka et al., 2015). And in the aforementioned study by Wallace and Sadalla (1966), self-punishment still occurred when the transgression was not known to others. Nevertheless,

it may simply be more effective or more frequent when the audience is aware of what has occurred and is in a position to validate the affirmation (i.e., "it's ok, you are forgiven"). In any case, if self-punishment has the potential to communicate something to others that might earn perpetrators forgiveness, what exactly is that message?

Self-Punishment Can Address Observers' Symbolic Justice Concerns

While restoring moral identity may be paramount to perpetrators of transgressions (SimanTov-Nachlieli et al., 2013), victims and third parties may have other concerns. Rather, others are motivated to punish transgressors (Nagin, 1998; Paternoster, 2010) and may demand that perpetrators repair the harm done (Exline, Deshea, & Holeman, 2007; Witvliet et al., 2008). How exactly would engaging in self-punishment transform observers' outrage into forgiveness and social acceptance? One way to answer this question is to view self-punishment as an act of justice restoration (as argued by van Bunderen & Bastian, 2014). According to justice restoration theory, victims and third parties are motivated to remedy *symbolic justice concerns*, that is, the symbolic harm caused by the injustice. These are twofold: Transgressions (1) undermine the victim's status and power in the group, and (2) violate shared moral values (Okimoto & Wenzel, 2008). I now consider how self-punishment may address these goals.

First, research indicates that although observers of injustice may demand repair, they primarily wish to see the perpetrator suffer through punishment as retribution (Baumeister, 1997; K. M. Carlsmith, 2006), which speaks to the desire to address status/power concerns (Okimoto & Wenzel, 2008). Self-punishment derogates the transgressor (Exline, Root, Yadavalli, Martin, & Fisher, 2011), returning status/power to the victim and moral community who are empowered to accept or reject the transgressor (Okimoto & Wenzel, 2008). Through this redistribution of status/power, self-punishment thus validates the victim's worth (Murphy, 2007) and re-establishes the validity of the broader moral order. In

line with this, self-criticism has been conceptualised as a submissive cue that can signal to others that the transgressor "gives up" and is no longer a threat to the social ranking order (Sloman, Price, Gilbert, & Gardner, 1994; Zuroff et al., 1999). Self-punishment may thus communicate that a transgressor does not feel superior to others, taking one's rightful place in the social hierarchy. Furthermore, since the status/power balance has been re-established through the offender's self-punishment, the suffering debt paid, others may be less likely to retaliate with their own punishment.

Second, self-punishment may address observers' justice concern about the violated values. Research has indicated that self-punishment is perceived by others as a sign of the transgressor's willingness to comply with social norms (Tanaka, Ohtsuki, & Ohtsubo, 2016). In this sense, self-punishment acts as a signal of remorse (Gold & Weiner, 2000). The act of voluntarily punishing oneself is a strong and costly signal that one endorses the values at stake and is willing to reform. Moreover, by labelling their own behaviour as immoral, self-punishers can assure observers that the values they all hold (i.e., the shared values that define the group) are legitimate (Bibas & Bierschbach, 2004; also see Okimoto & Wenzel, 2009).

In sum, self-punishers might regain social acceptance by communicating a commitment to group values and a relinquishment of status/power. However, even an act as seemingly unambiguous as an apology is often challenged by observers (Skarlicki, Folger, & Gee, 2004; Struthers, Eaton, Santelli, Uchiyama, & Shirvani, 2008). Thus, there are likely to be some caveats to this strategy. For example, when transgressors select the nature and severity of their own punishment, others may perceive a lack of true suffering, or even that transgressors are gaining pleasure from their behaviour. Or, perhaps more critically, observers may harbour suspicions about the motives underlying the self-punishment.

There is an intersection here between the intrapersonal (identity-regulating) and interpersonal functions of self-punishment. If self-punishment is perceived as an act of moral

repair in which the offender is thinking through one's wrongdoing, then observers' justice concerns may be readily allayed (i.e., perceiving that, "yes, the transgressor is truly committed to reform"). On the other hand, if observers interpret self-punishment as an act of avoidance, then they may doubt that offenders are willing to revise their values. This leaves the threat to shared values largely unaddressed. Therefore, perceptions of justice might be enhanced when self-punishment is perceived as being motivated by genuine moral repair, as opposed to moral cleansing.

Summary and Overview

In summary, I propose that self-punishment can be conceptualised as two distinct processes that can defuse the threat to moral identity posed by one's wrongdoing. Either self-punishment reflects an act of moral cleansing, through which transgressors can swiftly cleanse their guilty conscience; or it is one of moral repair, through which transgressors actively engage with their wrongdoing by affirming the wrongness of the act. To the extent that transgressors are perceived as engaging in genuine moral repair, then observers should be convinced that they are truly committed to their group and the rules that define it—facilitating forgiveness. In this thesis I examine why people self-punish and self-punishment's effects on both transgressors and observers.

As a starting point, in Chapter 2 I review the difficulties of conducting self-punishment research, suggesting that due to these obstacles the current literature may not adequately reflect self-punishment as it occurs naturalistically. I explore some of the important themes in the experience of self-punishment from the perspective of self-punishers themselves using a qualitative approach. I find evidence of the same tension identified in the literature: Self-punishment seems to reflect both avoidance and exploration of guilt.

In light of the methodological hurdles identified in Chapter 2, in Chapter 3 I review a set of criteria researchers should consider in conducting self-punishment research in the

laboratory. I pilot four different self-punishment paradigms in order to determine which of these tasks may be valid and reliable for further experimental study.

In Chapter 4 I explore the model of self-punishment as an identity-regulating process, according to which self-punishers may be acting out of a desire to *cleanse* or to *repair*—moderated by the extent to which transgressors acknowledge the threat to their moral identity. I test this model across four experimental studies by assessing the effects of self-punishment on emotional, cognitive, and interpersonal outcomes relevant to the two hypothesised processes, and across various self-punishment paradigms.

Chapter 5 marks a shift in perspective, from the effect of self-punishment on transgressors to its effect on observers. In this chapter I elaborate on the interpersonal functions of self-punishment from a justice framework. I suggest that self-punishment is a beneficial strategy for transgressors to gain forgiveness from third parties, and findings indicate that this effect appears to be relatively robust—yet subject to some degree to the attributions made about transgressors' motivation for their self-punishment.

Finally in Chapter 6 I discuss the significance and implications of my findings in the context of the existing literature, and provide avenues for future research that would extend and complement the research reported here. Suffering may indeed purify, and the particular ways in which self-punishment restores a perpetrator's moral identity has significant implications for perpetrators, victims, and observers of wrongdoing.

CHAPTER 2: Why Do We Self-Punish? Perceptions of the Motives and Impact of Self-Punishment Outside the Laboratory

People are motivated to protect their self and social image (DeWall et al., 2011; Leary, Raimi, Jongman-Sereno, & Diebels, 2015). We have cultivated various self-serving biases in order to maintain a positive view of ourselves and to present ourselves favourably to others (e.g., self-enhancement and self-protection, Alicke & Sedikides, 2009). In the context of immoral behaviour, we tend employ defensive techniques to evade responsibility for our actions and thereby protect our self and social integrity (Bandura, 1990; Gausel & Leach, 2011; Woodyatt & Wenzel, 2013a). However, this is not always the case. In response to their own transgressions some people engage in practices of self-punishment. Rather than protecting the self from threats of failure they appear to indulge in them, subjecting themselves to aversive experiences such as physical pain or psychological self-blame. Indeed, self-punishment holds an enduring place in intuitive conceptions of justice and responses to sin (van Bunderen & Bastian, 2014).

But why do people actively cause themselves harm, pain or discomfort, and in doing so identify themselves as transgressors, risking both their self and social image? Though empirical research is scarce, existing data suggest that self-punishment, when experimentally induced, may reduce guilt (Bastian et al., 2011; Inbar et al., 2013). Others have suggested that it may be a communication of remorse to victims or observers (Nelissen, 2012; Tanaka et al., 2015; Watanabe & Ohtsubo, 2012). There is some question, however, as to how accurately these studies reflect self-punishment as it occurs outside of experimentally induced settings. The present study thus employs a qualitative approach to examine whether existing psychological theories are commensurate with self-punishers' understanding of their behaviour.

Limitations of Experimental Methods in Self-Punishment Research

Conducting self-punishment research is difficult. Researchers have creatively constructed a range of experimental procedures in order to examine the function of self-punishment; nevertheless, these procedures are limited in several ways—and these limitations lead to the question of whether these experiments accurately represent when, why, and how people punish themselves. Limitations apply to three stages of the research process: the induction or priming of a transgression; the opportunity (or rather, imposition) of self-punishment; and, finally, the meaning attributed to the self-punishment task.

Due to principles of ethical research, experimental research on self-punishment is limited in how benign its transgression and self-punishment manipulations may be. As a presumed motivating cause for self-punishment, researchers normally first induce participants to commit a moral transgression or feel guilty in some way before providing the opportunity to self-punish. Examples in the literature include autobiographical recall of past wrongs (Bastian et al., 2011; Inbar et al., 2013), hypothetical wrongs (Nelissen & Zeelenberg, 2009), poor performance on group tasks (Nelissen, 2012; Nelissen & Zeelenberg, 2009), and a forced unfair allocation in resource games (Ohtsubo et al., 2014; Tanaka et al., 2015; Watanabe & Ohtsubo, 2012). In the context of these minor, distal, and at times unintentional transgressions, it is unclear whether participants would indeed be motivated to punish themselves. In this way, it is not clear if the experimental research captures when people choose to self-punish in reality.

Similarly, self-punishment tasks that follow transgression manipulations are necessarily mild and artificial. Researchers have examined a number of self-punishment behaviours that are relatively simple to produce in the laboratory, including denial of pleasure (Nelissen & Zeelenberg, 2009), giving up one's money (Ohtsubo et al., 2014; Tanaka et al., 2015; Watanabe & Ohtsubo, 2012), and willingness for or infliction of physical pain (Bastian

et al., 2011; Inbar et al., 2013; Nelissen, 2012; van Bunderen & Bastian, 2014). Little is known about what individuals themselves define as self-punishment and how different these expressions may be to experimental paradigms. As a result, it is not clear whether experimental procedures reflect how people choose to self-punish in naturalistic settings.

Due to constraints on how self-punishment may be investigated in the laboratory, researchers have been compelled to devise subtle and obscure self-punishment tasks that often create incidental confounds, either because of the nature of the task specifically or because of people's attribution about the task. Such confounds give rise to logical alternate explanations of the findings. First, the nature of self-punishment tasks themselves can be confounded by how self-inflicted the aversive stimulus truly is. Though between-subjects experimental designs are desirable for making causal inferences, in the case of self-punishment the requirement for participants in the experimental condition to punish themselves (relative to a control condition) is complicated by the fact that it should be self-inflicted. It is not enough to simply punish participants with a negative stimulus, since this would reflect third-party punishment, not self-punishment. Importantly, the experience must be felt to be in participants' control as much as possible; yet, if it were fully under their control they would be able to choose to take the option of punishing themselves, or not, and thus self-select into the self-punishment conditions, undermining principles of experimental randomization. Or, participants might be given control over the intensity of the self-punishment (e.g., the measured time that one's hand is left in a bucket of iced water, Bastian et al., 2011; selected electric shock voltage, Inbar et al., 2013), where again the design would be turned into a correlational one that limits causal inference, in addition to the problem of significant interindividual variation in sensitivity to physical stimuli (Fagius & Wahren, 1981; Nielsen, Staud, & Price, 2009).

Additionally, confounds can be introduced into experimental self-punishment research through the uncontrolled attributions participants may make when faced with the task. For example, in Nelissen's (2012) study, the measure of self-punishment was participants' willingness to receive electric shocks for incorrect answers on a knowledge test, where they were explicitly told that aversive stimuli could improve cognitive performance. A higher selected shock level, then, can be interpreted by the participant not as a measure of punishment, but rather a threat of aversive consequences that serves as an incentive to motivate their performance on the knowledge test. Another confound in past studies is that the self-punishment may result in benefits to others. For instance, punishing oneself by forfeiting tickets in a lottery (Nelissen & Zeelenberg, 2009) has the indirect effect of increasing the chances of other players (including the ostensible transgression victim) winning the prize. Likewise, the amount of money sacrificed as self-punishment in research by Tanaka et al. (2015) presumably went back to the experimenter; as a consequence, the measure could reflect altruism or amend-making instead of, or in addition to, self-punishment. In this way participants may be able to interpret these tasks in various ways: as an opportunity to repair, a type of game or an opportunity to show off their physical toughness, for example. This criticism is important because the way participants interpret why they are doing the particular experimental task is likely to lead to different outcomes, and thus the consequences or functions of self-punishment may differ in naturalistic settings compared to experimental settings.

These three sets of limitations may in turn impact on each other to affect the observed outcomes. At the very least, then, we would need to be cautious in interpreting findings from such experimental methods as to what they tell us about self-punishment as it occurs in life. We would be wise to complement these methods with research that observes self-punishment in naturalistic settings, which is the aim of the present study.

Existing Research Findings

Experimental research has primarily focused on the emotion-regulating function of self-punishment. Some researchers have suggested that self-punishment arises from feeling guilt and is used as a means to alleviate guilt, in line with a substantial qualitative and quantitative literature that points to affect regulation as a key motivator for physical self-harm (Chapman et al., 2006; Klonsky, 2007). Research supports the notion that guilty and shameful individuals are more likely to punish themselves (Bastian et al., 2011; Inbar et al., 2013; Nelissen & Zeelenberg, 2009; Tanaka et al., 2015; Wallington, 1973), and that in turn self-punishment appears to reduce guilt (Bastian et al., 2011; Inbar et al., 2013). If self-punishment does reduce distress, it might do so through distraction or dissociation from one's "tainted" identity, as is seen in masochism (Baumeister, 1988). Alternatively, self-punishment may reduce distress as a result of a socially learned model of pain and justice (van Bunderen & Bastian, 2014). People may learn to punish themselves through the modelling provided when parents punish them as children, and punishing oneself may reduce external punishment (Stanciu, 2015). People may thus be socialised to feel relieved after punishment. However, the mechanism for a reduction in guilt is not yet clear (Baumeister et al., 1994).

Moreover, some studies have found that psychological self-punishment (e.g., self-condemnation or self-criticism) is associated with *more* psychological distress (Fisher & Exline, 2006; Whelton & Greenberg, 2005). Interestingly, these studies employed non-experimental methods. That is, self-punishment was measured, not imposed. This begs the question of whether this design—where one's punishment is genuinely self-imposed—has more accurately captured the experience of self-punishment, or whether the two conflicting accounts can be somehow consolidated.

Setting aside the intrapersonal effects of self-punishment, other research has pointed to the interpersonal implications of self-punishment. Nelissen (2012) found that self-punishment is most likely to occur in the presence of a victim, relative to no audience or a third party. Another study demonstrated that individuals overtly criticising themselves were perceived as more vulnerable and favourable for future interaction by observers (Powers & Zuroff, 1988). Together these findings suggest that self-punishment has social value, but what exactly might that social value be?

If self-punishment does reduce distress, there are clear social implications. Theoretically, a reduction in guilt should inhibit reconciliatory action, since the improvement in one's self-image can undermine the motivational drive that guilt has in instigating compensatory behaviour to victims (Baumeister et al., 1994; Woodyatt & Wenzel, 2014). Similarly, self-punishment's social signal may be a communication of remorse that pre-empts and reduces the threat of external punishment, allowing transgressors to release themselves of further reconciliatory action.

However, it is possible that self-punishment has a more reparative social function. Nelissen and colleagues (Nelissen, 2012; Nelissen & Zeelenberg, 2009) argue that self-punishment is a form of reconciliatory action in itself: a show of remorse to victims in order to repair the relationship, elicited only when there is no direct way to repair the harm. Their empirical data, however, only provide limited support for their claim. In Nelissen and Zeelenberg's (2009) Study 1, for example, post-hoc analyses (though not explicitly reported) appeared to indicate that self-punishment was not significantly different whether repair was possible or not; Study 2, by design, did not allow for their claim to be tested. More recently, van Bunderen and Bastian (2014) examined downstream social consequences of self-punishment, finding a differentiated effect on victim reparation. For those low on victim justice sensitivity (i.e., who were less likely to view themselves as victims of injustice), self-

punishment did not reduce the motivation for repair. Indeed, while not anticipated by the authors there was a curious trend in the opposite direction, with self-punishment *increasing* repair (though the simple slope was not statistically significant). It may be that for some individuals, self-punishment might not be so much ego-protective as instead reflecting a genuine attempt to declare one's wrongdoing and commitment to the moral order; in these cases, self-punishment sustains the motivation to engage in reconciliatory behaviour.

Study 2.1[1]

Existing experimental research on self-punishment is limited by ecological and construct validity issues in self-punishment experiments. Plausible functions have been inferred from observed result patterns, but how accurately do they reflect the phenomenon? How do individuals describe and understand their own self-punishment behaviour?

The aim of the current study was to extend self-punishment research beyond the laboratory. We adopted a qualitative approach to explore, describe and interpret the lived experience of self-punishment from the perspective of self-punishers. This constitutes a test of whether naturalistic accounts of self-punishment are consistent with the existing literature.

Method

An online survey was undertaken to explore individuals' experience and understanding of self-punitive behaviours. Key measures were open-ended items, though some quantitative measures were included as well to complement the observations.

Participants

After receiving approval from the university ethics board, we approached a general community sample of Australian residents via an online survey pool (Qualtrics). Respondents gave their informed consent prior to commencing the survey, which outlined their right to

[1] Though this is the first study reported in this thesis, Study, Figure, and Table numbering follow Chapter numbering for consistency.

withdraw at any stage, confidentiality/anonymity, and the potential for the topic material to cause some emotional discomfort.

Participants were asked whether they could recall ever having punished themselves, as a screening item for inclusion. We cast a relatively wide net, aiming for about 100 self-punishers. This *N*, while larger than used in traditional qualitative research, permitted reliable use of our quantitative measures, and allowed greater breadth in terms of the contexts, circumstances and meanings of self-punishment we hoped to uncover. Of 264 individuals entering the survey, 127 were unable to recall any self-punishment behaviour, and thus were excluded from the study. Fifty-seven participants were removed for failing validity checks.[2] The final sample consisted of 80 adults (52.5% female, age ranging from 18 to 80 years, mean age 35.3, *SD* = 16.4), most of whom were Australian citizens (93%), with Australian or New Zealand nationality (70%; other sizeable nationalities were 11% mixed Australian/European and 6% European).

Procedure and Materials

Participants completed an online survey assessing perceptions of self-punishment. As self-punishment is linked to self-conscious emotions of guilt and shame, we judged that the privacy and anonymity of an online format may reduce social desirability response bias and increase the openness with which respondents discuss sensitive topics (Rhodes, Bowie, & Hergenrather, 2003). Moreover, we decided against in-depth interviews at this point (necessarily based on limited cases) and aimed at a larger sample (of necessarily briefer descriptions), for the reasons explained in the previous section—that is, aiming for breadth.

After answering some basic demographic items, survey participants were asked to recall their self-punishment behaviours (limited to three instances). The following items

[2] There were two automatic validity checks that instructed participants to follow a simple instruction, for example "select the first option for this item", that eliminated 47 respondents. A further 10 were manually removed because they entered invalid (nonsense) responses into text fields.

explored the event that triggered the most serious of these instances. This included a free recall of the transgression, and quantitative ratings on characteristics such as severity, importance of the relationship with the victim, guilt, and shame (all measured on 7-point rating scales from "not at all" to "very much," with the relationship item including an eighth option allowing "n/a- I didn't hurt a specific person"). The self-punishment was rated in its unpleasantness, also on a 7-point rating scale.

Following this, open-ended questions probed participants' understanding of a single instance of self-punishment. The first item was the broadest, constituting the central question of the research (Creswell, 1998): *Why do you think you chose to punish yourself in this instance?* Subsequent items were designed to explore the effects indicated in the literature regarding intrapersonal and interpersonal goals, i.e., *How did the self-punishment make you feel? How did punishing yourself change the way you thought about or saw yourself? How did the self-punishment affect your relationships with others (e.g., the person you hurt)?*

The final section of the survey moved away from participants' own behaviour and assessed general perceptions of self-punishment (i.e., *Generally speaking, why do you think people self-punish?*). By dissociating themselves from the behaviour, it was hoped that participants would respond to the general item with a slightly different perspective; for example, perhaps they would be more critical of others' behaviour relative to their own.

Analytic Approach

We conducted a phenomenological investigation of the experience of self-punishment. In order to most faithfully describe self-punishers' accounts, for the most part we took what they said at face value, although we also understand that participant's accounts are essentially perceptions and beliefs rather than facts. Qualitative research can adopt a range of epistemological approaches, spanning positivist/realist approaches that prioritise the standards of reliability and objectivity, through to radical constructionism, which rejects the

possibility of discovering any objective "truth." Hammersley (2002) rejects the idea that qualitative research must take either a naively realist approach, which fails to examine how people experience and understand their social world, or a radically constructivist approach, which falls into a relativist trap whereby all experiences are treated as unstable and ultimately impervious to analysis. Instead, he proposes a "more subtle form of realism" (Hammersley, 2002, p. 73). Hammersley's approach can be seen as a form of contextual constructionism (Madill, Jordan, & Shirley, 2000; Parker, 1994), which acknowledges that participants' narratives are always partial and subjective, but at the same time allows for the identification of consistent patterns of meaning that emerge from the data. Here, we follow Hammersley's approach and accept participants' views as being "more or less" (2002, p. 73) accurate reflections of their experiences of self-punishment. This approach is consistent with our aim—as the first qualitative study in the area—to give voice to self-punishers, who may not have been represented in experimental research. To further facilitate this aim, the thematic analysis was inductive; analysis was a process of coding data without trying to fit them into pre-existing theories, so that the themes identified were strongly grounded in the data (Glaser & Strauss, 1967/2012; Patton, 1990).

Though we wanted to capture any experiences missing from the literature, we also sought to confirm whether naturalistic accounts of self-punishment are consistent with the existing literature. Thus, guided by the nature of past findings, we attended to information about the emotional, cognitive, and interpersonal motivations and outcomes of self-punishment. This is reflected in the design of several of the survey questions (e.g., *How did punishing yourself change the way you thought about or saw yourself?)*, and undoubtedly played a part in the way the data were interpreted. Nevertheless, we were not invested in proving or disproving any particular theory. Additionally, we were not interested in rigidly fitting the data to existing theories; as we have pointed out, there might be significant

knowledge gaps in the current literature due to the inherent problems in conducting self-punishment research. Indeed, this was our motivation for carrying out the research. In this way, our preconceptions were "permeable" (Madill et al., 2000; Stiles, 1993): We sought to consider how the data fit with past findings and how it might change our understanding of self-punishment, such that our findings would have reflexive validity. As we will detail shortly, the findings both validated some aspects of the existing literature while also revealing some neglected experiences that necessarily expand our theoretical understanding of self-punishment.

We conducted a thematic analysis following the guidelines proposed by Braun and Clarke (2006). Open-ended items were first analysed either individually or in blocks of items, depending on the nature of the question being asked. For example, the broad item asking why people self-punish was analysed as a single item, whereas questions relating to the outcomes of the self-punishment were grouped (e.g., items about its effects on feelings and thoughts coded in one block). Some data were analysed simply (e.g., coding negative or positive for effect on relationships), whereas items eliciting detailed responses were subject to thematic coding. First, the full data corpus was read several times, during which time the researcher assigned initial *descriptive* codes to segments of information. Initial codes were then collated into coherent semantic themes based on interpretation of the data—this step was where the *interpretative* analysis began to occur. Finally, the dataset was examined holistically; themes that appeared consistently were designated as overarching themes, and the links between themes were explored. Themes were developed and revised in an iterative process through engagement with the data, making sense of semantically related and/or contradictory information to produce coherent and meaningful themes. Ad hoc analyses were then conducted to explore the themes extracted.

Reliability of the thematic analysis was verified by having a second independent researcher code data from a randomly selected 25% of participants, using the frameworks developed through the inductive analysis (see Appendix A). In order to provide additional support for the trustworthiness of the results, the second coder was unaware of the study aims as well as the broader literature about self-punishment. A reliability estimate was calculated for each item or set of items. Where categories were mutually exclusive, inter-rater reliability was calculated using Cohen's kappa (all $\kappa > .80$). Where coding schemes allowed for multiple categories, reliability was calculated using percentage agreed responses (all > 80%).[3]

Descriptive quantitative statistics provide context and support for the qualitative results.

Results

Descriptive Data

Self-punishment behaviours. Of the 264 participants initially sampled, 137 were able to identify engaging in self-punishment (52%) before excluding the 57 who failed attention checks.[4] Text responses were analysed to determine what kinds of behaviours individuals engaged in as self-punishment. Behaviours were coded and categorised into behaviour types using terminology from existing literature and common-sense notions. Table 2.1 displays the frequency of self-punishment behaviours reported in the sample.

[3] For each block, each participant response was given a weighting of 1, which was counted as agreement (1) or disagreement (0). Where at least one coder identified more than one theme, each identified theme was given equal weighting, with the total of all weightings always equal to 1; i.e., 1/ *n*, where *n* is the largest number of codes identified by either coder. E.g., if a response was coded by rater 1 as theme A ($n = 1$), and by rater 2 as both themes A and B ($n = 2$), then a score of 0.5 agreement and 0.5 disagreement would be given (i.e., yielding a 50% agreement rate).

[4] This figure should not be considered an underlying rate of self-punishment in the community. In the introduction page to the survey, we described our aim as being "interested in whether people might punish themselves for their transgressions." This description may have led self-punishers to self-select into our survey, biasing estimates upwards.

Table 2.1

Frequencies of Reported Self-Punishment Behaviours

Type	Frequency	Participant examples
Psychological torment[a]	28	• *Mental torture* • *Being really hard on myself about the situation and how I behaved*
Physical self-harm	22	• *Cut my wrists* • *Made myself take an ice-cold shower*
Denial of pleasure[b]	19	• *Restricted luxuries* • *Went without things I wanted*
Food restriction	18	• *Skipped a meal* • *I didn't eat for a week*
Social isolation	15	• *Secluded myself for an extended period of time from the opposite sex* • *Withdrawn from everyone*
Self-sabotage[c]	13	• *Drinking copious amounts of alcohol* • *Not cared for myself*
Other/ Unspecified[d]	9	
TOTAL	124[e]	

[a]Verbal/cognitive self-deprecation. [b]Restricting activities or goods to withhold pleasure. [c]Behaviour not necessarily aversive in itself but disadvantages the self. [d]Category contained few members (7%), therefore the scheme adequately described the sample. [e]Each identified instance of self-punishment was counted, including where participants described several discrete behaviours. Therefore this total exceeds the total number of participants in the sample.

What types of transgressions lead to self-punishment? A broad range of transgressions was reported, with the most common category being a fight or argument. Transgressions were rated as moderately guilt provoking (M = 5.58, SD = 1.46) and shame provoking (M = 5.53, SD = 1.54), and rated as somewhat severe (M = 4.55, SD = 1.79). Participants tended to transgress against quite important relationships (M = 5.98, SD = 1.46), though many participants reported no specific person had been hurt by the transgression

(36.3%), suggesting that some self-punishers may feel like they fail at things generally rather than in response to a specific transgression. This finding was unexpected, but warranted further investigation. Transgression text descriptions were examined and coded as either identifying a specific act, or general failures across time. Interestingly, 29% of respondents reported general failures, although most (61%) could identify a specific act (10% were unclassifiable).

Thematic Analysis

We identified three overarching themes and one sub-theme describing the effects and perceived functions of self-punishment: self-punishment as an emotion regulation strategy, as a normalised behaviour, as an opportunity for reflection and learning, and as moral restoration (sub-theme). The final thematic map is presented in Figure 2.1. Only three participants' accounts were not explained by these themes at any point in their responses.[5] Each of the themes will be presented in the following section, illustrated with text responses from participants (presented in italics; case, age, and gender in parentheses).

Theme 1: Self-punishment as emotion regulation. Consistent with the research literature, most participants described emotion in their narratives of self-punishment. These reports were generally limited to describing the emotion as the precursor to the behaviour, pointing to emotion as the proximal perceived cause of self-punishment. Guilt was the most commonly cited emotion, but there were also references to anger and disappointment.

[5] Two of these participants provided very short responses, stating that they didn't know why they or others punish themselves, and that it did not have any effects on the way they felt or saw themselves. This may reflect that some people self-punish instinctively (consistent with Theme 2), or it could be that these participants were unwilling to share more information with us. Accordingly, we could not analyse these cases in any more depth. The third participant described a suicide attempt. We felt this case was unique because of its severity and it could not be analysed in the context of more generalised self-punishment behaviours.

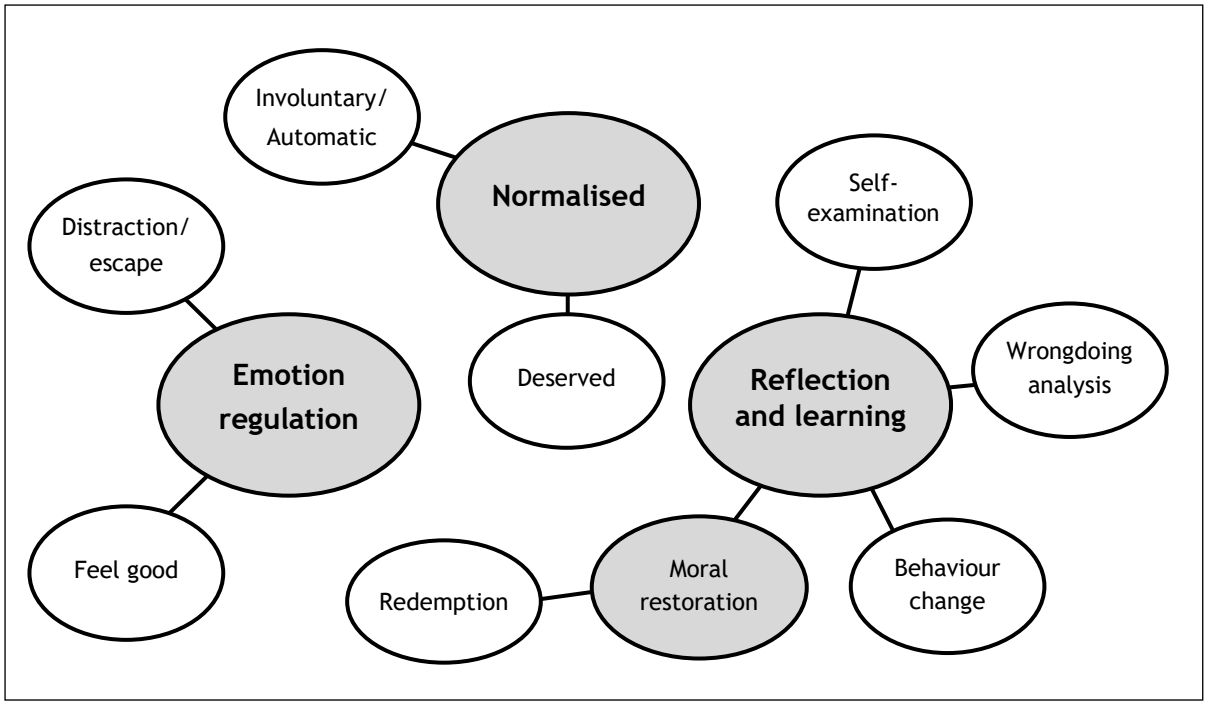


Figure 2.1. Thematic map illustrating perceived self-punishment functions

A number of participants expressed that self-punishment may reduce the emotions related to the transgression. In one participant's words: *Catharsis* (P2: 28, male). Moreover, some participants pointed to possible ways that self-punishment might reduce negative emotions: some suggesting that the experience of self-punishment can distract from other unwanted feelings, for example, *to take their mind of* [*sic*] *the reason for it* (P24: 47, female), and *one pain takes away another pain* (P65: 51, female). The source of the reason for self-punishment was expressed as something undesirable, to be repressed or to escape from, e.g., [self-punishment] *gives them an escape* (P20: 19, male).

Almost all the emotion regulation explanations, however, were found in response to why people self-punish in general, rather than when participants were describing the motivation for their own self-punishment. This suggests that the narrative of self-punishment as a way to purge guilt may be a general perception people hold, but this is not necessarily their own conscious experience.

Furthermore, it was unclear whether self-punishment was a successful strategy to reduce undesirable emotions. There was a notable tone of scepticism around this motivation for some participants, for instance: *I think people self-punish because it gives them a reason to live and it* ***feels as if*** [emphasis added] *you are making all the guilt go by punishing yourself* (P37: 19, male). Indeed, few participants reported that the self-punishment made them feel better. As expressed by this participant: *I felt guilty; wanted to be punished for my wrongdoing as I felt it eases the guilt*. … [My self-punishment made me feel] *even worse; felt that I couldn't even punish myself properly* (P58: 25, male). Perhaps going some way to explain why self-punishment could not always relieve emotional distress, some participants suggested that self-punishments could exacerbate the very self-blame from which they might be seeking to escape, e.g., *It reinforced my shame* (P34: 30, male). This notion—that self-punishment can be an expression and affirmation of one's wrongdoing rather than evasion of it—will be further explored in Theme 3.

Theme 2: Self-punishment as a normalised behaviour. For many participants, self-punishment was described as a normalised behaviour: an expected and inevitable response to one's perceived wrongdoing. Rather than a norm prescribed to others per se (i.e., social/interpersonal norms; though this was at times implied), participants tended to describe this as a norm prescribed at the individual level—as a personal standard of behaviour (Kahneman & Miller, 1986). For example, participants described being compelled by a sense that they deserved their punishment, e.g., *It was an appropriate punishment* (P17: 18, male). While this might seem as though participants were complying with some notion of just deserts, no mention was made of justice or balance. Rather, the deservingness made the punishment inevitable. As expressed by one participant, *deep down inside, I think we all know that we're constantly doing wrong things every day, and deserve some form of*

punishment. I don't think it's healthy to inflict it on ourselves, but I think ***it's to be expected*** [emphasis added] (P50: 20, male).

Normalised behaviours are those that are consistent over time (Kahneman & Miller, 1986), thus they can become habituated and feel automatic or involuntary. A sense of self-punishment as being automatic and unconscious was evident across self-punishers' accounts. For example, a participant explained: *I don't think it* [my self-punishment] *was necessarily a choice I made, but rather something I just started doing* (P50: 20, male). A lack of choice was thus central to this theme, participants expressing that self-punishment was "natural"—one could say inherent—to who they are: *I didn't choose to punish myself in this way. It was a natural reaction* (P1); *I just did it naturally. … that's how people cope* (P68: 34, male). When asked why they thought people self-punish in general, this self-punisher described the utter inevitability of her behaviour: *What else is there to do really?* (P73: 38, female).

Most participants conveying their self-punishment as normalised reported feeling worse as a result of self-punishment. This suggests that self-punishment, when experienced as a normalised response, may be part of a self-reinforcing cycle whereby self-punitive behaviours become habitual after some time even if they are self-defeating. Illustrating this cycle, the following participant described their self-punishment: [I punished myself] *because of my self-confidence issues and mental illness*. … *It just triggered even more self-punishments*. … *My depression got worse and I entered a spiral of negative thoughts* (P61: 25, female).

Theme 3: Self-punishment provides an opportunity for reflection and learning. A large number of participants expressed that self-punishment provides an opportunity to reflect on their wrongdoing and affirm the wrongness of that act. For instance, the following participant describes their self-reflection process: *I feel as if I became a bigger man and my brain registered the acts committed before the punishment as negative so next time a similar*

situation arises, if ever, I will be able to handle the situation better (P7: 32, male). In this sense, the process of self-examination in self-punishment equips transgressors to better navigate future moral choices. In another participant's words: *It's a chance to fully investigate why you acted in that way and most importantly to never let yourself down again as you did on that occasion. Also to think of outcomes before you act, this has helped me enormously* (P76: 77, female). This process bears a strong similarity to value affirmation, whereby individuals affirm that the violated value is important to one's identity (Woodyatt & Wenzel, 2014). This sentiment is echoed directly in one participant's explanation of why they self-punished: *affirmation of responsibility to make sure decisions are correct* (P56: 45, male).

Rather than trying to avoid guilt, this theme suggests that self-punishment can provide psychological resources to actively deal with the source of one's guilt. However, this is unlikely to be an easy task—admitting responsibility for one's wrongdoing could further elicit feelings of guilt or shame (as has been found following value affirmation tasks, Woodyatt & Wenzel, 2014). Indeed, participants who indicated that they learnt from their self-punishment often reported feeling worse as a result. As one participant remarked: [My self-punishment was] *a way of teaching myself not to behave like that around people I care about. … It made me feel guilty and question how good of a person I am* (P79: 19, female). Thus, this theme suggests that for some people, self-punishment is a way of confronting one's wrongdoing and deconstructing the reasons for one's guilt and shame, rather than avoiding it.

Sub-theme: Self-punishment can restore morality. This sub-theme reflects the idea that self-punishment as learning can offer redemption or atonement. Participants who described learning from their transgression often also described a sense of moral restoration. For example, as one participant expressed, *I needed to redeem myself* (P37: 19, male); and according to another, *it gave me time to reflect on my actions as to not do it again in the*

future. … I grew as a person, became more mature, thought about other people and how my actions could impact on them (P54: 29, male). Morally restoration was closely tied to accounts of self-punishment as a learning process; it was the knowledge participants gained about themselves through their self-examination that made them feel morally restored. For example, reflecting on their self-punishment experience, one respondent explained:

> *It made me feel good that I had the willpower to step back from doing something that was clearly hurting others. … I was proud of myself. It definitely changed me for the better. … I became a much better person with a stronger moral compass.* (P35: 26, female)

The process of self-reflection and accepting responsibility for one's actions could at times provide a positive moral self-image, but as revealed in the previous section, this was not always a pleasant process. In one participant's words: [My self-punishment] *made me feel terrible*. … [as a result of the self-punishment I have] *been more thoughtful as to what I say and how I act* (P78: 50, male).

Further exploration of themes. We conducted exploratory analyses in order to investigate links between themes and with other relevant variables. An initial question was whether particular features of transgressions were more likely to give rise to different motivations for, or effects of, self-punishment. Relationships between transgression variables and self-punishment functions are displayed in Table 2.2. The more guilt participants reported they felt about their transgression, the less they reported an emotional regulation motivation for their self-punishment, and the less they found their self-punishment made them feel better. Meanwhile, a higher threat from one's transgression (i.e., higher levels of guilt and shame, and a closer relationship with the victim) was associated with reflection and learning outcomes, and to some extent, a restored sense of morality. Though it is difficult to draw conclusions based on these necessarily messy data (i.e., using dichotomous variables

that were not created for this purpose), these findings tentatively suggest that the function of self-punishment may depend on the extent of the threat posed by the transgression.

Table 2.2

Correlations Between Transgression Variables and Self-Punishment Motivations and Effects

	Point-biserial correlation			
Coded categories	Guilt	Shame	Relationship	Severity
Motivations				
Emotion regulation	-.24*	-.03	-.18	-.13
Normalised	.004	-.08	-.08	-.20^
Learning	-.09	.03	.11	.11
Effects				
Feeling better	-.34*	-.22*	-.51**	-.33*
No change to emotion	-.25*	-.33*	-.22	-.18
Feeling worse	.01	-.05	-.06	-.12
Learning	.22^	.25*	.34*	.19
Restore morality	.10	.20	.26^	.17

Note. All self-punishment variables are coded 0 or 1 for presence in item or item block. $N = 80$ for all cells except relationship importance column, where $N = 51$ (excluding those reporting no specific person hurt).

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

Second, we investigated whether each theme was associated with a particular interpersonal outcome. Due to small cell sizes it was not possible to explore this in relation to each motivation and outcome. One exception was learning as an outcome. Comparing the proportions of participants reporting positive, negative, or no effect on relationships within these cells, a pattern emerged: Those reporting a learning effect tended to also report positive relationship effects (27% positive, 60% no effect, 13% negative), relative to those not reporting a learning effect (12% positive, 49% no effect, 39% negative). A chi-square test of independence revealed that the difference in proportions was significant, X^2 (2, $N = 71$) = 6.45, $p = .04$. This suggests that learning from the transgression may facilitate relationship repair, though the current analysis does not permit a causal claim.

Participants largely ignored the possible interpersonal dimensions of self-punishment; hence no themes focused on interpersonal functions. When asked explicitly about the effects of self-punishment on their relationships, most participants stated that there was no effect, or that nobody knew about the self-punishment (in many of these cases, participants stated, "n/a"). For these individuals, self-punishment was a private affair that did not concern or affect others. Another portion of individuals reported negative effects on their relationships because others did not understand or sympathise with the self-punisher. For example, as one participant expressed, [my self-punishment] *Made them think I was fucked up* (P13: 18, female). In another's words, *They just thought I was trying to get attention* (P40: 48, female). In contrast, a small number of participants reported positive effects on their relationships, mostly through gaining forgiveness, e.g., *It made my brother realise that I was very sorry and I honestly did not mean in any way to hurt him* (P37: 19, male). Thus, self-punishment could sometimes restore morality in others' eyes—a sign of remorse as argued by Nelissen (2012)—though this was not commonly reported.

Discussion

As the first study to examine accounts of spontaneous self-punishment rather than through experimental manipulation, the current study provides novel and naturalistic descriptions of the experience of self-punishment. The findings reveal the heterogeneity of self-punishment; not only did participants describe a broad number of self-punishment types (e.g., from physical harm to food restriction), but we also identified three distinct themes indicating its perceived functions and effects. These themes were: self-punishment as emotion regulation, the notion that self-punishment is normalised, and self-punishment providing an opportunity for reflection and learning (with a sub-theme of moral restoration). A key contribution of the current research is the perception that self-punishment provided an opportunity to reflect on one's behaviour and enact behavioural change in a process of moral

repair. This characterisation is distinct to past experimental findings focusing on alleviation or avoidance of distress. All three themes, however, carry implications for self-punishment theory and research.

Does Self-Punishment Relieve Distress?

The experience of self-punishment is unmistakably emotionally charged. Emotions—particularly guilt—as catalysts for self-punishment were prominent in participants' explanations for their behaviour. This is consistent with the current characterisation of self-punishment as guilt- and shame-induced (Nelissen & Zeelenberg, 2009; Tanaka et al., 2015). However, participants tended to report self-punishment making them feel worse rather than better, suggesting that the narrative of self-punishment as a way to expunge one's guilt (Bastian et al., 2011; Inbar et al., 2013) is not the typical experience outside the laboratory.

It is possible that for some, self-punishment provides some initial relief via pain's ability to distract, but after some time the unresolved conflict could resurface. Research tracking self-harmers' emotions indicates such a pattern: Initial relief of distress supplies negative reinforcement to sustain self-harm, despite the negative consequences that result from neglecting the source of the problem (Brown, Williams, & Collins, 2007, see also Chapman et al., 2006). This is in line with research indicating that unresolved shame can be problematic (Lewis, 1971; Scheff, 1994), as well as broader psychological research on thought suppression and emotional avoidance (Abramowitz, Tolin, & Street, 2001). For individuals trying to repress their wrongdoing, self-punishment may similarly "short-cut" guilt or shame processing, denying transgressors the opportunity to resolve the moral crisis and resulting in grief later down the track. The typically narrow timeframe of experimental approaches may thus be a limitation as it does not permit capturing longer-term effects.

There may be other methodological features of experimental research that could explain the discrepancy between past findings and the narrative of a negative emotional cycle

identified in the current research. Only 52% of participants could recall an instance of self-punishment. In experimental paradigms all participants are randomised to self-punishment conditions. As highlighted earlier, this could be problematic because manipulating self-punishment renders it somewhat involuntary, arguably capturing a different phenomenon. Compounding this issue, participants who do not typically engage in self-punitive behaviour may lack the genuine motivation to punish themselves, potentially influencing the effects of self-punishment tasks.

Also, prior research that has found a reduction in guilt used physical self-punishment (Bastian et al., 2011; Inbar et al., 2013), which may not generalise to other self-punishment types. For instance, psychological torment may be more liable to increase negative emotionality, since the act of telling oneself that one is a bad person may itself be reinforcing of that view; the act of self-punishment in this case is directed towards the negative self-view. This is consistent with research finding psychological self-punishment is related to psychological distress (Fisher & Exline, 2006; Whelton & Greenberg, 2005). In contrast, physical self-punishment may more readily provide a distraction from emotional and psychological distress by directing attention to physical pain. Moreover, these tasks could be easily perceived as tests of perseverance rather than a focus on the pain itself. Thus, some self-punitive responses may be more or less associated with a negative emotional cycle.

If self-punishment represses unresolved conflicts, how can we explain its adaptive function? In other words, under what conditions might this strategy be beneficial? It may be advantageous to withdraw from moral conflicts when they pose a low threat to the self and there is no benefit to be gained from exploring them in depth. This is consistent with Tetlock et al.'s (2000) speculation that when transgressions are less serious, people may be content with a more simple defence. In these cases, self-punishment can be a relatively low cost act that restores one's self-integrity and minimises others' desire for retribution, without the

burden of a comprehensive self-examination or victim restitution. The current data provide some preliminary evidence for this proposition: Participants who felt less guilty about their transgression tended to report an emotion regulation motivation, and were more likely to report feeling better as a result of their self-punishment (feeling better was also negatively associated with shame, severity, and relationship importance). Perhaps this strategy becomes maladaptive when the source of one's guilt is too severe to be easily brushed off—in which cases it may return unresolved. It is interesting to keep in mind the criticism that experimental methodology is limited to inducing relatively benign transgressions. The lower severity acts used in prior research may have biased results towards a distress-alleviating effect, rather than activating other possible functions of self-punishment.

Self-Punishment May Reflect Problem Resolution

One of the major themes identified was the perception that one could reflect and learn from one's behaviour through an act of self-punishment, facilitating moral restoration and moral behaviour. Far from the avoidant characterisation implicated in the emotion regulation theme, this theme speaks to a problem-focused interpretation of self-punishment.

From this perspective, for some individuals self-punishment may constitute a symbolic affirmation of commitment to the values violated by the transgression (Woodyatt & Wenzel, 2014). Such a characterisation is broadly in line with Nelissen's (2012) conjecture that self-punishment is a signal of remorse to victims, a proclamation that one does not condone the wrongdoing. Whether the punishment reaffirms transgressors' commitment to values in others' or merely their own eyes, self-punishment may be fostered by a desire to repair one's moral identity. Research indicates that meeting this need for moral identity by affirming the values violated by the transgression can enable moral engagement and reconciliation (Woodyatt & Wenzel, 2014; Woodyatt, Wenzel, & Ferber, 2017). Participants in the current study who learnt from the transgression reported more positive relationship

effects (relative to those who did not report learning), suggesting that, for some people at least, self-punishment may affirm moral values and the relationships defined by these values.

Though some participants believed that self-punishment allowed them reflection and learning, the precise mechanism for such a process is unclear. Considering social isolation, food restriction or denial of pleasurable activities, a straightforward case can be made: Self-punishment can provide a physical and psychological space in which to contemplate one's deeds. In fact, using self-denial in order to reflect and gain insight is a persistent theme in many religions, which view asceticism as a path to spiritual enlightenment. For the case of psychological torment, such rumination could be useful in analysing complex problems (Andrews & Thomson, 2009). For physical self-punishment, some researchers have argued that pain captures attention and brings cognitive resources online for effective problem solving (Bastian et al., 2014). Therefore, learning can potentially occur in different type of self-punishment via different mechanisms.

Though reflection and learning might be helpful, it is a psychologically demanding process. This strategy is therefore only advantageous when there is a clear benefit to be gained by resolving the moral conflict, rather than by sidestepping the self-threat. For example, the transgression may be particularly severe or involve an important relationship, where there are more expectations of, or benefits to be gained from, repair and reconciliation. The current data support this hypothesis, as learning was positively associated with some of the moral threat markers. However, this remains speculative at this stage.

The Possible Role of Trait Self-Punitiveness

There is a possible link between the emotion regulation and normalised themes: With each self-punitive response that alleviates distress, it is reinforced and becomes normalised over time. This link between habituated behaviour and emotion regulation was noted in a qualitative analysis of self-harm accounts, where "addiction" to self-harm was linked to the

short-term emotional relief it provides (Wadman et al., 2016). In this way, self-punitiveness may be—or become—a disposition. Providing some empirical support for this claim, a large proportion of self-punishing individuals in the current study did not identify a specific person being hurt by their transgression. Rather than being linked to particular transgressions, perhaps some self-punishers have a tendency to find fault in themselves in general.

In line with the idea that there is a self-punitive personality, self-criticism (Powers & Zuroff, 1988)—spontaneous cognitive or verbal expressions of self-derogation, similar to psychological torment—is considered one dimension of trait perfectionism. Thus, self-punitiveness could be part of a broader tendency to self-criticise, manifested in the moral domain. There is some evidence consistent with this proposition. Perfectionism has been linked to suicidal ideation and maintenance of eating disorders such as Anorexia Nervosa (Cassin & von Ranson, 2005; Chang, Watkins, & Banks, 2004); in the current study, there were reports of self-harm and restricted calorie intake as self-punishment. Additionally, some researchers have detected an association between self-critical perfectionism and increased feelings of guilt and shame (Stoeber, Harris, & Moon, 2007; Tangney, 2002), such that, by extension, self-punishment-prone individuals could be driven by self-critical thinking.

A clinical implication here is that it may be difficult to modify self-punitive behaviours. Highly habituated or reinforced behaviours—particularly those that are underpinned by personal norms—are influential yet can be difficult to change (Aarts & Dijksterhuis, 2000; Hiler, 2015; Mallett, Bachrach, & Turrisi, 2009). Thus, clinicians working with individuals who self-punish in destructive ways may benefit from exploring and breaking down individuals' perceptions of self-punishment as a "normal" or functional response.

Interestingly, some research indicates that self-critical people are driven by two distinct motivations: wanting to self-improve, and wanting to harm the self for failures (P.

Gilbert, Clarke, Hempel, Miles, & Irons, 2004), mirroring the current findings indicating that self-punishment may repair moral identity while at other times it may be a more ruminative act of self-blame. Future self-punishment research may benefit from exploring notions of self-critical perfectionism in the moral domain, and testing whether such a trait is associated with self-punishment.

Future Directions

A question arising from the findings is whether self-punishment has one function or several. It appears contradictory that self-punishment can be simultaneously an exploration of one's problems and at the same time an attempt to suppress or avoid them. Does self-punishment have more than one function, occurring under different circumstances? In answering this question, research should explore moderators, that is, the conditions that may give rise to one effect over the other.

As we have speculated, levels of threat (e.g., transgression severity) might moderate effects: Lower threat may lead to emotion regulation while higher threat may lead to reflection and learning. Of course, the criticism remains that researchers are limited by ethical constraints in eliciting guilt; therefore, manipulating threat to a sufficiently high level to detect between-groups differences may be difficult. However, within current designs, threat could be measured and examined as a moderating variable, testing whether effects of self-punishment differ depending on perceived transgression severity.

Evidently, some of the methodological challenges we have noted here are more difficult to surmount. Experimenters will inevitably impose self-punishment manipulations to some degree; consequently they face the difficult task of creating the illusion of a self-inflicted experience as much as possible. They should therefore emphasise the voluntary nature of any tasks, as well as increasing participant engagement with the task (e.g., including multiple trials or opportunities for self-punishment). Moreover, care must be taken to

minimise the presence of confounds as much as possible, matching control conditions on extraneous variables such as benefits to the victim or others. Additionally, researchers should continue devising and testing a range of self-punishment tasks so that results can be generalised across types. For example, the current research indicated that spontaneous psychological self-punishment is particularly prevalent, though the experimental literature has largely eschewed this form.

Finally, it should be noted that although participants mostly ignored interpersonal functions of self-punishment, these might yet be implied down the track. According to a moral regulation model, moral emotions and appraisals of moral self-worth help to monitor interpersonal relations, prompting action when damage has been done (Leary & Baumeister, 2000; Sachdeva et al., 2009; J. L. Tracy & Robins, 2004). The psychological processes underlying self-punishment may therefore complement any interpersonal functions. For example, reflecting and learning might motivate amend-making in the long run. While we intentionally chose to analyse the phenomenological experience of self-punishment, the current design is perhaps not well placed to test longer-term implications, as subjective self-reports of self-punishment may focus on the more immediate and salient aspects of the self-punishment (i.e., the intrapersonal effects).

In fact, the inability or reluctance of participants to identify interpersonal functions could be consistent with all three themes: Those seeking to regulate emotion through self-punishment are perhaps consumed by their own intrapsychic experience, rather than paying attention to interpersonal outcomes; those reporting automatic and involuntary self-punishment responses perhaps feel that any interpersonal outcomes are incidental and irrelevant to their experiences; and those reflecting and learning about their wrongdoing are focused on introspection and perhaps making things right by their own measure rather than by others'. This does not mean that self-punishment does not have any interpersonal

outcomes, but highlights that from the self-punisher's perspective, these are not the most relevant parts of their experience.

Moreover, the current study was limited by the inability to follow up responses and probe into interpersonal functions, as in a traditional qualitative interview format. For instance, one could further question individuals who stated that nobody knew about their self-punishment about what they would *want* communicated by their act, or what it would mean to them if others knew about their self-punishment. Indeed, obtaining additional data through in-depth interviews could help to validate many aspects of the current findings; the themes identified in the present study provide many promising directions for other researchers, quantitative and qualitative alike, to follow up.

Limitations

The findings of this study should be considered in light of a number of limitations. First, findings are based on the perceptions and experiences of a specific set of individuals who were open to discussing their self-punishment. This group may not have included less forthcoming individuals who potentially engage in self-punishment for a different set of reasons. Moreover, online research studies can be influenced by the digital divide—the educational, economic, racial, and gender disparities between those who use and do not use the web (Rhodes et al., 2003). It should be noted that the current sample was primarily composed of Australian and New Zealand nationals, who may have specific cultural beliefs regarding self-punishment.

The current study was an exploration of self-punishers' conscious perceptions of their past experiences. Evidently, these recollections may not be entirely accurate, despite participants' best intentions to answer honestly. Moreover, if one is to consider the process of data collection as an interaction between the interviewer and interviewee (Potter & Hepburn, 2005), then the survey questions may have led participants to give undue weight to certain

interpretations of their experience. For example, asking how self-punishment changes the way one sees oneself prompts a consideration of how self-punishment is related to one's self-view, as well as suggesting a change in the self-view. Neither of these concepts may have been in fact relevant to participants' experiences at the time of their self-punishment. In fact, even the broadest question (*why do you think you chose to punish yourself?*)—asked to participants first, in order to minimise leading questions—implies (and leads participants to search for) a particular reason for their behaviour. It should be kept in mind that the current findings may not generalise to other observation contexts.

Similarly, participants' responses may have been influenced by a social desirability bias. In trying to explain their behaviour, participants may have been more likely to highlight functions they considered more pro-social, such as self-punishment as an opportunity for reflection and learning. Equally, participants may have been hesitant to point out interpersonal functions of self-punishment, lest they reveal (or admit to, in their mind's eye) a calculated and inauthentic dimension to their behaviour. Though we hoped the anonymity of the online format would limit a desirability bias, it is also possible that the format afforded participants a better chance to revise their responses in a way that was desirable to *their own* self-concept. No overt study of human behaviour, however, is completely immune from this charge, since self-revision may be unconscious, spontaneous, and shaped by the context in which one finds oneself.

In line with this thinking, it may be interesting to approach the topic from a different perspective, exploring narratives about self-punishment on a different level that steps away from the subtle realist perspective adopted here. For example, in line with more active forms of constructionism, one could see participants as narrators that are trying to make sense of their behaviour and self-concept through the act of telling their stories (Holstein & Gubrium, 1995). It may be of interest to qualitative researchers to expand our understanding of self-

punishment by investigating the role of storytelling in (for example) peoples' moral redemption narratives in self-punishment. Of course, this type of analysis would require more in-depth data than those in the current study.

Moreover, we acknowledge that qualitative methods are subject to researcher bias, and consider the findings in the context of criteria for good qualitative research (S. J. Tracy, 2010; Yardley, 2000). For example, we provided a transparent account of our expectations and understanding of self-punishment effects in the context of past research and its limitations. Though we do not deny that our theoretical background guided our research aims and analysis, given that the findings provided novel and contradictory perspectives relative to previous research—expanding our understanding of self-punishment—they are unlikely to be a mere product of our own expectations. Furthermore, the current research gives primacy to self-punishers' own conscious accounts of their behaviour, in an area where these voices may not yet have been heard. In any case, future studies could explore the consistency of the current findings by triangulating multiple research approaches and methods.

Conclusion

Despite its limitations, this study has provided one of the first accounts of spontaneous self-punishment outside the laboratory. Through a qualitative approach, we have gained a richer understanding of what self-punishment looks like and what it might mean to those who engage in it naturalistically. The findings offer a starting point for investigating the themes identified in self-punishers' accounts, particularly those that have thus far been neglected in—or contradicted by—the psychological literature, such as the long-term effectiveness of self-punishment in regulating emotion, and the conflict between avoidance and approach-oriented roles of self-punishment. In following up the present insights with experimentation, researchers are urged to bear in mind methodological hurdles in conducting self-punishment research.

CHAPTER 3: Validating Experimental Self-Punishment Tasks

Given the themes discovered in the qualitative research (Chapter 2), it is clear that there are dimensions of self-punishment that have been previously overlooked. Before any of these propositions can be tested experimentally, however, the criteria for experimental self-punishment paradigms ought to be reviewed in light of the various problems in conducting self-punishment research. In this chapter, I first review the features of self-punishment tasks that researchers may wish to consider. I then pilot four self-punishment tasks and judge them against these criteria. Thus, the purpose of this chapter is to establish a foundation and justification for the experimental self-punishment research that follows in Chapter 4.

Necessary Criteria for Self-Punishment Tasks

I propose that the two necessary and defining criteria of a self-punishment task are that it should be (1) self-inflicted and (2) aversive. Without these two elements, the task would neither be a punishment, nor self-inflicted (i.e., a "self-punishment").

Self-punishment manipulations in experimental designs are necessarily imposed to some degree by the experimenter; therefore, such tasks should attempt to create an illusion of a self-inflicted experience as much as possible. One way to simulate this experience is to maximise participants' engagement in the task. That is, the task should have sufficient length and/or variability so that participants feel that the *degree* of pain (or discomfort, or whatever constitutes the aversive nature of the task) they expose themselves to is self-inflicted, if not the experience as a whole. Previous paradigms have achieved this to varying degrees. The ice bucket task allows variability in length of time one's hand is held in the water (Bastian et al., 2011; van Bunderen & Bastian, 2014). Similarly, money or points can be deducted from oneself to varying degrees (Nelissen & Zeelenberg, 2009; Ohtsubo et al., 2014; Tanaka et al.,

2016; Tanaka et al., 2015; Watanabe & Ohtsubo, 2012; Zhu et al., 2017). Rating scales related to hypothetical or forecasted experiences (Nelissen, 2012; Nelissen & Zeelenberg, 2009; Slepian & Bastian, 2017) are not as immersive in the negative experience itself, but allow variability at the least.

Another way researchers could enhance participants' sense that they are genuinely choosing to punish themselves is to increase participant engagement with the task (e.g., by including multiple trials or opportunities for participants to punish themselves). Increasing the number of times participants choose to punish themselves (or not) may lead them to feel that they have more control over the task as a whole. Last, researchers could emphasise the voluntary nature of the task to participants.

Meeting the second fundamental criterion of self-punishment—an aversive experience—is a more complicated matter. Ethical considerations constrain the amount of pain that can be inflicted on participants. In an effort to work within these boundaries, researchers have creatively devised tasks that do not cause too much discomfort. Such tasks may involve non-noxious physical punishments (Bastian et al., 2011; Inbar et al., 2013; van Bunderen & Bastian, 2014), intended self-punishment for anticipated experiences that do not eventuate (Nelissen, 2012; Slepian & Bastian, 2017; Wallace & Sadalla, 1966), hypothetical experiences that can vary in aversiveness (Nelissen & Zeelenberg, 2009; Slepian & Bastian, 2017), and denial of points or money in lab-based games (Nelissen & Zeelenberg, 2009; Ohtsubo et al., 2014; Tanaka et al., 2016; Tanaka et al., 2015; Watanabe & Ohtsubo, 2012; Zhu et al., 2017).

In devising these subtle punishments, however, confounds have been inadvertently introduced that undermine the validity of the measures. For example, money deducted from oneself in economic games is implicitly given back to the

experimenter, or in the case of points that are used for lotteries, "self-punishment" results in a benefit for the other "players" in the game (sometimes the ostensible transgression victim). Consequently, the self-punishment tasks in these paradigms may be interpreted as altruism or amend-making rather than punishment. Therefore, self-punishment tasks should only have negative consequences for the transgressor (i.e., the participant), no gain for others, and be devoid of any confounds that may result in alternate interpretations of the task.

Can Self-Punishment Be Used as a Measured Variable?

As well as judging self-punishment tasks against the above criteria, researchers might also wish to use scores on self-punishment tasks as measures of self-punishment intensity. For example, a researcher might wish to be able to claim that a participant who holds his or her hand in an ice bucket for 30 seconds has self-punished more than another participant who only did so for 10 seconds. However, this relies on the assumption that scores are indeed meaningful measures of self-punishment intensity (i.e., that these scores vary as a function of objective self-punishment intensity). This assumption is yet to be confirmed or assessed.

As a preliminary test of whether scores on self-punishment tasks can be used as a measured variable, researchers could assess whether *degree of wrongdoing* is reliably related to the intensity with which one punishes oneself. If we begin from the premise that self-punishment is a response to wrongdoing—as the notion of self-punishment implies a perceived sense of deservingness, that one has done something to warrant it—then theoretically it should be more likely to occur as a result of a transgression. In terms of experimental design, this requires that participants in experimental conditions in which they are induced to transgress should score higher on self-punishment measures, relative to those who are not induced to transgress (a

test at the between-subject level of analysis). Similarly, it should follow that measured degree of wrongdoing (e.g., moral emotions as guilt or shame) is associated with more intense self-punishment (i.e., a test at the within-subject level of analysis).

Past studies provide mixed findings in relation to the claim that scores on objective self-punishment measures map onto degree of wrongdoing. Nelissen and Zeelenberg (2009) found that transgressors self-punished more than participants who did not transgress in Study 2, yet in Study 1 this was not so. In a later study, Nelissen (2012) found that transgressors were more likely to self-punish than those who did not transgress when the victim was present (audience presence was manipulated as well as guilt in this study), though they self-punished at the same rate for the other two audience conditions. Other research found that guilty participants inflicted greater electric shocks on themselves than did sad or control (neutral) group participants, and that levels of guilt generally correlated with shock level, though these correlations were not statistically significant (Inbar et al., 2013). Transgressors in Zhu et al. (2017) deducted more points from themselves than those who had not transgressed, and measured guilt predicted the number of deduction points (Study 2 and Study 3). Bastian et al. (2011) found that transgressors held their hand in a bucket of iced water for a longer period of time than those who did not commit a transgression—while using the same task in a later study yielded no such difference (van Bunderen & Bastian, 2014). Thus, though there is some concordance, self-punishment scores are not consistently related to wrongdoing.

There may be a fundamental problem with using scores on self-punishment tasks to measure the degree of self-punishment inflicted. The conflicting findings detailed above may be indicative of individual variation in sensitivity to pain stimuli (Fagius & Wahren, 1981; Nielsen et al., 2009) and perhaps to aversive experiences

more generally. Thus, it might be problematic to regard such scores as reliable, objective measures of self-punishment intensity. Though there might be some underlying relationship between wrongdoing and self-punishment, individual differences might obscure scores on objective self-punishment measures. It is unclear whether some tasks are more resistant to this charge than others (e.g., perhaps this is more of a problem for physical stimuli). The validity of self-punishment scores as objective measures of self-punishment intensity warrants further investigation.

Do All Self-Punishment Tasks Have the Same Effects?

Another question resulting from the literature on self-punishment is how universal the existing findings are. Self-punishment may take on a multitude of forms, as evidenced in the qualitative data collected in the present research (Chapter 2), including physical pain, psychological pain, denial of pleasure, social isolation, self-sabotage and food restriction. Yet experimental studies have employed few of these forms, partly due to ethical constraints of the research process. In fact, only two types of self-punishment have been empirically trialled: physical pain (Bastian et al., 2011; Inbar et al., 2013; Nelissen, 2012; van Bunderen & Bastian, 2014) and denial of pleasure/money (Nelissen & Zeelenberg, 2009; Ohtsubo et al., 2014; Tanaka et al., 2015; Watanabe & Ohtsubo, 2012; Zhu et al., 2017). For this reason, the present study will explore more diverse forms of self-punishment and test whether they are homogenous in their effects on variables such as guilt.

According to Chapter 2, psychological self-punishment may be one of the most pervasive forms of self-punishment, yet it has not been examined experimentally. Thus, it appears a particularly interesting form to further explore. However, this form of self-punishment carries an additional risk regarding construct validity and the subsequent interpretation of results. Specifically, there is a possible

confound between measures of psychological self-punishment and outcome measures of psychological wellbeing due to overlap between constructs as well as common method variance (Campbell & Fiske, 1959; Podsakoff & Organ, 1986). When using psychological self-punishment as a predictor variable (e.g., self-condemnation) and a negative psychological state as a dependent variable (e.g., guilt, self-esteem), it becomes difficult to reliably separate the dependent variable from the predictor variable, particularly if they are measured using similar methods.

For example, in one study (Fisher & Exline, 2006), self-condemnation (the predictor variable) is described as a self-directed negative emotional state, comprising of items such as "I feel like a bad person", "I feel angry at myself", and "I feel hateful toward myself", which closely resemble the outcome measures of self-esteem, emotional stability, depression, anger, and anxiety (all rating scales). Rather than demonstrating effects of self-condemnation, the psychological wellbeing items could reflect further self-condemnation; there could be insufficient discriminant validity between these measures. This might account for the finding that self-punishment was associated with *more* psychological distress (Fisher & Exline, 2006), which runs counter to experimental research using physical self-punishment (Bastian et al., 2011).

Efforts should be made to overcome this issue, for instance by obtaining independent and dependent variables from different (i.e., procedurally independent) sources (Podsakoff & Organ, 1986). To illustrate, a self-deprecating message that one chooses to send to oneself (or not) as a psychological self-punishment might be later followed by a self-report measure of guilt—ensuring sufficient differentiation between the predictor and outcome variables. However, though this might go some way to mitigate this problem, it will not entirely eliminate it.

Study 3.1

I piloted four different self-punishment tasks that each tapped into a different type of aversive experience: *cold stress*, *auditory pain*, *cognitive exhaustion*, and *self-criticism*. Thus expressions of both physical and psychological pain were captured. Each task was designed with the proposed criteria of self-punishment tasks in mind: allowing a self-determined experience of pain; and minimising the role of potential confounds that undermine the aversive experience (i.e., avoiding alternate attributions about the task, such as a benefit to the experimenter, demonstration of strength to others, etc.).

The primary aim of Study 3.1 was to assess these four tasks on the necessary criteria of self-punishment tasks, that is:

- Pain that is as self-inflicted as possible, by ensuring engagement with the task and variability within the measure (Analysis 1); and
- A purely aversive experience, free of confounds (Analysis 2).

The study also explored the proposition that degree of wrongdoing is related to increased use of the self-punishment task—constituting a test of whether scores on such self-punishment tasks can be reliably used as continuous measures of objective self-punishment intensity (Analysis 3). Furthermore, the tasks were compared on a small number of outcome variables in order to examine whether different self-punishment types have are different in their effects (Analysis 4).

Method

Sample and Design

The sample size was determined largely by the primary aim of the study (Analyses 1 and 2)—which amounted to within-cell descriptive statistics (i.e., of self-punishment score variability, and mean scores on rated aversiveness). Approximately

50 participants per self-punishment type was judged as sufficient to provide a reasonable indication of how the self-punishment tasks fared on these characteristics. In addition, this sample size granted adequate statistical power to conduct exploratory analyses of the secondary aims (i.e., Analyses 3 and 4). A power analysis (using G*Power 3; Faul, Erdfelder, Lang, & Buchner, 2007) indicated that 25 participants per cell would achieve 74% power at an alpha level of .05 to detect a medium effect for the difference between transgression and no transgression groups on self-punishment intensity measures (Analysis 3)—noting that in prior studies finding such differences, effect sizes were large.[1] Similarly, the sample would yield 79% power to detect a medium overall effect of self-punishment type on outcome measures (Analysis 4).

Two hundred and one undergraduate students (mean age 23.89, *SD* = 8.55, 75% female, 25% male) from Flinders University participated in exchange for course credit. The study was completed in person in the laboratory on campus, where participants were seated in private cubicles. Participants were randomly allocated to one of eight conditions, in a 2 (recall: no transgression vs. transgression) x 4 (self-punishment type: cold stress vs. auditory pain vs. cognitive exhaustion vs. self-criticism) factor between-subject design.

Procedure and Materials

Transgression recall. After answering some demographic variables, participants were asked to write in detail about a time when they "did something to hurt someone else" (transgression condition), or about their purchases when they last went grocery shopping (no transgression condition).

[1] $\eta^2 = .23$ (Study 2, Zhu et al., 2017); $d = 2.89$ (Bastian et al., 2011).

Pre-test guilt and shame. Prior to punishing themselves, participants reported their feelings of guilt and shame, which were embedded within the Positive And Negative Affect Schedule (PANAS; D. Watson, Clark, & Tellegen, 1988). Responses were on 7-point rating scales (scale anchors "Strongly disagree" and "Strongly agree").

Self-punishment. Following the transgression manipulation and emotion ratings, participants completed one of four self-punishment tasks: cold stress, auditory pain, cognitive exhaustion, or self-criticism. In order to avoid demand effects, the self-punishment purpose of the task was not made explicit to participants.

Cold stress (n=50). The ice bucket task was selected because it has already been used in previous self-punishment research as a traditional measure of physical pain (Bastian et al., 2011; van Bunderen & Bastian, 2014). This would also permit a comparison to previous findings. Participants were presented with a bucket of iced water (0°–2°C) and instructed to "hold your non-dominant hand, up to your wrist, in this water for as long as you can *or want to*. *You are free to take your hand out whenever you've decided you've had enough; there is no minimum time required*." The variation to the original instruction (additional words in italics) was designed to emphasise the voluntary nature of the task. A message to this effect was given to all participants across the four self-punishment tasks. The experimenter used a stopwatch to measure how long participants held their hand in the water, up to a maximum of two minutes.

Auditory pain (n=52). In the auditory paradigm, participants exposed themselves to a series of non-noxious short bursts of white noise (used as an aversive stimulis in aggression research, e.g., Anderson & Murphy, 2003; Thomaes, Bushman, Stegge, & Olthof, 2008). The task had two stages: "calibration" and "exposure."

Participants were first instructed to "calibrate" a sound measure by adjusting the volume of a series of noise bursts over seven trials. An initial test noise burst—the lowest volume burst—was fixed at a volume that was clearly detectable but not too unpleasant (approximately 60 dB HL), in order to avoid a startle response. Each subsequent trial allowed the participant to increase or decrease the volume by a single level (there were a total of seven volume levels); they could also leave the intensity unchanged. In the second stage (exposure), participants were asked to choose which volume level from the calibration stage they would like for "exposure" over a number of trials. After each burst, participants elected whether they would like to repeat the same burst, up to a maximum of four bursts. The auditory task was designed in such a way to increase participants' perceptions of the self-inflicted nature of the punishment. Their interaction with the noise volumes across different stages of the task aimed to enhance their sense of control over the experience, and thus the degree to which it was considered self-inflicted rather than imposed by the experimenter.

Noise bursts were 500 millisecond Waveform Audio File Format clips of broadband noise created and manipulated to different volume levels using tone generator software. Noise from the clips was measured by comparison to a calibrated audiometer (Amplivox 270) and confirmed with a sound level meter (IEC 651 type 2 standard) in dB(A). The maximum possible noise burst intensity was well within safe levels for ear damage, comparable to those used in previous research (approximately 100 dB HL).

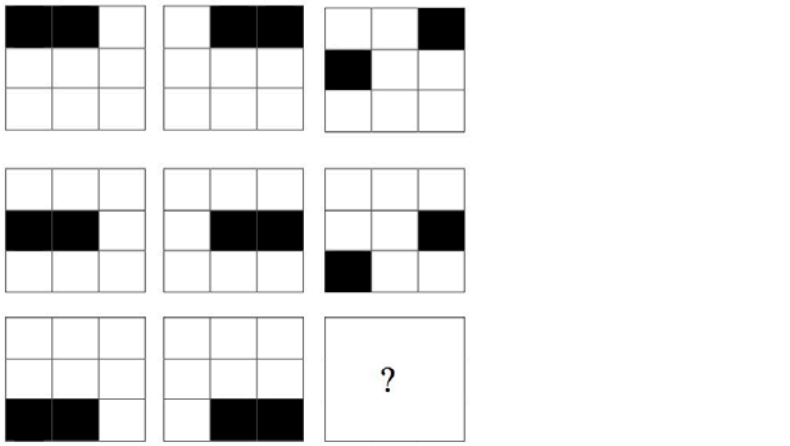
Cognitive exhaustion (n=50). Participants in the cognitive exhaustion condition were asked to re-type lines of text on the computer. Each line presented the same words but varied in letter capitalisation in order to make the task cognitively demanding, for example, "WritinG liNes caN be bOrIng, diffiCult, and anNoyiNg."

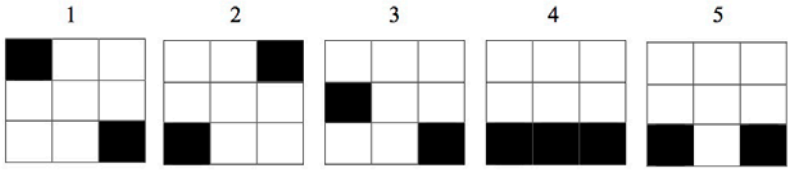
Lines were presented on a series of pages. The number of lines displayed on each page was varied, and each page included an option to continue or opt out of the task at the bottom, such that it was not obvious how many total lines the task involved. This was done to give the illusion that there was potentially an infinite (or a large) number of lines, and rather than feel pressured to diligently complete the entire task, participants would feel free to exercise their choice in how many lines to write.

Self-criticism (n=49). The fourth self-punishment condition was a form of psychological self-punishment, in which participants could reprimand themselves with negative feedback for their incorrect answers on an online test (modelled on Callan, Kay, & Dawtry, 2014, Study 7).

Participants were asked to solve a series of Raven's matrices-type puzzles by identifying the missing element in a pattern of shapes. An easy example item was first presented to ensure participants understood the nature of the task, followed by 10 test puzzles. Seven of the 10 test puzzles were impossible to solve, that is, they contained a logical sequence but the real solution to the puzzle was not included as an answer option. Thus, all responses to these items were incorrect and participants would have a chance to punish themselves in these cases. In order to reduce participants' suspicions that these puzzles were impossible, three of the 10 test puzzles included a valid response. A countdown timer limited each puzzle to 30 seconds (a failure to provide a response within this time was treated as an incorrect response). For each puzzle, when an incorrect response was selected, participants were given the option to "punish" themselves for their incorrect response with "negative feedback" or to skip the punishment. When participants opted to self-punish, the text "YOU'RE WRONG!" was presented on a page in large red font before proceeding to the next puzzle (see Figure 3.1).

(a)





(b) **YOU'RE WRONG!**

Figure 3.1. An example of a puzzle presented in the self-criticism task (a), along with the self-criticism issued on subsequent page (b). The correct answer to this puzzle is (1). In an unsolvable puzzle, this response option would not be present.

Post-test measures. Following their respective self-punishment tasks, participants rated how aversive they perceived it to be ("Please rate how unpleasant the task was", scale anchors "Not at all unpleasant" and "Very unpleasant"). At this stage they again completed measures of guilt and shame (e.g., "I feel guilty at the present moment"), as well as a broader sense of self-esteem ("I have high self-esteem"; Robins, Hendin, & Trzesniewski, 2001). Responses to these three measures were on 7-point rating scales, with scale anchors "Strongly disagree" and "Strongly agree" unless otherwise specified. Guilt and shame measures were employed because

self-punishment research has tended to focus on these two moral emotions; self-esteem was included as a more neutral identity-repair measure.

Results

Analysis 1: Assessing the Variability of Self-punishment Measures

Scores for all self-punishment intensity measures were examined. A high variability in the measure would suggest that participants were choosing how much they were punishing themselves, while low variability might indicate that the experiment itself was influencing their use of the task, possibly undermining how self-inflicted the punishment is or feels. Descriptive statistics for objective intensity measures and subjective unpleasantness ratings across the four self-punishment tasks are presented in Table 3.1.

Table 3.1

Descriptive Statistics for Objective Self-Punishment Intensity Measures

Measure	*n*	*M (SD)*	Kurtosis (*z*-score)
Cold stress	50		
Time in water (sec)		39.24 (32.59)	2.62
Auditory	52		
Calibration mean vol		2.38 (1.31)	-1.81
Exposure level vol		2.00 (1.80)	3.86
Exposure quantity		3.77 (0.70)	14.24
Cognitive exhaustion	50		
Number of lines		16.74 (7.97)	-2.21
Self-criticism	49		
Total reprimands		3.98 (3.39)	0.00
Reprimand rate		44.85 (27.34)	-2.24

Cold stress. The ice bucket task produces a single objective intensity measure: the amount of time participants held their hand in the water. The measure

showed good spread, ranging from 8 seconds to the maximum two minutes, with most participants (84%) removing their hand within one minute (see Figure 3.2). The kurtosis statistic indicated that the distribution was slightly leptokurtic, signifying that scores were somewhat clustered together around the mean.

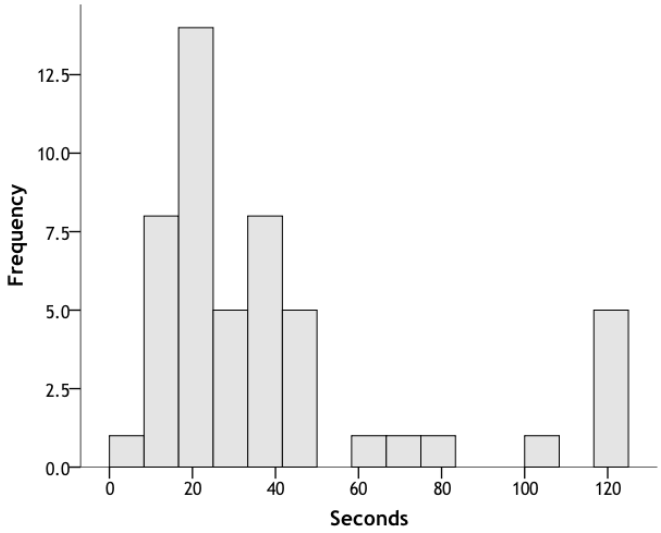


Figure 3.2. Histogram of cold stress self-punishment measure: time in iced water.

Auditory pain. For the auditory task there are three objective self-punishment intensity measures: the mean volume level across the seven trials in the calibration stage, the volume level of the burst selected for exposure, and the number of bursts selected for exposure. See Figure 3.3 for frequencies across these three measures.

There was good spread for mean volume level across the seven trials in the calibration stage, with scores observed on all possible values (mean volume level 1 to 4.86[2]) and a kurtosis score consistent with a relatively normal distribution. Volume level and number of bursts for exposure, however, did not show as much variability, reflected in significant positive kurtosis statistics: Most participants (67.3%) only exposed themselves to the lowest volume burst, and most (88.5%) tended to expose themselves to the maximum number of bursts (four).

[2] As participants could only increase the volume one level at a time, the maximum mean volume level across the seven trials was 4.86 ([2]+[3]+[4]+[5]+[6]+[7]+[7] / 7). Values correspond to volume level labels.

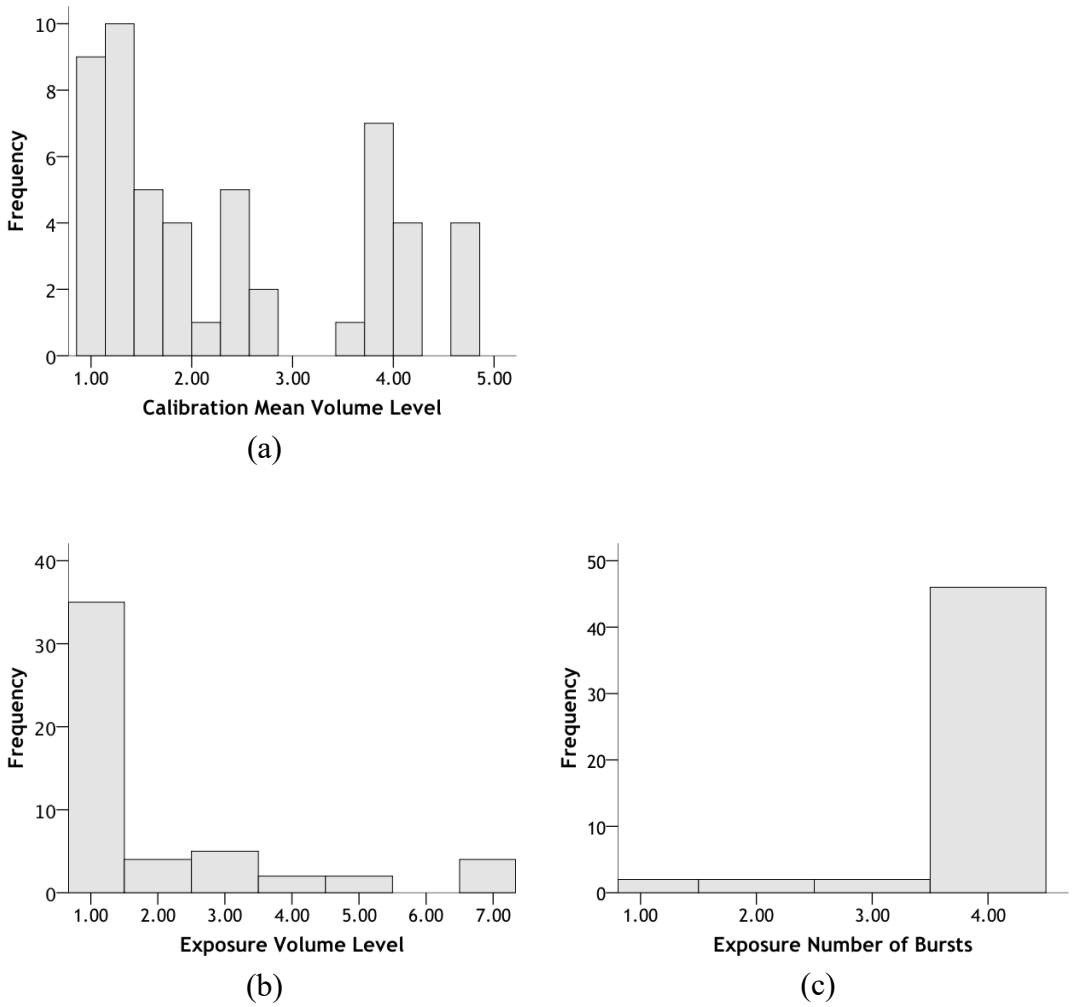


Figure 3.3. Histograms of auditory self-punishment measures: mean volume level in calibration stage (a), volume level selected for exposure (b), and number of bursts selected for exposure (c).

Cognitive exhaustion. The cognitive exhaustion task produces a single objective intensity measure: the number of lines written by the participant before opting out of the task. As can be seen in Figure 3.4, a substantial portion of participants (36%) wrote the maximum number of lines (25 lines)—suggesting a ceiling effect—though there was some spread across number of lines for the remaining participants. In fact, the distribution was slightly platykurtic, indicating a fairly flat distribution across all observed scores. Thus, the measure shows promise but there is scope to increase the number of lines presented to participants.

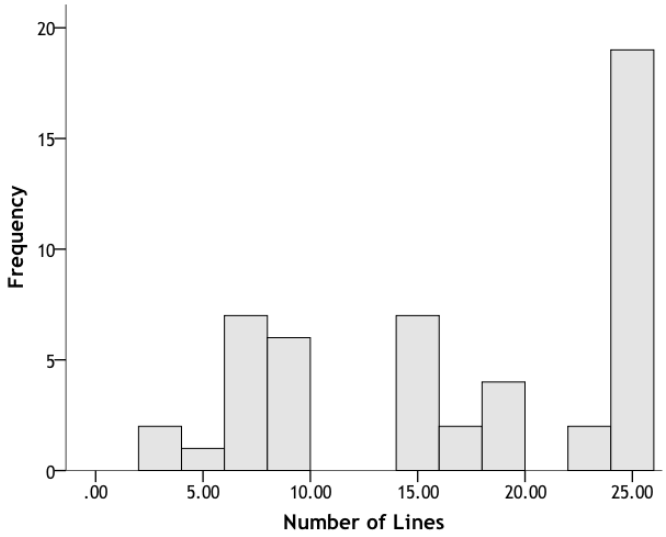
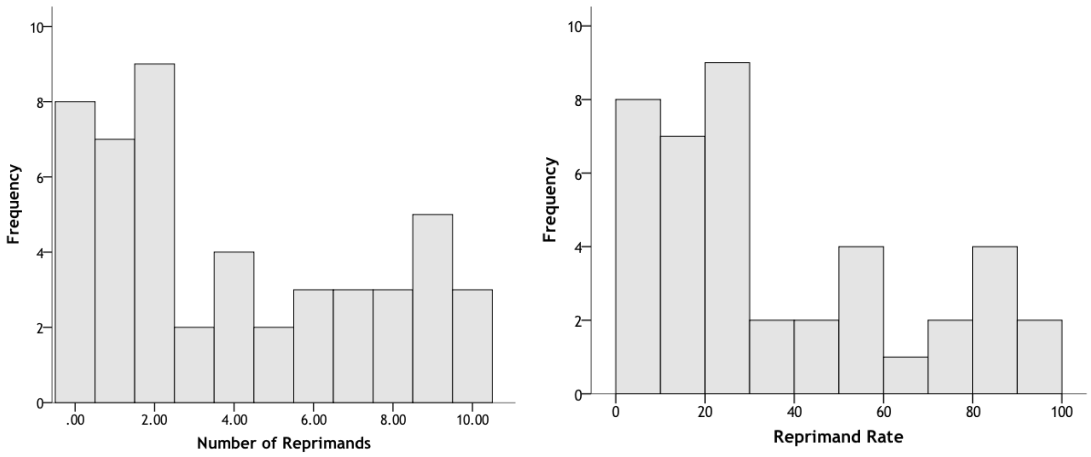


Figure 3.4. Histogram of cognitive exhaustion self-punishment measure: number of lines written.

Self-criticism. The self-criticism task produces a raw number of times participants chose to reprimand themselves. However, although every participant had the opportunity to punish on seven occasions (the number of impossible puzzles), others may have had more chances to do so if their responses to the genuine three puzzles were also incorrect. Thus, a self-punishment rate was calculated (number of reprimands ÷ number incorrect). This second measure is arguably a better gauge of whether participants felt compelled by the task to punish themselves, since it accounts for the number of opportunities they were given to do so. Both measures showed good spread (Figure 3.5), with scores across all possible values of the measures and a slightly negative kurtosis statistic for reprimand rate. Number of reprimands showed no kurtosis.



(a) (b)

Figure 3.5. Histograms of self-criticism measures: raw number of reprimands (a) and reprimand rate (b).

For Analysis 1, we can also consider an additional measure of self-punishment intensity: participants' *subjective* ratings of how unpleasant the task was. Given possible variability in sensitivity to *objective* measures of self-punishment, it may be useful to consider whether there was variance in self-punishment intensity from a subjective perspective. Table 3.2 shows subjective unpleasantness statistics for all conditions. Results indicated that all four tasks showed good spread in subjective self-punishment intensity approximating normal distributions. The auditory task and self-criticism tasks were more platykurtic than the cold stress and cognitive exhaustion tasks, suggesting slightly more even distributions across observed scores.

Table 3.2

Descriptive Statistics for Subjective Self-Punishment Unpleasantness Scores

Self-punishment task	*n*	*M (SD)*	Kurtosis (*z*-score)
Cold stress	50	5.06 (1.73)	0.59
Auditory	52	3.65 (1.67)	-1.92
Cognitive exhaustion	50	4.46 (1.53)	-0.65
Self-criticism	49	3.96 (1.78)	-1.49

Analysis 2: Assessing the Unpleasantness of Self-punishment Measures

Mean scores for task unpleasantness were subject to a one-way analysis of variance (ANOVA) with *Tukey* post-hoc tests (see also Table 3.2 for means and standard deviations). The cold stress task was significantly more unpleasant than the auditory and self-criticism tasks (both ps < .01), but not significantly more unpleasant than the cognitive exhaustion task (p > .05). The auditory, self-criticism, and cognitive exhaustion tasks did not differ significantly from one another (ps > .05). These findings suggest that the cold stress task is particularly promising on the criterion of unpleasantness. But there was no benchmark or comparison point from which to judge whether a task was unpleasant enough to constitute an aversive experience; it is not clear whether the other three tasks are unfit for use as self-punishment tasks. Nevertheless, it is encouraging that mean scores for all tasks tended to sit around the midpoint of the scale or above, suggesting that they were all considered unpleasant to some degree.

Unpleasantness of the tasks may also be assessed by examining whether those who punished themselves to a greater extent (defined objectively) also found the task more unpleasant (i.e., indicating a dosage effect of self-punishment). However, correlations between unpleasantness ratings and objective self-punishment intensity measures were weak, inconsistent, and not significant (see Table 3.3). This could suggest that the tasks were not aversive, though this assumes that objective self-punishment scores are reliable measures of self-punishment intensity—a claim that will be revisited in Analysis 3.

Table 3.3

Pearson Correlations Between Objective Self-Punishment Intensity and Subjective Unpleasantness

Measure	Unpleasantness
Cold stress (*n* = 50)	
Time in water	-.11
Auditory (*n* = 52)	
Calibration mean vol	.02
Exposure level vol	.25
Exposure quantity	.05
Cognitive exhaustion (*n* = 50)	
Number of lines	-.08
Self-criticism (*n* = 49)	
Total reprimands	.15
Reprimand rate	.10

Analysis 3: Self-Punishment as a Measured Variable: Does Degree of Wrongdoing Lead to More Intense Self-Punishment?

The proposition that higher degree of wrongdoing is related to self-punishment was examined next. At the between-subject level, this equates to a test of whether those who recalled a transgression self-punished to a greater extent than those who did not recall a transgression. First, to confirm that the manipulation elicited the sorts of feelings that would be associated with recalling transgressions, transgression and no transgression groups were compared on pre-test guilt and shame scores (i.e., before they were allocated to a self-punishment task). Results confirmed that the manipulation was successful, with those in the transgression condition feeling guiltier ($M = 4.76$, $SD = 1.83$) than those in the no transgression condition ($M = 1.89$, $SD = 1.32$); they also felt more ashamed ($M = 4.54$, $SD = 1.19$) than those in the no transgression condition ($M = 1.76$, $SD = 1.19$).

Descriptive and test statistics for the differences between control and transgression conditions on objective self-punishment measures are displayed in Table 3.4. Nonparametric tests were used for these analyses as cell sizes were small and most of the self-punishment measures violated the assumption of normality (i.e., Shapiro-Wilk normality test $p > .05$, along with skewness statistics > 2; Kim, 2013). There were no significant differences between the transgression conditions on any objective intensity measure. In fact, for the cold stress task and the three auditory task measures, means were in the opposite direction than would be expected; that is, those who had transgressed tended to punish themselves less than those who had not transgressed. Thus, there is little evidence that there is a relationship between committing a transgression and more (objective) self-punishment at the between-subject level.

Table 3.4

Differences Between Control and Transgression Conditions on Self-Punishment Intensity

	M (*SD*)		Mann-Whitney test	
Measure	No transgression	Transgression	*U*	Sig.
Cold stress				
Time in water	42.12 (36.06)	36.36 (29.17)	295.50	.74
Auditory				
Calibration mean vol	2.72 (1.30)	2.06 (1.26)	238.50	.07
Exposure level vol	2.28 (1.90)	1.74 (1.70)	268.00	.13
Exposure quantity	3.88 (0.44)	3.67 (0.88)	312.50	.41
Cognitive exhaustion				
Number of lines	16.42 (8.00)	17.04 (8.08)	292.00	.69
Self-criticism				
Total reprimands	3.79 (3.12)	4.16 (3.68)	285.00	.76
Reprimand rate	43.46 (35.89)	46.19 (39.37)	288.50	.82

To investigate the claim at the within-subject level, correlations between pre-test emotions (measured after the transgression recall but before self-punishment) and each of the objective self-punishment intensity measures were examined (see Table 3.5). Only two significant correlations emerged: Shame was associated with *less* self-criticism. Remaining correlations were weak and inconsistent; the measures did not vary in a meaningful way with guilt and shame. Thus, within-subject analyses largely accorded with the between-subject analyses, demonstrating that those experiencing more guilt or shame for one's transgression did not reliably self-punish—according to objective intensity—to a greater extent than those experiencing less guilt or shame.

Table 3.5

Pearson Correlations Between Emotions and Objective Self-Punishment Intensity

Measure	Guilt	Shame
Cold stress (n = 25)		
Time in water	.03	.11
Auditory (n = 27)		
Calibration mean vol	.11	.22
Exposure level vol	.05	.06
Exposure quantity	-.15	-.08
Cognitive exhaustion (n = 26)		
Number of lines	-.04	.05
Self-criticism (n = 25)		
Total reprimands	-.21	-.50*
Reprimand rate	-.18	-.45*
Pre-test variables (n = 103)		
Guilt	1	
Shame	.78**	1

Note. Correlations presented are for those in the transgression condition.[3]

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

[3] Guilt and shame measures were collected for the entire sample, but following Inbar et al. (2013) the transgression condition is the most relevant to test this hypothesis. Using the entire sample, none of the correlations between guilt/shame and the self-punishment measures are significant.

Neither transgression salience nor measured guilt and shame were reliably associated with self-punishment intensity. Based on these results, one could argue that objective self-punishment scores are not reliable measures of self-punishment intensity. However, if there is substantial variability in sensitivity to objective measures of self-punishment, it may be more revealing whether participants subjected themselves to more unpleasant self-punishment in response to guilt and shame, where unpleasantness is defined subjectively. Indeed, those feeling more guilt and shame as a result of their transgression tended to perceive their punishment as more unpleasant, particularly for the auditory pain and self-criticism tasks—though these correlations were not statistically significant (see Table 3.6). The exception to this was the cold stress task, for which the correlations were negative (also not significant).

Table 3.6

Pearson Correlations Between Emotions and Subjective Self-Punishment Unpleasantness

Self-punishment task unpleasantness	Guilt	Shame
Cold stress (*n* = 25)	-.25	-.29
Auditory (*n* = 27)	.22	.34
Cognitive exhaustion (*n* = 26)	.17	.23
Self-criticism (*n* = 25)	.32	.21

Note. Correlations presented are for those in the transgression condition.[4]

Analysis 4: Testing the Homogeneity of Self-Punishment Tasks

A final question resulting from the literature on self-punishment is whether different self-punishment tasks have the same effects. The current study did not have a control condition of no self-punishment from which to compare absolute effects of the various tasks. Nonetheless, as a first step this study would be able to compare the

[4] Participants in the no transgression condition were not induced to feel guilt or shame (i.e., there was no wrongdoing); as a result, their scores are not relevant and might introduce noise. Running the correlations with the entire sample reduces the size (but not direction) of the correlations.

effects of each task relative to one another. Post-test scores of guilt, shame, and self-esteem were subject to one-way ANOVAs with *Tukey* post-hoc tests, with self-punishment condition as the between-subject factor. The groups did not significantly differ on any outcome measure (see Table 3.7).

It is worth noting, however, that the cell sizes were small, and thus it is possible that there was not enough statistical power to detect the (generally) small effects for the individual comparisons.[5] There was some consistency in the pattern of means across the three variables whereby those in the cognitive exhaustion task had higher guilt and shame scores, and lower self-esteem scores, than those completing the other tasks; the largest effect sizes are indeed between the cognitive task and the other three tasks (see Figure 3.6). However, without a baseline condition of "no self-punishment" the nature of any potential differences are ambiguous.

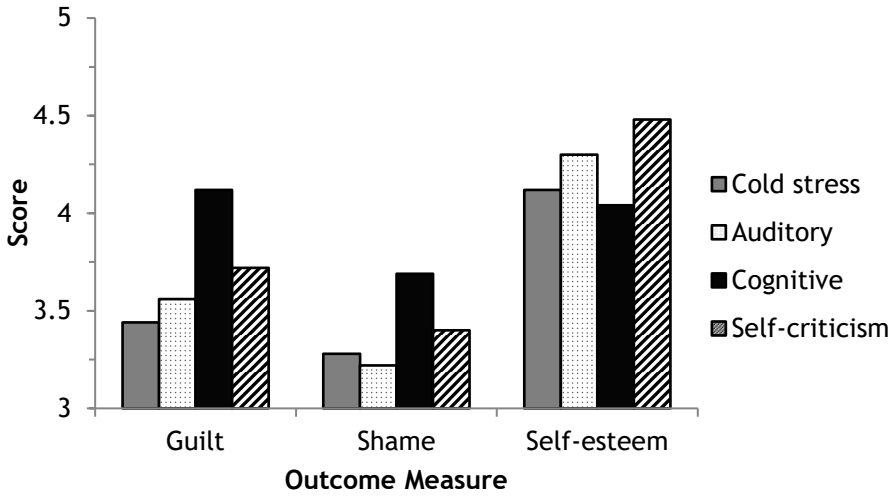


Figure 3.6. Differences between self-punishment tasks on dependent variables

[5] A power calculation using G*Power 3 (Faul et al., 2007) indicated that there was less than 20% statistical power to detect *d*s between 0.10-0.30 (within the range of most of the comparisons).

Table 3.7

Differences Between Self-Punishment Tasks on Dependent Variables

	M (*SD*)				ANOVA *F*-test	
Dependent variable	Cold stress (*n* = 25)	Auditory (*n*= 27)	Cognitive (*n* = 26)	Self-criticism (*n* = 25)	*F*	*p*
Guilt	3.44 (1.67)	3.56 (1.60)	4.12 (1.77)	3.72 (2.07)	*F*(3,99) = .71	.55
Shame	3.28 (1.82)	3.22 (1.85)	3.69 (1.83)	3.40 (2.00)	*F*(3,99) = .33	.81
Self-esteem	4.12 (1.64)	4.30 (1.27)	4.04 (1.73)	4.48 (1.64)	*F*(3,99) = .40	.76

	Comparison effect size (*d*)					
Dependent variable	Cold vs. Aud	Cold vs. Cog	Cold vs. SC	Aud vs. Cog	Aud vs. SC	Cog vs. SC
Guilt	0.07	0.40	0.15	0.33	0.09	-0.21
Shame	-0.03	0.22	0.06	0.26	0.09	-0.15
Self-esteem	0.12	-0.05	0.22	-0.17	0.12	0.26
Mean	0.07	0.37	0.14	0.25	0.10	0.21

To corroborate the above findings, pre- and post-test guilt and shame scores were each subjected to a mixed ANOVA with self-punishment condition as a between-subjects factor and guilt (or shame) scores as the within-subjects factor (two time points). For guilt, there was no significant interaction effect, indicating that the changes in guilt between the two measurement times were equal across self-punishment tasks, $F(3,99) = 0.10$, $p = .96$, partial $\eta^2 = .003$. Similarly, changes in shame did not depend on the self-punishment task, $F(3,99) = 0.66$, $p = .58$, partial $\eta^2 = .02$. Self-punishment reduced participants' guilt and shame across all four self-punishment tasks (see Table 3.8). It is also worth noting that those in the cognitive exhaustion task had slightly elevated guilt and shame scores prior to the self-punishment manipulation, which may account for the pattern of means seen for post-self-punishment guilt and shame. However, one-way ANOVAs indicated that any pre-test differences across conditions on guilt and shame scores were not significant, $F_{guilt}(3,99) = 0.56$, $p = .64$; $F_{shame}(3,99) = 0.85$, $p = .47$.

Table 3.8

Pre- and Post-Self-Punishment Guilt and Shame Scores

	Guilt			Shame		
	Pre	Post	*p*	Pre	Post	*p*
Condition	*M (SD)*	*M (SD)*		*M (SD)*	*M (SD)*	
Cold stress (*n* = 25)	4.56	3.44	< .001	4.44	3.28	< .001
	(1.85)	(1.66)		(1.87)	(1.82)	
Auditory (*n* = 27)	4.70	4.56	.003	3.56	3.22	< .001
	(1.64)	(1.74)		(1.60)	(1.85)	
Cognitive exhaustion (*n* = 26)	5.15	5.00	.02	4.12	3.69	.002
	(1.71)	(1.72)		(1.77)	(1.83)	
Self-criticism (*n* = 25)	4.60	4.16	.05	3.72	3.40	.03
	(2.16)	(2.29)		(2.07)	(2.00)	

Discussion

Suitability of the Four Tasks for Further Self-Punishment Research (Analyses 1 and 2)

Four different types of self-punishment were piloted in this study to determine their suitability for future self-punishment experiments. Two a priori criteria were proposed for self-punishment tasks: that it be (a) self-imposed and (b) aversive. Cold stress, auditory pain (particularly the calibration stage) and self-criticism showed good variability in their objective outcome measures. The latter two showed the most variability in terms of subjectively-defined self-punishment intensity, suggesting that participants exercised their choice in what extent to punish themselves—meeting the first criterion of the task being self-imposed. The cognitive exhaustion task, on the other hand, was less variable; participants tended to complete the task in its entirety. Though this does not necessarily mean that the task did not feel self-imposed, it might suggest that some participants did not utilise the task as much as they would have liked to, potentially undermining the potency of any subsequent effects. It may be beneficial to increase the duration of the cognitive exhaustion task (i.e., number of lines that participants can write) to further assess this measure.

All four tasks were rated as unpleasant to some degree, constituting a preliminary confirmation of the second necessary criterion for self-punishment tasks—that the task should be aversive. Cold stress stood out as more aversive than the other tasks.

Objective Self-Punishment Scores Might Not Be Reliable Measures of Self-Punishment Intensity (Analysis 3)

Despite most of the objective intensity measures showing good variability, there were some indications that these scores could be problematic as measures of self-punishment intensity. More intense self-punishment on these objective measures was not consistently correlated with subjective perceptions of task unpleasantness (i.e., subjective and objective perceptions of intensity did not match up). In addition, those who had transgressed were not

any more likely to self-punish relative to those who had not transgressed; in fact, means for two of the tasks (cold stress and auditory) showed the inverse pattern. The pattern observed for the cold stress task—whereby transgressors self-punished to a lesser extent than non-transgressors—is consistent with the results obtained by van Bunderen and Bastian (2014), but inconsistent with an earlier study using this task (Bastian et al., 2011). Moreover, for those who had transgressed, higher levels of felt guilt and shame did not consistently lead to more self-punishment. These results are inconsistent with a prior study that found that levels of guilt could significantly predict number of points deducted from the self (Zhu et al., 2017). On the other hand, the current findings are in line with a study that did not find significant correlations between guilt and electric shock level (Inbar et al., 2013). Overall, the current results indicating that degree of wrongdoing did not reliably vary with objective self-punishment intensity is consistent with the mixed results in research by Nelissen and colleagues, in which transgressors did not always punish themselves more than non-transgressors (Nelissen, 2012; Nelissen & Zeelenberg, 2009).

One explanation for the results (and indeed, for the inconsistency across the entire literature) is that group differences may be obscured by individual variation in sensitivity to pain stimuli (Fagius & Wahren, 1981; Nielsen et al., 2009). Interestingly, the results of the current study suggest that individuals may vary not only in their sensitivity to cold stress and auditory pain, but also in relation to more psychological forms of self-punishment (cognitive exhaustion and self-criticism). If this were the case, measures of objective self-punishment intensity would be generally uninformative; individual differences in sensitivity may diminish any between-group differences. It is possible, however, that such measures could be used within individuals (i.e., a person's self-inflicted punishment for a transgression at Time 1 could be compared to the same person's degree of self-punishment for no transgression at

Time 2)—though this assumes that individual sensitivities are relatively stable across time. This is worthy of further exploration.

Alternatively, one could use subjective measures of self-punishment unpleasantness as indicators of self-punishment intensity. In this way, though the objective measures of self-punishment may be unreliable between participants, the tasks may nevertheless function as a subjective experience of self-imposed punishment. That is, irrespective of varying sensitivities to pain, participants may seek to make the task unpleasant for themselves as a response to wrongdoing. It should be noted, however, that while correlations between guilt and shame and *subjectively*-defined self-punishment intensity were relatively consistent (particularly for auditory pain, cognitive exhaustion and self-criticism), these were not statistically significant. Further research with a larger sample could help determine whether there is any merit to this argument.

Possible Homogeneity of Self-Punishment Tasks (Analysis 4)

Results indicated that the self-punishment tasks, though arguably tapping into a diverse range of pain types, did not significantly differ in their effects on guilt, shame, and self-esteem. Self-punishment appeared to "reduce" guilt and shame across all four tasks. However, one cannot conclusively speak of these differences across time as "reductions" or "increases," since the pre- and post-test measures of guilt and shame differed in their wording. Consequently, the nature of the changes is unclear. Nevertheless, one could tentatively contend that the self-punishment types examined here are relatively homogenous in their effects on emotion and self-esteem. Of course, it is possible that other forms of self-punishment not employed in the current study have distinctive effects.

Limitations

The current study is limited by its sample size. Cell size was sacrificed in the interest of piloting a range of self-punishment tasks. This limitation, however, does not speak too

strongly to the assessment of the self-punishment meeting the two necessary criteria for self-punishment tasks (Analyses 1 and 2); 50 participants per cell is large enough to get a sense of how participants employed the measures (i.e., variability) and how they rated the measure (mean unpleasantness scores).

In contrast, sample size may have influenced the tests of between-group differences that were undertaken to address the exploratory Analyses 3 and 4, which relied on smaller cell sizes. For Analysis 3, there was no evidence that transgression and no transgression groups differed on the degree to which they subsequently punished themselves. However, this finding cannot be chalked up to sample size limitations. First, in studies that found significant differences on similar measures, the effect sizes were large (Bastian et al., 2011; Zhu et al., 2017). There was ample power in the present study to detect effects of such magnitudes. Second, and more importantly, the current results indicated mean differences in the opposite direction as to what was found in the studies cited above—particularly noteworthy in the case of Bastian et al. (2011), in which the authors used the same ice bucket paradigm as the present study.

Moreover, correlations at the within-subject level were inconsistent (i.e., correlations between emotions and objective self-punishment varied in direction across measures). In contrast, correlations between emotions and subjective self-punishment unpleasantness were more consistent in their direction and size, though these correlations were not statistically significant (note these were small correlations, likely beyond the statistical power in the study to detect). Thus, it seems unlikely that the conclusion—that transgressors are no more likely to self-punish on objective measures than non-transgressors—is driven by sample size limitations. Rather, it seems likely that individual differences in sensitivity to punishment stimuli diminish the predictive power of the objective self-punishment intensity measures.

With regard to Analysis 4, as already mentioned, the small sample size may have prevented the detection of meaningful differences between self-punishment tasks on outcome measures. Thus, self-punishment researchers may wish to test the generalisability of their findings by employing various self-punishment types.

Conclusion

The present study reviewed the criteria for, and features of, experimental self-punishment paradigms. Four types of self-punishment tasks were designed and piloted for use in quantitative experimentation, with results indicating various advantages and disadvantages of each task. Auditory pain and self-criticism showed the most variability on measures of intensity (defined both objectively and subjectively), suggesting that the aversive experiences in these tasks are inflicted voluntarily—meeting the criterion that self-punishment tasks be genuinely self-imposed. All tasks were considered aversive—meeting the second criterion—though cold stress stood out as the most aversive task. Of these three tasks, however, only those in the auditory pain and self-criticism groups appeared to punish themselves more (intensity as defined subjectively) when they felt more guilt and shame as a result of their wrongdoing.

While by these criteria, the auditory pain task and the self-criticism seem similarly promising, the latter is potentially problematic due to a possible confounding of psychological self-punishment and the negative psychological states that are measured as dependent variables following these tasks. Thus, though all four tasks warrant further research, it can tentatively be suggested that the auditory task appears particularly promising. However, though this task may be useful as a manipulation of self-punishment, objective intensity scores (e.g., mean volume of bursts) might not be reliable as quantitative measures of self-punishment intensity between individuals.

CHAPTER 4: Expression or Evasion of Guilt? Two Ways Self-Punishment Resolves the Threat to Moral Identity

Committing a wrongdoing poses a threat to one's *moral identity*. Individuals may feel that their sense of being a moral and good person, accepted by others as such, is under attack (Haidt, 2003; Leary, 2004; Shnabel & Nadler, 2008; J. L. Tracy & Robins, 2004). But while self-forgiveness comes easily to some, for others immoral acts are more difficult to move past. Plagued by guilt and self-reproach, some people inflict punishment on themselves following a perceived moral failure (Nelissen & Zeelenberg, 2009; Nyström & Mikkelsen, 2013; Tanaka et al., 2015). Whether of a physical nature (as exemplified by self-flagellation) or more subtle manifestations (e.g., social self-isolation or psychological self-condemnation), self-punishers actively seek retribution for their immoral act. That is, instead of avoiding castigation, self-punishers actively invite it. Given that as humans, we tend to act in our own defence, present ourselves in the best possible light (e.g., Alicke & Sedikides, 2009; Bandura, 1990), and avoid pain (Higgins, 1997), what could motivate this type of behaviour?

As discussed in Chapter 1 and reflected in self-punishers' accounts (Chapter 2), the idea that suffering can provide a sense of atonement for one's sins is one that holds a place in lay understandings of justice. The rhetoric of atonement and purification provides some indication as to the appeal of self-punishment: that these acts can symbolically restore one's moral identity. A closer examination of the connection between moral identity and self-punishment is warranted, given the central role that moral identity plays in moral behaviour more generally (Aquino & Reed, 2002; Blasi, 1993; Sachdeva et al., 2009). In line with this, researchers have found that self-punishment can reduce guilt (Bastian et al., 2011; Inbar et al., 2013). But how exactly does self-punishment resolve the threat to moral identity posed by one's wrongdoing?

The current research proposes that self-punishment may actually reflect not one, but two distinct processes that resolve the threat to moral identity: *moral cleansing* and *moral repair*. On one hand self-punishment may restore moral identity by excusing one's immoral behaviour (morally "cleansing" oneself of moral fault for the transgression). On the other hand, self-punishment might actively confront and repair one's moral identity by recommitting oneself to the values violated by the transgression (moral repair). As we will review below, though these two mechanisms both attempt to address the threat to moral identity, they are likely to have different emotional, cognitive, and interpersonal implications.

Individuals Attempt to Resolve Threats to Moral Identity Through Moral Cleansing or Moral Repair

Threats to identity may be resolved in many ways, the tools available to do so being remarkably diverse (Tesser, 2001). These strategies, however, differ in their mechanisms and outcomes. One way scholars have classified behavioural responses to guilt and shame is by defining them as either "defend" or "repair" oriented behaviours, in line with broader literature on motivational orientations (Gausel et al., 2012; Schmader & Lickel, 2006). We propose that expressions of self-punishment might represent either a defensive or reparative approach to resolving the threat to moral identity.

First, transgressors can address the threat to moral identity indirectly through defensive strategies that are aimed at escaping or hiding from the threat. For example, transgressors can disengage from their wrongdoing by minimising the harm done or shifting blame to others (Bandura, 1991, 1999; Sykes & Matza, 1957). These actions can be categorised as *moral cleansing*: behaviours that attempt to restore moral identity by ridding oneself of responsibility for the transgression in some way. Moral cleansing is presently defined as a self-enhancing process that restores moral identity by casting off self-examination. In this process, transgressors do not pause to consider why they have done

wrong, or how they can undo the harm. Rather, they seek to swiftly purge themselves of their guilt (and other aversive emotions and cognitions). Studies finding a reduction in guilt following self-punishment (Bastian et al., 2011; Inbar et al., 2013) are consistent with the proposition of self-punishment as moral cleansing.

Self-punishment might reduce guilt—and resolve the threat to moral identity—by affirming alternate sources of positive identity; that is, values or qualities that are not related to the transgression itself. For example, affirming one's toughness, agency, or that one is a righteous administrator of punishment (Rothschild et al., 2015). That is, the act of punishing (rather than being punished) might restore a self-punisher's identity. This is in line with research showing that punishing others can make one feel "morally just" (Adams, 2011). These types of affirmations are unrelated to the moral failure, or attempt to detract from the notion of the self as the perpetrator. Therefore, self-punishment as moral cleansing can be considered an evasion of guilt rather than an expression of it (Carveth, 2006), analogous to physical cleansing behaviours that purify by "wiping the slate clean" (S. W. Lee & Schwarz, 2011; Zhong & Liljenquist, 2006).

Yet there is some evidence that self-punishment does not always cleanse transgressors of their wrongdoing. Self-criticism and self-condemnation have been linked to increased psychological distress (Fisher & Exline, 2006; Whelton & Greenberg, 2005). These findings could suggest that self-punishment may be unsuccessful at cleansing away one's wrongdoing despite a belief that it will do so; that is, an affective forecasting failure as seen when individuals anticipate (yet do not gain) satisfaction from punishing others (K. M. Carlsmith, Wilson, & Gilbert, 2008). Alternatively, these findings could point to a different conceptualisation of self-punishment. Thus, we propose that self-punishment might also be understood as *moral repair*: a proactive, approach-oriented response through which transgressors orient themselves towards repair of the wrongdoing. Rather than sidestepping

the threat to their moral identity, perhaps self-punishers are instead confronting the source of the threat. They might want to assert that the immoral behaviour is not representative of their moral character; rather, they restore their moral identity through the act of calling out their misbehaviour as a wrongdoing and accepting the need to be punished for it.

Implicated in this proposed process of moral repair is a recommitment to the moral values violated (Woodyatt & Wenzel, 2014). By punishing oneself for the violation, self-punishers can remind themselves that the violated value is indeed a value of importance to their identity. Thus, self-punishers who are undergoing moral repair are engaging, rather than disengaging, from their wrongdoing. In doing so, self-punishment reinforces their identification with—and commitment to—the collective social order (Tanaka et al., 2015). Yet resolving the threat to moral identity in this manner is unlikely to be an easy process. Unpacking one's wrongdoing through value affirmation can induce further feelings of guilt and shame (Woodyatt & Wenzel, 2014). This is consistent with the finding that self-condemnation does not always reduce distress (Fisher & Exline, 2006; Whelton & Greenberg, 2005).

Moral cleansing and moral repair have different interpersonal implications. Following a transgression, there is often a chance to engage in reconciliatory actions (e.g., apologise or make amends). If moral cleansing provides an escape from one's wrongdoing, it would also dissolve the need for such actions. Releasing oneself of one's guilt undermines its motivational power for offenders to address others' needs (O'Keefe, 2000; Woodyatt & Wenzel, 2014; Zhong & Liljenquist, 2006). Indeed, defensive responses are so named because they motivate social withdrawal (Gausel et al., 2012). In line with this, simply making oneself feel better without emotional and cognitive processing of a wrongdoing is related to self-centredness, a lack of empathy, and a diminished interest in victim reconciliation (Tangney et al., 2005; Woodyatt & Wenzel, 2013a; Woodyatt et al., 2017).

On the other hand, if self-punishment regulates moral identity through a process of moral repair whereby transgressors recommit themselves to moral values, it is likely to promote reparation and reconciliation. Affirming violated values implies an acceptance of one's responsibility, of the hurt caused to others. These appraisals could lead to more feelings of guilt and shame. If self-punishment does not reduce guilt, then the motivational drive to amend the hurt caused by one's actions would also remain intact (Baumeister et al., 1994; J. M. Carlsmith & Gross, 1969; De Hooge et al., 2008; Gino et al., 2015; Tangney et al., 2007). Confirming this, research indicates that value affirmation maintains (or increases) the desire to engage in prosocial behaviour and reconcile with victims (Tetlock et al., 2000; Woodyatt & Wenzel, 2014; Woodyatt et al., 2017). Self-punishment might thus be a starting point for moral repair through which transgressors can strengthen their resolve to make things right.

A study by van Bunderen and Bastian (2014) found evidence for both interpersonal outcomes of self-punishment: defend and repair. In this study, self-punishment only replaced victim compensation for individuals high in victim justice sensitivity (a measure of how sensitive individuals are to identifying themselves as victims of injustice). Meanwhile, for those low in victim justice sensitivity, self-punishment did not replace compensation. In other words, to the extent that the transgressor was likely to engage in a defensive appraisal of the incident (i.e., seeing oneself as the victim), self-punishment re-balanced the scales of justice and reduced reparative intentions (consistent with moral cleansing). In contrast, those less likely to react defensively were not simply cleansed of their wrongdoing by punishing themselves. These individuals remained motivated to engage in interpersonal restitution following self-punishment—consistent with moral repair.

Acknowledgment of Moral Need Determines the Strategy Employed

What determines whether self-punishment cleanses or repairs? We propose that a key factor moderating these effects is the acknowledgment of a moral need. Empirical evidence

indicates that shame acknowledgement is a key process in value affirmation (Woodyatt & Wenzel, 2014). Thus, to the extent that transgressors acknowledge the violation and accept the need to engage in a critical self-examination, the self-punishment that emerges may be motivated by moral repair. Those who deny that need will be motivated to use self-punishment as a simple strategy to cleanse away their guilt. Indeed, van Bunderen and Bastian's (2014) results indicated that the tendency to see oneself as a victim (i.e., denying the moral need) was associated with reduced restitution following self-punishment.

Perpetrators may be more likely to acknowledge a moral need under particular conditions. The extent to which they do so is likely to depend on how important, obvious or pressing the need is, since such threats more relevant and *meaningful* to one's identity—and to one's social survival (J. L. Tracy & Robins, 2004). For instance, people might tend to seek moral repair when a transgression is more severe; in these cases a single action is not sufficient to resolve the threat to moral identity (Tetlock et al., 2000). A minor transgression can be swept under the rug, but a more blatant violation—a warning signal too loud to ignore—may not be so easily excused by affirming an unrelated quality. This argument is consistent with Tetlock and colleagues' observation that sacred values may be more difficult to trade off with other values (Tetlock et al., 2000).

In contrast, moral cleansing may be more successful when there are fewer expectations for action, or more importantly, ramifications for inaction (e.g., retaliation from others). Thus, from a functional perspective, it would make sense to choose the least costly path when the repercussions for inaction are low—one that restores moral identity but requires nothing more than a symbolic gesture of self-punishment. Such an act might be sufficient to transgressors and to others that justice has been done, effectively neutralising the threat and with it the need for further action. Providing some support for these predictions, Fisher and Exline (2006), whose results were consistent with moral repair, asked participants

to recall a “fairly serious” offence, whereas the two studies finding a reduction in guilt (consistent with moral cleansing) did not make such a specification (Bastian et al., 2011; Inbar et al., 2013). Participants in the latter studies may have recalled fairly benign transgressions, which may have motivated them to simply cleanse themselves of their wrongdoing through an act of self-punishment.

The Present Research

In summary, previous research points to two possible ways that self-punishment regulates one’s moral identity: one that seeks to minimise further engagement with one’s immoral act, and one that seeks to repair it. We propose that the two pathways are compatible, and the function of self-punishment is likely to be moderated by the extent to which an individual acknowledges a moral need. The broad aim of the following studies is to demonstrate this differentiated effect, not only in terms of emotional outcomes (e.g., guilt), but a more detailed analysis of moral cognition (e.g., moral engagement, empathy, commitment to values, responsibility taking) and downstream effects on reparative action. The hypothesis is as follows:

Self-punishment will have either a moral cleansing or moral repair function, moderated by the extent to which a person acknowledges a moral need (see Figure 4.1):

- Low moral need will result in self-punishment as moral cleansing: decreased commitment to values, moral engagement, acceptance of responsibility, guilt, empathy, and reparation.
- High moral need will result in self-punishment as moral repair: increased commitment to values, moral engagement, acceptance of responsibility, guilt, empathy, and reparation.

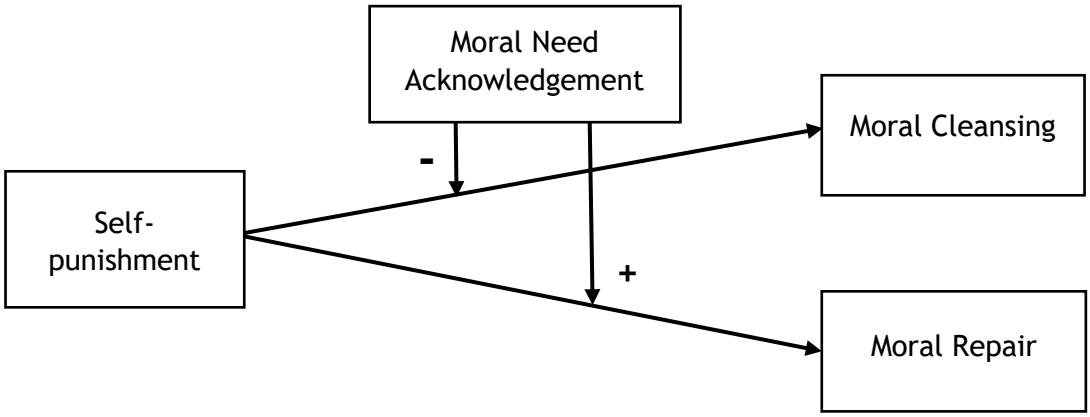


Figure 4.1. Schematic depiction of the proposed theoretical model.

We report four studies that explore the identity-regulating processes underlying self-punishment. In Studies 4.1 and 4.2, we test the proposed model across two different transgression contexts. In Study 4.3 we seek to generalise the findings to a different self-punishment task. Finally, in Study 4.4 we attempt to provide an experimental test of the causal role of moral need acknowledgment in determining self-punishment outcomes.

Study 4.1

Prior studies finding a reduction in guilt following self-punishment—supporting a moral cleansing function—have one important feature in common that limits the generalisability of their results: They utilise self-punishment tasks that could be easily construed as endurance tasks. Faced with the opportunity to punish oneself with iced water (Bastian et al., 2011) or electric shocks (Inbar et al., 2013), participants may be tempted to use the task to show off their physical toughness. These types of task are therefore well placed to affirm values or qualities unrelated to the transgression, providing a boost to one's ego. It remains unclear if all instances of self-punishment have the same effect. To eliminate this possible confound, we developed a novel auditory self-punishment paradigm: blasting oneself with loud noises. This task minimises an endurance dimension as there is perhaps less of an intuitive sense of toughness in withstanding an auditory pain stimulus (relative to

physical pain), balancing the potential to facilitate both moral cleansing and moral repair goals. Moreover, the privacy granted by this task limits the possibility of self-punishment being a communication to victims or third parties—allowing us to test the moral identity-regulating functions in isolation.

In order to further engage the identity processes potentially underlying self-punishment, we used a non-relational transgression. If identity is an internalised self-regulation mechanism, a victim need not be present or implicated in a transgression itself to activate this mechanism. A short video was used to induce a transgression, or rather, to make one's transgressing salient. The video was abridged from *Glass Walls,* a documentary produced by People for the Ethical Treatment of Animals. It depicts the mistreatment of animals in the meat industry, revealing cruelty to animals such as chickens and cows. As consumer choices contribute to the maintenance of this industry, this can be considered a moral transgression on the participant's part. This video has been used to elicit guilt in past research (Wenzel, Woodyatt, & McLean, 2017). Previously self-punishment has been studied strictly in response to relational transgressions (i.e., where there is an equal moral agent as a victim; nonhuman animals, by contrast, are objects of moral value but cannot respond in the same way as moral agents). Perhaps due to this past focus on relational transgressions, some researchers have highlighted interpersonal motivations such as victim restitution (Nelissen, 2012). A non-relational transgression may thus provide a different perspective on self-punishment.

Method

Sample. Though we had no reference point with which to determine the size of our interaction effect, G*Power 3 (Faul et al., 2007) suggested a sample size of 64 at 80% power to detect a small to medium effect for both directions of the simple slopes (0.3 and -0.3 regression coefficients), at an alpha level of .05. This sample size would also provide 99%

power to detect a main effect consistent with Bastian et al.'s (2011) reduction in guilt following self-punishment. This power calculation provides a guide for the minimum sample requirement for all studies in this chapter.

We recruited 72 first-year psychology students at Flinders University (68% female, 32% male, mean age 20.63 years (*SD* = 5.65), who participated in return for course credit. Vegetarians or those with sensitivity to loud noises were not eligible for participation in the study.

Procedure. The study was a between-subjects design with two cells (control vs. self-punishment). After completing some demographic variables, participants "transgressed" by watching the movie about mistreatment of animals in the meat industry. Participants then proceeded to the auditory stimulation task (i.e., self-punishment); a paradigm that exposed them to a series of non-noxious short bursts of noise (see research by Anderson, Bushman and colleagues, e.g., Anderson & Murphy, 2003; Thomaes et al., 2008).

Participants were randomly assigned to one of two conditions for the auditory task. All participants were first instructed to "calibrate" a sound measure by adjusting the volume of a noise burst as they pleased, over a series of seven trials. An initial test noise burst was fixed at a low intensity. Each subsequent trial allowed the participant to increase or decrease (or maintain) the noise level by a single volume level. In the self-punishment condition, participants were informed at the beginning of the task that they would be exposed to *loud* noises. In this condition, the initial test burst was the minimum volume participants could elect across all trials; they could only increase or maintain this volume. This level was clearly detectable but not too unpleasant (to avoid a startle response): approximately 60 dB HL. In the control condition, participants were told that they would be exposed to *quiet* noises. In this condition, the same initial test burst was the maximum volume participants could elect across all trials; from there they could only decrease or maintain this volume.

In both conditions, following their respective calibration stages participants moved onto an exposure stage. They were asked to choose which noise level from the calibration stage they would like for "exposure" over a number of trials. Noise level options differed between conditions to reflect the set of louder or quieter levels used in their relevant calibration stage. Last, participants elected how many times to repeat the chosen noise level for exposure, up to a maximum of four bursts.

Noise bursts were 500 millisecond Waveform Audio File Format clips of broadband noise created and manipulated to different volume levels using tone generator software. Noise from the clips was measured by comparison to a calibrated audiometer (Amplivox 270) and confirmed with a sound level meter (IEC 651 type 2 standard) in dB(A). The maximum possible noise burst intensity was well within safe levels for ear damage, comparable to those used in previous research (approximately 100 dB HL).

The task was designed in such a way to increase participants' perceptions of the self-inflicted nature of the punishment. That is, though a starting level was given, their interaction with the noise volumes across different variations of the task (i.e., the calibration stage and then the exposure stage) aimed to increase their sense of control over the experience, and thus the degree to which it was considered self-inflicted rather than imposed by the experimenter. Following the auditory stimulation task, all participants completed the outcome measures.

Outcome measures. Unless indicated, all outcome measure items reported in this series of studies were measured on a 7-point rating scale.

To tap into acknowledgement of a moral need, we measured how threatened participants perceived their moral identity was by the transgression. We appealed to moral identity threat at the person level ("I feel that I am a [moral]/[good] person"), the social level ("I feel that I am a good member of my community") and at a more universal level ("I feel

that I have acted in line with my moral values"). The scale values were reversed such that higher scores on this measure indicated higher moral identity threat (α = .69).

Guilt and shame were assessed using single item measures with the wording, "When I think about my food choices I feel [guilty]/[ashamed]." A single-item measure gauged participants' overall acceptance of responsibility ("I feel responsible for what occurs on industrial animal farms"). A single-item measure evaluated empathy ("I can appreciate how farm animals must think and feel").

Seven items comprised a commitment to values scale. This scale was inspired by Aquino and Reed's (2002) internalisation/symbolisation formulation of moral values importance, but reframed to reflect commitment to animal welfare specifically (e.g., "I am committed to animal welfare," "Farm animals deserve our moral concern," "I want others to know that I believe animals are our responsibility"; α = .90).

Moral engagement was assessed using six items from the Genuine Self-Forgiveness subscale of the Differentiated Process Scale of Self-Forgiveness (Woodyatt & Wenzel, 2013b). Items were amended to match the context, for example, "I am trying to think through why I eat meat products" (α = .70). The original scale had seven items, but one item was not included in Study 4.1 because it was not suited to the transgression context. This seventh item was added into the survey battery from Study 4.2 onwards.

Reparation consisted of five items reflecting desire and intention to change eating behaviours and support animal welfare outcomes (e.g., "I really hope to change my food choices," "I intend to reduce my meat consumption," "I would support legislation to change food production practices"; α = .80).

Three behavioural measures of reparation were also used. The first item asked if participants were interested in donating money (any amount of their choice) to an animal welfare organisation. The second item asked participants if they would be willing to spend

some time writing emails to governmental bodies to lobby on behalf of animal rights (they could decline this offer, or nominate either 5, 10 or 15 minutes of their time). A third item asked participants if they would like to sign up to a newsletter from the same organisation (they could decline or accept this offer).[1]

A manipulation check was included to determine whether the auditory task was indeed perceived as unpleasant.

Results

Results of the manipulation check confirmed that participants in the self-punishment condition ($M = 4.67$, $SD = 1.72$) found the task significantly more unpleasant than those in the control condition ($M = 2.86$, $SD = 2.22$), $t(65.97) = -3.86$, $p < .001$.

Main effects of self-punishment. Fewer than 15% of participants volunteered any money or time to contribute to animal rights, thus these two behavioural measures were not analysed as planned. For the third behavioural measure, participants in the self-punishment condition (72%) were no more or less likely than those in the control condition to sign up for the animal rights newsletter (72%). Descriptive statistics for all remaining outcome measures are presented in Table 4.1 (see also Appendix B for inter-item correlations). No significant differences were found between group means on any of the variables.

[1] Four additional measures were initially included in the survey battery but are not reported here. These four measures were dropped after Study 4.2. A three-item measure of justice restoration and a two-item measure of avoidance were created, but there were no main effects of self-punishment condition on either of these two measures, nor were there moderated effects on these measures in line with the predicted model. In addition, we created a seven-item self-victimisation measure, and an eight-item measure of neutralisation, but these measures showed poor internal reliability and factor structure and thus were not considered valid for further analysis. See also Footnote 2.

Table 4.1

T-Test Statistics for Differences Between Experimental Conditions (Study 4.1)

	M (*SD*)				
Dependent variable	Control	Self-punishment	*t* (df)	*p*	*d*
Moral identity threat	2.51 (0.92)	2.73 (0.89)	1.04 (70)	.301	0.24
Guilt	4.03 (1.75)	4.14 (2.29)	0.23 (65.39)	.818	0.05
Shame	3.69 (1.75)	3.75 (2.13)	0.12 (70)	.904	0.03
Responsibility	3.61 (1.89)	4.11 (1.72)	1.18 (70)	.244	0.28
Empathy	5.92 (1.36)	6.06 (1.15)	0.47 (70)	.641	0.11
Values	5.61 (1.08)	5.69 (1.15)	0.29 (70)	.775	0.07
Moral engagement	4.60 (0.95)	4.92 (1.01)	1.36 (70)	.179	0.33
Reparation	5.42 (0.93)	5.49 (1.29)	0.27 (70)	.786	0.06

Note. Control condition $n = 36$, self-punishment condition $n = 36$.

Moral need as a moderator of self-punishment effects. Next, we tested our proposed model of self-punishment. Since the moderator (moral identity threat) was measured subsequent to the self-punishment manipulation, we first determined whether self-punishment might have influenced scores on this measure. Moral identity threat did not differ significantly across experimental conditions (see Table 4.1), thus we proceeded with our analysis. Results revealed a range of moderated effects on guilt, shame, responsibility, commitment to values, and reparation. See Appendix C for a tabulated record of all tested moderations.

There was a significant interaction between experimental condition and moral identity threat on shame, $B = 0.67$, $SE = 0.24$, $p = .01$. The overall model (including main effects and the interaction effect) explained 16% of the variance in shame, $F(3,68) = 4.31$, $p = .01$. Following Aiken and West's (1991) procedure, simple slopes for the association between condition and shame were calculated for low (-1 SD below the mean) and high (+1 SD above the mean) levels of identity threat. Consistent with predictions, at low levels of identity threat, self-punishment decreased shame (i.e., self-punishment "cleansed" shame), $B = -0.64$,

SE = .031, *p* = .04. At high levels of identity threat, self-punishment marginally increased shame (in line with moral repair), *B* = 0.57, *SE* = .031, *p* = .07. These simple slopes are plotted in Figure 4.2 to illustrate the nature of the interaction; note that all subsequent moderations followed the same pattern.

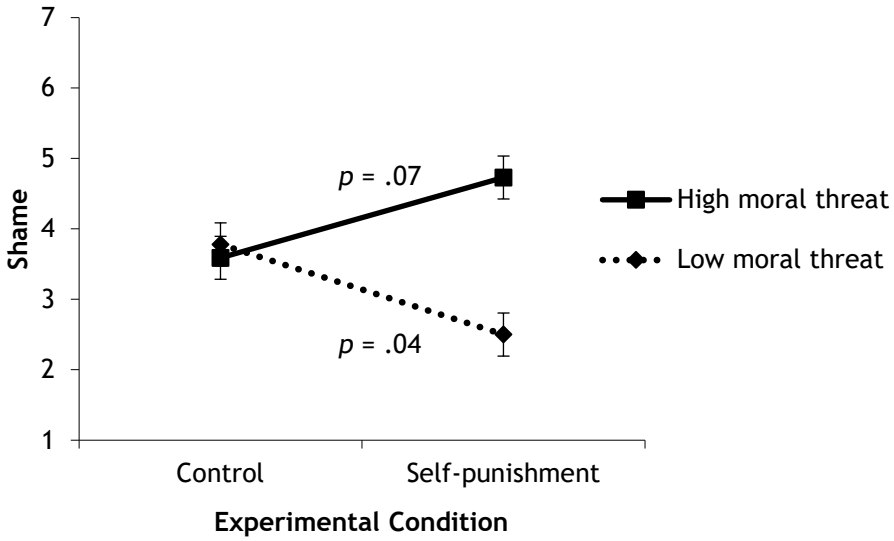


Figure 4.2. Simple slopes with standard error bars, illustrating the interaction effect between self-punishment and moral threat on shame (Study 4.1). Self-punishment reduces shame for those reporting low threat (-1 SD above the centred mean), and increases shame for those reporting high threat (+1 SD above the centred mean).

Identity threat marginally moderated the relationship between self-punishment and guilt, *B* = 0.46, *SE* = 0.26, *p* = .08, explaining 10% of the variance in guilt, *F*(3,68) = 2.54 , *p* = .06. At low levels of threat self-punishment tended to reduce guilt, *B* = -0.42, *SE* = 0.33, *p* = .21. At high levels of threat, this relationship trended in the opposite direction, *B* = 0.41, *SE* = 0.33, *p* = .22; however, neither simple effect was statistically significant.

Similar interactions were found when considering outcome variables of responsibility, *B* = 0.65, *SE* = 0.23, *p* = .01, and commitment to values, *B* = 0.30, *SE* = 0.14, *p* = .04. Each model accounted for 13% (*F*[3,68] = 3.30, *p* = .03) and 14% (*F*[3,68] = 3.57, *p* = .02) of the

variance in the outcome variables, respectively. When participants perceived a low identity threat, self-punishment did not significantly influence responsibility, B = -0.33, SE = 0.29, p = .26, or commitment to values, B = -0.20, SE = 0.18, p = .27. When threat to moral identity was high, self-punishment increased responsibility, B = 0.85, SE = 0.29, p = .01, and commitment to values, B = 0.35, SE = 0.18, p = .06.

The relationship between self-punishment and reparation was also moderated by identity threat, B = 0.33, SE = 0.14, p = .02, the model explaining 13% of the variance in reparation $F(3,68) = 3.30$, p = .03. Neither of the simple slopes reached statistical significance, but the pattern of effects was consistent with predictions. At low identity threat self-punishment did not influence reparation, B = -0.29, SE = 0.18, p = .13, at high identity threat it marginally tended to increase reparation, B = 0.31, SE = 0.18, p = .09. The interaction suggests that self-punishment may either strengthen or weaken reparation, depending on one's perception of moral need.

Identity threat did not moderate the relationship between self-punishment and empathy or between self-punishment and moral engagement.

Discussion

There were no main effects of self-punishment on a range of emotional, cognitive, and behavioural intention variables. These results are contrary to past findings indicating that self-punishment reduces guilt (Bastian et al., 2011; Inbar et al., 2013). The moderated effects detected instead suggest that the meaning and function of self-punishment varies according to the acknowledgment of moral need. Specifically, when participants perceived a substantial threat to their moral identity, self-punishment increased commitment to values, guilt, shame, acceptance of responsibility, and reparation. This picture is consistent with a moral repair process. In contrast, when there was little identity threat, only then did self-punishment diminish commitment to values, guilt, shame, responsibility, and reparation, pointing to a

moral cleansing process. In summary, there was evidence for both functions of self-punishment, moderated by moral need. To the extent that individuals felt they had done something substantially immoral, threatening their self-concept and demanding action that could be ignored, self-punishment was a tool to affirm their values and examine their act (i.e., moral repair). On the other hand, for those who were less threatened, self-punishment provided a way to cleanse themselves of the implications of the transgression.

However, there was a major limitation to Study 4.1: The identity threat moderator was measured following the self-punishment manipulation. The independent variable should not influence the moderator variable; under the current design this could not be guaranteed, though scores did not differ between conditions. Furthermore, it could be that the findings are only relevant to non-relational transgressions. We sought to further validate the model in response to traditional interpersonal transgressions.

Study 4.2

Study 4.2 used the same design as Study 4.1, except for two key changes. We had two primary aims: to test for the moral need moderation in a traditional interpersonal transgression context, and to provide a more robust test of our model of self-punishment. To facilitate the former aim, we substituted an interpersonal transgression for the animal video. To facilitate the latter aim, the moral identity threat measure was moved from a post-manipulation measure to a pre-manipulation measure, where it would be better placed to test for its moderating role in shaping the nature of self-punishment. We also expanded measures of moral need by adding measures of guilt and shame prior to self-punishment, as well as perceived transgression severity and the closeness between participants and their victims. Appraising a transgression as severe and one that jeopardises an important relationship should lead to a higher moral identity threat, since it is more relevant and meaningful to one's

identity (J. L. Tracy & Robins, 2004). In turn, this should lead to higher levels of guilt and shame, thus these may be considered emotional markers of moral need.

Method

Participants. Eighty-nine participants were recruited from the student participation pool at Flinders University, 70% female, 30% male, mean age 21.35 years (*SD* = 5.81).

Procedure and measures. Participants were induced to transgress by recalling a recent interpersonal transgression. Using this type of transgression with our auditory self-punishment measure will thus constitute a test of the generalisability of previous self-punishment findings to another type of pain (Bastian et al., 2011; Inbar et al., 2013).

In order to elicit sharp emotional responses and motivation to engage in identity-repair strategies such as self-punishment, we elicited only recent transgressions, using a funnel procedure (Woodyatt & Wenzel, 2013b). We first asked participants whether they could recall a time they had "hurt, offended, or done wrong by another person" in the last two days. If they could not recall such an incident, then we probed for one in the last week, then the last two weeks, and finally within the last month (maximum time allowed). Participants were asked to write about what happened and then rated it on perceived severity ("How serious do you think what you did was?") and importance of the relationship ("How important to you is your relationship with this person?" with an option, "N/A- I didn't hurt a specific person," recorded as missing data). Additionally, measures of moral identity threat, guilt, and shame were collected at this point. The moral identity threat measure was identical to Study 4.1. Guilt and shame items from Study 4.1 were reworded to fit the interpersonal context (e.g., "When I think about what I have done I feel guilty").

The auditory task from Study 4.1 was used again as self-punishment. We tried to further distinguish the two experimental conditions by substituting pleasant piano tones for the lower volume white noise blasts in the control condition.

To avoid arousing suspicion by asking about guilt and shame twice, as per Inbar et al. (2013) a larger battery of post-test emotions was given in the form of the Positive And Negative Affect Schedule (D. Watson et al., 1988). This scale includes guilt and shame. Response scales were altered to a seven-point format to match the rest of the survey items.

In order to complement the measure of moral engagement, we measured moral disengagement with the seven-item pseudo self-forgiveness subscale of the Differentiated Process Scale of Self-Forgiveness (Woodyatt & Wenzel, 2013b). Example items: "I'm not really sure whether what I did was wrong," "I wasn't the only one to blame for what happened" ($\alpha = .80$).

To measure interpersonal reparation, we used Woodyatt and Wenzel's (2013b) six-item reconciliation scale, which includes items such as "I really hope to make things right with this person," and "I want to apologise to this person" ($\alpha = .91$).

Outcome measures of commitment to values and moral engagement were almost identical to those in Study 4.1, with only minor changes to item wording to match the interpersonal transgression context. The amended scales from Study 4.1 showed good internal consistency (commitment to values $\alpha = .80$; moral engagement $\alpha = .78$).[2]

Results

Participants in the self-punishment condition ($M = 4.17$, $SD = 2.05$) found the task significantly more unpleasant than those in the control condition ($M = 1.86$, $SD = 1.42$), $t(80.52) = -6.22$, $p < .001$. The change to the tones in the control condition was successful: The mean unpleasantness score for the control condition was lower than in Study 4.1.

[2] As in Study 4.1, we also included measures of avoidance, justice restoration, neutralisation and self-victimisation. Neutralisation and self-victimisation items were amended in an attempt to overcome the reliability issues identified in Study 4.1, but nonetheless continued to show poor internal reliability and validity. In addition, the avoidance and justice restoration measures showed poor internal reliability and validity in Study 4.2. Thus, these four measures were deemed unreliable for further analysis, and were subsequently dropped from the survey battery.

Table 4.2

T-Test Statistics for Differences Between Experimental Conditions (Study 4.2)

	M (*SD*)				
Dependent variable	Control	Self-punishment	*t* (df)	*p*	*d*
Moral need variables					
Moral identity threat	3.40 (0.96)	3.21 (0.88)	-0.94 (87)	.351	-0.21
Severity	4.74 (1.24)	4.07 (1.39)	-2.43 (87)	.017	-0.51
Relationship	5.14 (1.83)	5.50 (1.92)	0.91 (87)	.368	0.19
Guilt	5.16 (1.62)	4.96 (1.44)	-0.64 (87)	.527	-0.13
Shame	4.91 (1.82)	4.54 (1.83)	-0.94 (87)	.351	-0.20
Post-manipulation DVs					
Guilt	3.28 (2.02)	3.13 (1.76)	-0.37 (87)	.711	-0.08
Shame	2.84 (1.75)	2.70 (1.62)	-0.40 (87)	.691	-0.08
Responsibility	5.62 (1.25)	5.63 (1.65)	0.01 (87)	.994	0.01
Empathy	4.93 (1.83)	5.37 (1.58)	1.21 (87)	.228	0.26
Values	5.45 (1.06)	5.70 (0.85)	1.25 (87)	.215	0.26
Moral engagement	4.66 (0.93)	4.61 (1.17)	-0.25 (87)	.805	-0.05
Moral disengagement	3.21 (1.08)	3.07 (1.31)	-0.55 (87)	.582	-0.12
Reparation	5.11 (1.63)	5.40 (1.58)	0.86 (87)	.395	0.18

Note. Control condition $n = 43$, self-punishment condition $n = 46$.

Main effects of self-punishment. Descriptive statistics for mean scores across the two experimental conditions are presented in Table 4.2. There were no significant differences between the two conditions on any outcome measure.

As a further test of changes in guilt as a result of the manipulation, pre- and post-test guilt scores were subjected to a mixed ANOVA with condition as a between-subjects factor and guilt scores as the within-subjects factor (two time points). There was no significant interaction effect, indicating that the changes in guilt between the two measurement times were equal across conditions, $F(1,87) = 0.02$, $p = .89$, partial $\eta^2 = .00$. Similarly, shame was not differentially affected by the experimental manipulation, $F(1,87) = 0.28$, $p = .60$, partial $\eta^2 = .003$. Both groups experienced reductions in both guilt and shame (see Table 4.3), which

could have been an artefact of time, or perhaps because the pre- and post-test items were worded slightly differently. Thus we cannot speak conclusively of "reductions" per se, but we can declare that self-punishment did not have any main effects on guilt and shame relative to control.

Table 4.3

Pre- and Post-Manipulation Guilt and Shame Scores (Study 4.2)

	Guilt			Shame		
	Pre	Post	*p*	Pre	Post	*p*
Condition	*M (SD)*	*M (SD)*		*M (SD)*	*M (SD)*	
Control	5.16 (1.62)	3.28 (2.02)	<.001	4.91 (1.82)	2.83 (1.74)	<.001
Self-punishment	4.96 (1.44)	3.13 (1.76)	<.001	4.54 (1.83)	2.69 (1.62)	<.001

Moral need as a moderator of self-punishment effects. Before testing for moderations by moral need, we examined whether the moral need measures were related to one another—as one might expect. Correlations between the variables (Table 4.4) generally support the reasoning that the measures are related and may reflect different dimensions or markers of moral need. The exception to this pattern was the moral identity threat scale—it did not correlate significantly with any of the other moral need variables. It should also be noted that there was a statistically significant difference between experimental conditions on severity, which indicates the randomisation resulted in chance differences. Though ideally there should be no relationship between the predictor variable and the moderator variable, since the moral need variables were measured before the experimental manipulation we can be confident that there is no causal relationship between the two. We thus proceeded with the proposed moral need moderation. Each of the moral need variables is considered in turn (see also Appendix C).

Table 4.4

Pearson Correlations Between Moral Need Variables (Study 4.2)

Measure	1	2	3	4	5
1 Identity threat	1				
2 Severity	.17	1			
3 Relationship	-.14	.24*	1		
4 Guilt	.12	.52**	.25*	1	
5 Shame	.12	.50**	.35**	.82**	1

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

Severity. Severity moderated self-punishment effects on moral engagement, moral disengagement, and marginally on shame and responsibility. For moral engagement, B = 0.18, SE = 0.08, p = .04, the model explained 14% of the variance in moral engagement, $F(3,85) = 4.52$, $p < .01$. Simple slopes indicated that in response to low severity transgressions, self-punishment did not affect moral engagement, B = -0.19, SE = 0.16, p = .23, whereas for high severity transgressions self-punishment marginally increased moral engagement, B = 0.28, SE = 0.15, p = .07 (see Figure 4.3).

The interaction between severity and condition also predicted moral disengagement, B = -0.20, SE = 0.10, p = .04, the model accounting for 10% of the variance, $F(3,85) = 3.05$, p = .03. For low severity transgressions, self-punishment had no effect on moral disengagement, B = 0.15, SE = 0.19, p = .44, but for more serious transgressions self-punishment reduced moral disengagement, B = -0.40, SE = 0.18, p = .03. Thus self-punishment facilitated moral engagement only when individuals felt they had committed a particularly serious wrongdoing.

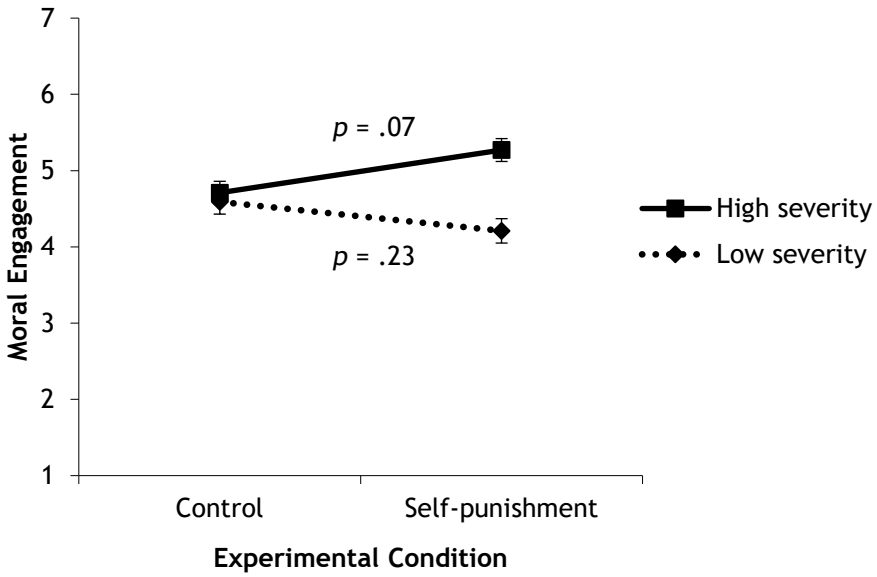


Figure 4.3. Simple slopes with standard error bars, illustrating the interaction effect between self-punishment and severity on moral engagement (Study 4.2), showing that self-punishment only increased moral engagement for those committing more serious transgressions.

The interaction between severity and self-punishment also marginally predicted post-manipulation shame, $B = 0.24$, $SE = 0.13$, $p = .08$. The overall model explained a significant 9% of the variance in shame, $F(3,85) = 2.74$, $p = .05$. For low severity transgressions, self-punishment did not influence shame, $B = -0.31$, $SE = 0.26$, $p = .24$. For individuals committing more serious transgressions, self-punishment was marginally more likely to increase shame, $B = 0.36$, $SE = 0.25$, $p = .08$.

Moreover, there was a marginally significant interaction between severity and condition on responsibility, $B = 0.20$, $SE = 0.11$, $p = .07$, the model accounting for 16% of the variance in responsibility, $F(3,85) = 5.52$, $p = .002$. For participants committing less serious transgressions, self-punishment did not influence responsibility, $B = -0.15$, $SE = 0.22$, $p = .49$. In cases of more serious transgressions, self-punishment marginally increased responsibility, $B = 0.40$, $SE = 0.21$, $p = .06$.

Relationship. Relationship importance, another proxy for moral need, also played a role in shaping self-punishment outcomes, moderating effects on moral engagement and marginally so on commitment to values and moral disengagement.

Condition and relationship interacted to marginally predict commitment to values, B = 0.10, SE = 0.05, p = .06. The model accounted for 11% of the variance in commitment to values, $F(3,85) = 3.44$, p = .02. When victims were less important to transgressors, self-punishment did not have an effect on commitment to values, B = -0.08, SE = 0.14, p = .56. In contrast, when participants rated the relationship as important, self-punishment boosted commitment to values, B = 0.30, SE = 0.14, p = .04.

Relationship significantly moderated the relationship between self-punishment and moral engagement, B = 0.13, SE = 0.05, p = .02, explaining 23% of the variance in moral engagement, $F(3,68) = 8.52$, $p < .001$. At low levels of relationship importance, self-punishment reduced moral engagement, B = -0.32, SE = 0.14, p = .03. For high importance relationships, self-punishment did not influence moral engagement, B = 0.18, SE = 0.14, p = .21.

Furthermore, relationship and condition interacted to predict moral disengagement, a marginally significant finding, B = -0.12, SE = 0.07, p = .06. The model accounted for 10% of the variance in moral disengagement, $F(3,68) = 3.19$, p = .03. When accompanied by low relationship importance, self-punishment tended to increase moral disengagement, B = 0.19, SE = 0.17, p = .28, but when accompanied by high relationship importance, self-punishment tended to reduce moral disengagement, B = -0.28, SE = 0.18, p = .12, though neither simple slope was statistically significant.

Guilt. There was a marginally significant interaction between condition and pre-self-punishment guilt on responsibility, B = 0.18, SE = 0.10, p = .08, with 11% explained variance in the model, $F(3,68) = 3.58$, p = .02. Though neither simple slope was statistically

significant, a pattern could be observed whereby self-punishment tended to reduce acceptance of responsibility for participants who did not feel particularly guilty about their actions, B = -0.24, SE = 0.21, p = .26, but for those feeling more guilty, it tended to result in more responsibility, B = 0.30, SE = 0.21, p = .16.

Shame and moral identity threat. Shame did not moderate any self-punishment effects, nor did moral identity threat.

Discussion

Results provided support for the two processes of self-punishment predicted by our model. When subject to lower moral need (less transgression severity, less important relationship violated, and less guilt experienced), self-punishment had a moral cleansing effect (less genuine self-forgiveness, responsibility, shame, and commitment to values, and more moral disengagement). When participants indicated higher moral need, self-punishment increased moral engagement, in line with moral repair.

However, we failed to replicate the moderation using the moral identity threat scale from Study 4.1. Rather, interaction effects in Study 4.2 were driven by severity, relationship closeness, and guilt. The reason for this discrepancy between the two studies is unclear. It could be that identity threat is not a valid measure of moral need, and the moderation effects found in Study 4.1 were invalid—potentially contaminated by the self-punishment manipulation (since in Study 4.1 moral threat was measured following the self-punishment manipulation). However, at this stage we cannot exclude that these findings were due to random variance. Further experimentation is needed to confirm the results.

Study 4.3

Study 4.3 was a conceptual replication of Study 4.2. We sought to again find evidence for the proposed model of self-punishment, generalised to another form of self-punishment: cold stress. We used the ice bucket task used in previous self-punishment research (Bastian et

al., 2011; van Bunderen & Bastian, 2014) as a more traditional measure of pain. However, we attempted to minimise the performance or endurance element to this task by making a subtle change to the instruction.

Method

Participants. Participants ($N = 91$) were again recruited from the student participation pool at Flinders University, 77% female, 23% male, mean age 22.46 years ($SD = 7.15$).

Procedure and measures. The procedure from Study 4.2 was followed, except for a small change to the transgression manipulation, and the substitution of the ice bucket task for the auditory self-punishment task.

For the transgression, rather than eliciting only recent transgressions, we asked participants to "think about a time when you did something to hurt someone else, something so bad that you still feel guilty about it today." This instruction allowed us to include those who had not transgressed recently, widening the participant pool.

Participants in the self-punishment condition were presented with a bucket of iced water (0°-2°C) and instructed to "hold your non-dominant hand, up to your wrist, in this water for as long as you want to. You are free to take your hand out whenever you've decided you've had enough; there is no minimum time required." The original instruction was simply to hold one's hand in the water "for as long as you can" (Bastian et al., 2011); our variation was designed to emphasise the voluntary nature of the task. The experimenter used a stopwatch to measure how long participants held their hand in the water, up to a maximum of two minutes. Participants in the control condition completed the task with a bucket of lukewarm water rather than iced water, but otherwise followed the same procedure.

Measures showed acceptable internal consistency (moral identity threat α = .71, commitment to values α = .68; moral engagement α = .88, moral disengagement α = .81, reparation α = .89).

Results

Scores on the manipulation check confirmed that participants in the self-punishment condition (M = 5.04, SD = 1.41) found the bucket task significantly more unpleasant than those in the control condition (M = 2.07, SD = 1.45), $t(89) = -9.90$, $p < .001$.

Main effects of self-punishment. There were two significant differences between the groups on the outcome measures (see Table 4.5): Self-punishment reduced responsibility and guilt. However, since the groups may have differed in how much guilt they started off with, the latter effect was followed up with a mixed ANOVA with pre- and post-manipulation measures of guilt. Results indicated that both groups experienced "reductions" in guilt (see Table 4.6), but these changes were not significantly different between conditions, $F(1,89) = 2.04$, $p = .16$, partial $\eta^2 = .02$, This suggests that self-punishment did not have as considerable an effect on guilt as the t test had indicated. Rather, the self-punishment group started with slightly less guilt to begin with, which partially contributed to their lower guilt scores following the manipulation. Changes between pre- to post-shame did not differ significantly between groups, F (1,89) = 1.13, $p = .29$, partial $\eta^2 = .01$.

Table 4.5

T-Test Statistics for Differences Between Experimental Conditions (Study 4.3)

	M (*SD*)				
Dependent variable	Control	Self-punishment	*t* (df)	*p*	*d*
Moral need variables					
Moral identity threat	2.84 (1.04)	3.01 (1.07)	0.79 (89)	.432	0.16
Severity	4.80 (1.53)	4.57 (1.64)	-0.66 (89)	.509	-0.15
Relationship	5.41 (2.05)	5.02 (2.12)	-0.89 (89)	.378	-0.19
Guilt	5.59 (1.85)	5.17 (1.87)	-1.08 (89)	.283	-0.23
Shame	5.16 (2.00)	5.11 (1.97)	-0.13 (89)	.900	-0.03
Post-manipulation DVs					
Guilt	3.41 (1.60)	2.32 (1.70)	-3.15 (89)	.002	-0.67
Shame	2.73 (1.63)	2.15 (1.67)	-1.67 (89)	.099	-0.35
Responsibility	6.34 (0.89)	5.49 (1.60)	-3.17 (72.81)	.002	-0.66
Empathy	5.91 (1.40)	5.66 (1.56)	-0.81 (89)	.423	-0.17
Values	6.11 (0.59)	5.92 (0.71)	-1.38 (89)	.171	-0.29
Moral engagement	5.37 (1.19)	5.22 (1.17)	-0.62 (89)	.535	-0.13
Moral disengagement	3.10 (1.07)	3.53 (1.39)	1.64 (85.76)	.103	0.35
Reparation	5.53 (1.40)	5.28 (1.42)	-0.83 (89)	.407	-0.18

Note. Control condition $n = 44$, self-punishment condition $n = 47$.

Table 4.6

Pre- and Post-Manipulation Guilt and Shame Scores (Study 4.3)

	Guilt			Shame		
Condition	Pre *M (SD)*	Post *M (SD)*	*p*	Pre *M (SD)*	Post *M (SD)*	*p*
Control	5.59 (1.85)	3.41 (1.60)	<.001	5.16 (2.00)	2.73 (1.63)	<.001
Self-punishment	5.17 (1.87)	2.32 (1.70)	<.001	5.11 (1.97)	2.15 (1.67)	<.001

Moral need as a moderator of self-punishment effects. Correlations between moral need variables indicated that guilt, shame, severity and relationship importance were related, though as found in the previous study, identity threat was not as strongly correlated with the

other measures (see Table 4.7). All of these variables were then used as moderators to test for the proposed model of self-punishment. Significant and marginal models are reported below; all were consistent with the predicted effects.

Table 4.7

Pearson Correlations Between Moral Need Variables (Study 4.3)

Measure	1	2	3	4	5
1 Identity threat	1				
2 Severity	.18	1			
3 Relationship	.07	.39**	1		
4 Guilt	.02	.57**	.23*	1	
5 Shame	.19	.50**	.19	.78**	1

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

Severity. Severity moderated the relationship between self-punishment and responsibility, $B = 0.17$, $SE = 0.08$, $p = .03$, explaining 31% of the variance in responsibility, $F(3,87) = 12.73$, $p < .001$. When participants perceived their transgression as low in severity, self-punishment resulted in less responsibility, $B = -0.66$, $SE = 0.17$, $p < .001$, while high severity transgressors maintained their sense of responsibility following self-punishment, $B = -0.12$, $SE = .17$, $p = .49$ (see Figure 4.4).

There was also a marginal interaction between self-punishment and severity on guilt, $B = 0.19$, $SE = 0.11$, $p = .08$, explaining 19% of the variance, $F(3,87) = 6.97$, $p < .001$. Self-punishment reduced guilt only when participants reported low severity transgressions, $B = -0.82$, $SE = .24$, $p < .001$; it did not significantly influence guilt when participants reported more serious transgressions, $B = -0.21$, $SE = 0.24$, $p = .37$.

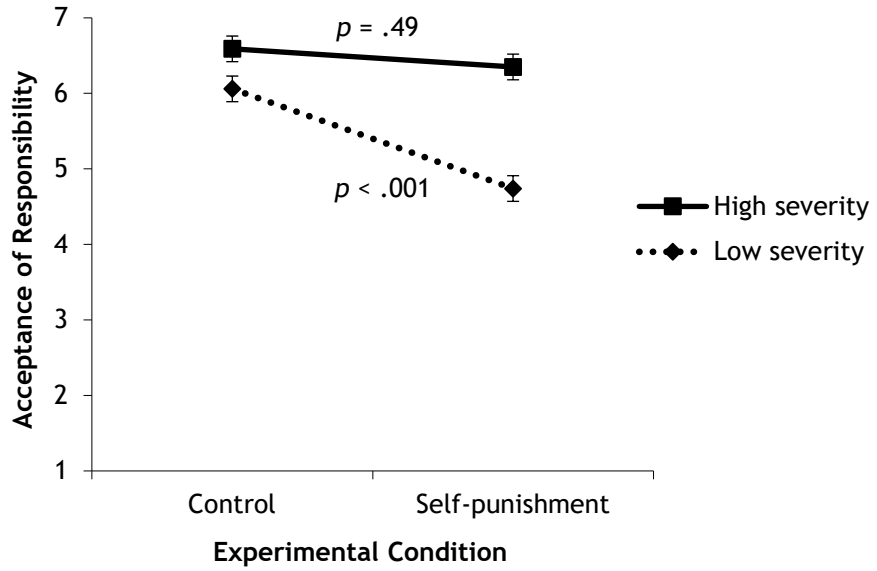


Figure 4.4. Simple slopes with standard error bars, illustrating the interaction effect between self-punishment and severity on acceptance of responsibility (Study 4.3), showing that self-punishment reduced responsibility only for those committing less serious transgressions.

Shame. Shame and self-punishment interacted to predict empathy, $B = 0.18$, $SE = 0.07$, $p = .02$, and moral disengagement, $B = -0.12$, $SE = 0.06$, $p = .03$; the models explained 16% of the variance in empathy, $F(3,87) = 5.39$, $p = .002$, and 37% of the variance in moral disengagement, $F(3,87) = 16.78$, $p < .001$. Consistent with our proposed model, those who were less shameful about their transgressions felt less empathy, $B = -0.47$, $SE = 0.21$, $p = .03$, and morally disengaged from their wrongdoing as a result of their wrongdoing, $B = 0.45$, $SE = 0.15$, $p = .004$. Meanwhile, for those who were more shameful, self-punishment tended to maintain their sense of empathy, $B = 0.23$, $SE = 0.21$, $p = .27$, and did not influence moral disengagement, $B = -0.04$, $SE = 0.15$, $p = .81$.

Moral identity threat. Identity threat interacted with self-punishment to predict guilt (marginal), $B = 0.30$, $SE = .16$, $p = .07$, and empathy, $B = 0.31$, $SE = 0.14$, $p = .03$; the models accounted for 17% of the variance in guilt, $F(3,87) = 5.88$, $p < .001$, and 8% of the variance in empathy, $F(3,87) = 2.46$, $p = .07$ (marginal overall model for empathy). When participants

perceived a low identity threat, self-punishment did not influence guilt, $B = 0.26$, $SE = 0.24$, $p = .29$, or empathy, $B = -0.19$, $SE = 0.22$, $p = .38$. With high threat, however, self-punishment increased guilt, $B = 0.88$, $SE = 0.24$, $p < .001$, and empathy, $B = 0.47$, $SE = 0.21$, $p = .03$.

Relationship and guilt. Effects of self-punishment did not depend on either relationship importance or guilt.

Discussion

Effects of self-punishment largely depended on the degree to which participants acknowledged a moral need. In line with predictions, when participants reported low moral need they used self-punishment to cleanse themselves of their wrongdoing. When participants reported high moral need, self-punishment appeared to affirm their wrongdoing, which maintained—and in some cases even strengthened—their sense of guilt, responsibility, victim empathy, and moral engagement. However, as moral need variables have only been measured in Studies 4.1–4.3, their causal role in motivating a particular function of self-punishment has not yet been determined. Thus, a final study was conducted to test for a causal relationship between moral need and self-punishment processes.

Study 4.4

The procedure for Study 4.4 was almost identical to Studies 4.2 and 4.3, but with two changes. First, we added a moral need manipulation by varying instructions for the autobiographical transgression recall. In this manipulation, we asked participants to either recall a time they hurt a person who was unimportant to them (e.g., a stranger) or a time they hurt a person who was close to them. Thus the study was a 2 (relationship: distal, close) x 2 (no self-punishment, self-punishment) factor between-subject design. We predicted that transgressions against more important relationships would be more meaningful to participants' identities, and be perceived as more serious, invoking more guilt and shame.

Theoretically, this should invoke a greater threat to one's moral identity, and transgressors should be more likely to acknowledge this as a moral need.

Second, we considered whether such perceptions of moral need might not be driven so much by objective features of the transgression but rather by individual differences in how threats tend to be perceived. That is, we sought to test whether personality traits could account for whether individuals used self-punishment to cleanse or to repair their moral identity. Therefore we also asked participants to complete a short personality measure that we predicted might also moderate self-punishment effects on the outcome variables. The perpetrator justice sensitivity scale (Schmitt, Baumert, Gollwitzer, & Maes, 2010) is a measure of the extent to which one sees oneself as the *perpetrator* of a wrongdoing; those high in this trait might be more willing to acknowledge their wrongdoing and a moral need.

Van Bunderen and Bastian (2014) found a differentiated effect of self-punishment on victim reparation using the *victim* justice sensitivity scale of this measure. For those who were more likely to view themselves as victims of injustice, self-punishment reduced their motivation to compensate their victim, while for those who scored lower on this scale, self-punishment did not dampen their motivation to make amends. Though these results are broadly consistent with our identity-regulating model of self-punishment, perpetrator justice sensitivity might be more closely related to moral need acknowledgment, thus we used the perpetrator sensitivity scale instead (Schmitt et al., 2010).

Method

Participants. As per the protocol for the previous studies, we recruited 114 students for participation, 76% female, 24% male, mean age 21.53 years (SD = 5.21).

Procedure and measures. Upon signup to the study, participants were instructed to complete a short survey online at least 12 hours prior to the laboratory session. The online survey included demographic variables and the 10-item perpetrator justice sensitivity scale

(Schmitt et al., 2010). Example items include, "I feel guilty when I enrich myself at the cost of others," and "It bothers me when I use tricks to achieve something while others have to struggle for it" (α = .91).[3]

Upon arrival to the laboratory, participants were randomly assigned to either the distal or close relationship transgression condition. Participants were asked to think about a time they hurt or victimised someone else, either: "someone who you did not know very well, someone who you were not close with, like a stranger or mere acquaintance, with whom you merely had a casual or business-like interaction" (*n* = 59); or "someone who you knew very well, someone who you were close with, like a friend or family member, someone with whom you had a personal relationship" (*n* = 55). This manipulation was previously used by Wenzel and Okimoto (2012). Participants then rated the transgressions on moral need variables as per the previous two studies (moral identity threat α = .74).

Following the moral need measures, participants were randomly assigned to self-punish (or not), using the auditory task from Studies 4.1 and 4.2. Finally, participants completed the same outcome variable battery as in the last two studies (moral engagement α = .78, moral disengagement α = .79, reparation α = .92), with the exception of the commitment to values scale. Since Study 4.1, commitment to values had not emerged as a significant outcome variable in the moderated self-punishment effect analyses. Thus, the scale was changed to reflect *shared* value consensus between the offender and victim, rather than abstract value commitment, as this is more directly relevant to the transgression and indeed how the concept of value recommitment had been previously conceptualised (Wenzel, Okimoto, Feather, & Platow, 2010; Wenzel, Woodyatt, & Hedrick, 2012). We used six items based on Wenzel et al. (2012), for example "The person I hurt and I would agree on what is right and wrong" (α = .86).

---

[3] Participants also completed trait measures of depressive self-criticism, and belief in a just world (Lerner & Miller, 1978), but these were included to test other hypotheses not reported here.

We also considered that effects could be influenced by the extent to which transgressors has already been forgiven (i.e., there is no outstanding identity threat that requires action). Thus, as control variables at the end of the study we also asked participants, "To what extent do you feel you have been forgiven for what you did?" (7-point rating scale from "not at all" to "very much") and "Are you still in contact with the victim?" (binary).

Results

Manipulation checks. See Table 4.8 for cell means and statistics for moral need variables. Participants in the close relationship condition reported transgressing against a more important relationship compared to those in the distal relationship condition. This translated to more serious transgressions in the close relationship condition relative to distal condition. However, this effect did not carry through to guilt, shame, or moral identity threat. This suggests that the manipulation was problematic in terms of manipulating moral need acknowledgment.

Replicating the previous studies, participants in the self-punishment condition (M = 4.68, SD = 1.65) found the task significantly more unpleasant than those in the control condition (M = 1.68, SD = 1.26), $t(104.60) = -10.93$, $p < .001$.

Table 4.8

T-Test Statistics for Differences Between Relationship Manipulation Conditions (Study 4.4)

	M (*SD*)				
Dependent variable	Distal	Close	*t* (df)	*p*	*d*
Moral identity threat	3.36 (0.92)	3.63 (1.22)	1.37 (112)	.173	0.25
Severity	3.71 (1.42)	4.93 (1.33)	4.72 (112)	< .001	0.89
Relationship	2.74 (1.69)	5.91 (1.61)	10.18 (111)	< .001	1.92
Guilt	5.41 (1.70)	5.33 (1.74)	-0.25 (112)	.806	-0.05
Shame	4.86 (1.81)	4.78 (1.84)	-0.24 (112)	.810	-0.04

Main effects and condition interactions. We next examined main effects of the self-punishment manipulation, as well as the predicted interaction between the two independent variables (relationship x self-punishment). There were neither main effects of self-punishment nor interaction effects on any of the outcome variables.[4] An exception was on guilt, where a significant interaction was observed, B = -0.41, SE = 0.16, p = .01. Going against predictions, simple effects revealed that self-punishment increased guilt for those in the distal relationship condition, $t(57)$ = -2.53, p = .01, while it did not affect guilt for those in the close relationship condition, $t(53)$ = 1.43, p = .16. Thus, our prediction was not supported.

We followed up the effect on guilt with a mixed ANOVA. The three-way interaction between self-punishment condition, relationship condition, and guilt (repeated measures factor) was not significant, indicating that changes in guilt did not depend on the manipulations, $F(1,110)$ = 1.76, p = .19, partial η^2 = .02. Mean scores revealed there were reductions in guilt across all four cells, but the slopes were of a similar magnitude across relationship conditions. The mean differences in post-guilt between the control and self-punishment conditions for those exposed to the distal manipulation might be partially explained by pre-manipulation differences between these two cells (i.e., the self-punishment condition had elevated pre-guilt scores; see Table 4.9 for cell means).

Table 4.9

Pre- and Post-Manipulation Guilt Scores (Study 4.4)

	Distal relationship			Close relationship		
	Pre	Post	*p*	Pre	Post	*p*
Condition	*M (SD)*	*M (SD)*		*M (SD)*	*M (SD)*	
Control	5.20 (1.92)	2.03 (1.33)	<.001	5.48 (1.81)	3.11 (1.76)	<.001
Self-punishment	5.62 (1.45)	3.03 (1.70)	<.001	5.18 (1.70)	2.43 (1.77)	<.001

[4] Adding the two control variables did not change the results for any of the outcome variables; thus, we report the unadjusted results.

The role of traits in self-punishment effects. The acknowledgement of a moral need may be less driven by objective features of a transgression context, but rather, by traits that make individuals more or less likely to acknowledge a moral need. First we examined whether perpetrator sensitivity was related to more moral need. Indeed, perpetrator justice sensitivity was positively correlated with perceived transgression severity ($r = .26$ $p = .01$), guilt ($r = .27$, $p = .003$), shame ($r = .35$, $p < .001$), and moral identity threat ($r = .26$, $p = .01$).

Next, we tested whether perpetrator sensitivity moderated any effects of self-punishment on the outcome variables. There was a significant interaction on moral engagement, $B = 0.25$, $SE = .10$, $p = .02$; the model explained 12% of the variance in moral engagement, $F(3,109) = 5.16$, $p = .002$. Consistent with predictions, self-punishment reduced moral engagement for participants low in perpetrator sensitivity, $B = -0.33$, $SE = .13$, $p = .01$, while those who were more likely to see themselves as a perpetrator of injustice maintained their moral engagement after their self-punishment, $B = 0.12$, $SE = .13$, $p = .35$ (see Figure 4.5). Similar trends were observed on guilt, responsibility, and commitment to values, but these did not reach statistical significance (see Appendix C for non-significant findings).

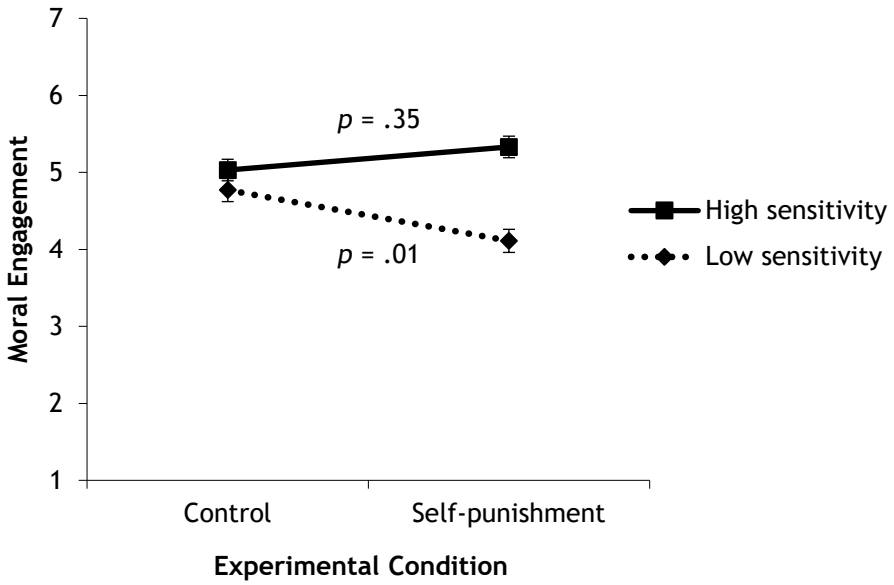


Figure 4.5. Simple slopes with standard error bars, illustrating the interaction effect between self-punishment and perpetrator sensitivity on moral engagement (Study 4.4), showing that self-punishment reduced moral engagement only for those low on perpetrator sensitivity.

Discussion

There were neither main effects of self-punishment, nor the predicted interaction by relationship closeness. However, the relationship manipulation appeared to be unsuccessful in terms of manipulating moral need. Though we successfully manipulated relationship closeness, and transgressions against close others were perceived to be more serious, this did not result in more identity threat or emotional markers of threat. Thus, while we did not observe any evidence for our model, the manipulation precluded a meaningful test of the causal role of moral need acknowledgment in determining self-punishment processes.

On the other hand, perpetrator sensitivity—the degree to which one tends to see oneself as the perpetrator of wrongdoing—was related to moral need variables. That is, those high in perpetrator sensitivity were more likely to feel guilt, shame, and perceive a threat to their moral identity. This trait then influenced the effect of self-punishment: For those low in perpetrator sensitivity, self-punishment reduced moral engagement, while it did not inhibit

engagement for those more likely to see themselves as a perpetrator of injustice. Individual differences might have thus introduced additional noise to the data that could have prevented the predicted effects to emerge. However, given this trait only moderated a single effect of self-punishment (on moral engagement), it is not clear whether this trait, at least, plays a critical role in self-punishment.

General Discussion

The present research contributes to the literature on self-punishment, introducing moral identity as a novel and integral way to understand self-punishment processes. Our findings provide evidence that self-punishment can resolve the threat to moral identity in two ways, by either evading or confronting the threat to one's moral identity. The extent to which transgressors acknowledged a moral need—acknowledging the violation and accepting the need to engage in a critical self-examination—moderated which process was activated. When transgressors failed to acknowledge a moral need, they simply cleansed themselves of their indiscretions, disengaging themselves from responsibility and concern for reparation. In contrast, when individuals acknowledged a moral need, then self-punishment promoted a more proactive and approach-oriented process whereby transgressors critically considered the implications of their wrongdoing for themselves and those they have hurt.

Contributions and Implications

The present research complements and expands on existing findings in the self-punishment and related literature. The divergent effects of self-punishment we have defined as moral repair and moral cleansing can make sense of conflicting research showing self-punishment appears to sometimes reduce and at other times promote psychological distress (Bastian et al., 2011; Fisher & Exline, 2006; Inbar et al., 2013; Whelton & Greenberg, 2005). We add to the literature by demonstrating that self-punishment has two faces: one of moral cleansing (letting oneself off the hook) and one of moral repair (confronting the threat to

moral identity). Self-punishment can thus be considered a “defend” or a “repair” behaviour, depending on the extent to which transgressors acknowledge a moral need. So far, the latter function has not been articulated in the self-punishment literature. Moreover, we extend the “defend” perspective by demonstrating that self-punishment not only reduces guilt, but can be understood as a more complex process which has consequences for moral engagement and victim reconciliation.

Acknowledgment of a moral need was key to the nature of self-punishment effects. We found evidence that more threatening transgressions (i.e., more serious transgressions against close relationships that evoke more moral threat and guilt or shame) were more likely to result in increased guilt, moral engagement, and regard for the victim. This draws parallels with Tetlock et al.’s (2000) work on sacred values—which the authors argue cannot be traded off with other values. Nevertheless, we could not prove that the relationship between transgression factors (severity and relationship importance) and moral need acknowledgment was causal. Rather, our results suggest that moral threat, and moral need acknowledgment itself, might be driven by factors outside of the transgression context—or those that we failed to account for and measure here. In line with previous findings (van Bunderen & Bastian, 2014), we found some evidence that trait-level factors could play a role in determining self-punishment outcomes; that is, people more sensitive to seeing themselves as a perpetrator of injustice were more likely to perceive moral threats as meaningful. Correspondingly, these individuals were more likely to employ self-punishment as moral repair than as a strategy to cleanse themselves of their indiscretions. Thus, the tendency for self-punishment to cleanse or repair moral identity may partly be an individual difference.

Past research has found that self-punishment can reduce guilt (Bastian et al., 2011), which is consistent with a moral cleansing function. We failed to replicate this finding on average (i.e., as a main effect of self-punishment) using an almost identical ice bucket task

(Study 4.3), or using a novel auditory self-punishment paradigm (Studies 4.1, 4.2, and 4.4). We suggest that physical self-punishment tasks—especially where endurance is emphasised—might bias effects towards one of the two pathways. Specifically, such tasks may encourage an affirmation of alternate, transgression-irrelevant qualities (e.g., toughness), and are thus more likely to lead to moral cleansing. By minimising the performance aspects of this task, we overall did not observe effects consistent with a moral cleansing hypothesis, and instead may have offered an opportunity for both pathways to be employed. This suggests that methodological factors can influence self-punishment processes, which researchers should consider when conducting self-punishment research in the laboratory.

The present research also has significant clinical implications. Though self-punishment can take on benign forms, it may also manifest itself in more harmful or destructive ways: self-harm, eating disorders, and substance abuse. Using insights from the present research, clinicians could attempt to diagnose individuals' motivation for self-punishing. By understanding why individuals are punishing themselves, clinicians have the opportunity to meet self-punishers' psychological needs through other means. If self-punishers are motivated to confront their wrongdoing, then they may benefit from learning alternate strategies to achieve moral repair. For example, transgressors might be encouraged and supported to apologise or provide reparation to their victims. Alternatively, Woodyatt and Wenzel (2014) have developed a *value affirmation task* that increases self-trust, self-forgiveness, and reconciliation, while facilitating genuine moral engagement with one's wrongdoing. In this simple task, transgressors are asked to write what value they feel they violated by their transgression, why this value is important to them, and a time in the past they have behaved consistently with this value. Encouraging transgressors to acknowledge their moral failing in a constructive manner such as this might reduce the desire to engage in more harmful self-punitive behaviours.

There is reason to believe that those self-punishing as avoidance might not always be able to easily "cleanse" themselves of their wrongdoing. Research reveals that self-harmers might become "addicted" to self-harm by the short-term emotional relief it provides; yet, after some time the unresolved conflict can resurface (Brown et al., 2007; Chapman et al., 2006; Wadman et al., 2016). This is in line with research indicating that unresolved shame and unacknowledged guilt can be problematic (P. Gilbert, 2000; Lewis, 1971; Scheff, 1994), as well as broader psychological research on thought suppression and emotional avoidance (Abramowitz et al., 2001). For individuals trying to hastily excuse their wrongdoing, self-punishment may similarly "short-cut" guilt or shame processing, denying perpetrators the opportunity to resolve the moral crisis and resulting in distress later down the track. In these cases, moral cleansing is not a viable long-term strategy. Instead, therapists could help clients shift to a moral repair mode by exploring and resolving feelings of guilt. This would involve encouraging an awareness of the transgression, an acknowledgement of responsibility, and the need to make amends (Narramore, 2002).

Limitations and Future Directions

Clarification of the construct of moral need is needed to better understand the specific emotions or cognitions that prompt moral repair and moral cleansing. Notably, moral identity threat was not correlated with the other moral need variables. Admittedly, each of the moral need variables were proposed to reflect a different aspect of a acknowledged moral need. Severity and relationship importance are features of the transgression that on average might lead to a more significant moral need; the identity threat measure reflects a more conscious perception of the need; and guilt and shame are emotional markers of the perceived moral need. Thus, though it might be expected that these variables be positively related, the strength of the associations may vary. Moreover, the operationalisation and wording of the items differed: The identity threat items were more general states ("I feel that I *am* a moral

person"), while appraisals of severity and relationship importance were necessarily tied to the transgression. Self-reported guilt and shame measures asked about one's current stated but also referenced the transgression ("*When I think about what I have done* I feel guilty"). Therefore it is not entirely surprising that the moral need measures were not strongly related.

Yet, it is unclear which of these variables best capture the acknowledgment of a moral need and are consequently responsible for the differentiated effect of self-punishment. The importance of each variable as a moderator was inconsistent across Studies 4.1, 4.2, and 4.3 (e.g., in Study 4.2, moral identity threat did not moderate any effects of self-punishment on the outcome variables). Echoing these irregularities, the theoretical literature is divided about which aspect of a moral threat is most important in understanding responses to immoral behaviour. Leary (2004) claims that *emotion* is more instrumental than cognitive measures of identity in motivating responses to transgressions. In contrast, Schnabel and Nadler (2008) argue that while emotions might be suppressed or downplayed, transgressors are most interested in remedying the threat to their *public* moral image (the perceptions that they believe others have of them), which might be relatively independent from guilt. Though we argue that the experience of feeling like a bad person or group member is largely internalised (concurring with J. L. Tracy & Robins, 2004)—after all, one can feel guilty even when nobody knows one has misbehaved—perhaps we could have included some items that target beliefs about others' perceptions, real or imagined (e.g., "what would others think about you *if they knew* about your wrongdoing?").

Alternatively, there may be more reliable measures of moral need that do not rely on self-report. For example, Cramwinckel, van Dijk, Scheepers, and van den Bos (2013) used cardiovascular indicators to measure moral threat. Interestingly, these measures did not correlate with the self-report measure of threat. Researchers could thus further investigate the relationship between implicit and explicit measures of threat and how these relate to

motivations for self-punishment. Moreover, it could be that as the transgression or self-punishment context is varied, different threat variables play more or less a role in participants' motivations for self-punishment. Future research could develop theory predicting which dimensions should come into play and when, in order to produce more consistent findings.

Given there was only a single moderated effect of perpetrator sensitivity on self-punishment effects, the role of traits in determining moral need is inconclusive. However, there are some potential avenues to test this relationship further. Given that our model focused on the *acknowledgment* of a moral need, we employed perpetrator justice sensitivity. Yet, perhaps the *denial* of a moral need is more influential here. Van Bunderen and Bastian (2014) found that victim justice sensitivity determined whether self-punishment inhibited or promoted victim compensation. Indeed, the definition given to the victim sensitivity subscale by the original authors echoes our formulation of processes involved in moral cleansing: "Victim sensitivity might […] reflect people's alertness to deprivation and their willingness to employ self-protective and egoistic interpersonal strategies" (Schmitt, Gollwitzer, Maes, & Arbach, 2005, p. 206). Further work needs to be done on elaborating the concept and measurement of moral need in the context of self-punishment, including exploration of both ends of the spectrum (acknowledgment versus denial) and measurement level (trait versus state).

A final qualification is the limited statistical power to detect small effects. Sample sizes for these studies were restricted by practical issues, which may have prevented detection of real effects. In particular, the experimental tasks were resource intensive. The ice bucket task (Study 4.3) required the experimenter to time how long participants left their hand in the water; moreover, to minimise the performance element to this task, each participant completed the task without an audience. At best, this meant participation was staggered, but

in reality it involved running one participant at a time. The data may have contained smaller yet meaningful effects that were not detected due to low statistical power (e.g., the non significant but consistent moderated effects by perpetrator sensitivity). Current findings could be tested using less resource-intensive tasks, such as those embedded in online surveys (e.g., see Chapter 3).

Conclusion

Conflicting views of self-punishment as either an expression of one's guilt, or as an evasion of it, both hold some truth. Far from being a simple emotion regulation tool, individuals use self-punishment to fulfil complex psychological needs. Researchers and clinicians should consider both faces of self-punishment to better understand and serve those who engage in self-punishment.

CHAPTER 5: Suffering for Justice: Self-Punishment and Third Party Forgiveness

> Clothed in long gowns reaching from head to foot, with no part of the face visible save the eyes, they paraded the streets, … chanting lustily the mournful verses of the "Miserere." To express the idea of sorrow for sin more forcibly, each penitent was provided with a whip well knotted or furnished with metal points, by means of which he lashed the exposed back and shoulders of the brother whom he followed. It was a weird but loathsome spectacle, from which sensible men turned away with mingled shame and indignation. But Henry of Valois was both interested and pleased. … If it could atone for moral delinquencies, the pain endured would be a cheap price to pay for the purchase of absolution. (Baird, 1896/2004, p. 38)

Wandering from village to village and whipping themselves publically, the self-flagellants of the Middle Ages sought divine redemption through their brutal expressions of self-punishment (E. Beard, 2013). Modern observers have similarly proposed that one can gain forgiveness from (more earthly) others through self-punishment. One scholar has argued that self-punishment is a communication of remorse to victims that helps to mend the relationship broken by a transgression (Nelissen, 2012), while others have suggested that it can redeem one's tarnished reputation in the eyes of one's social group more broadly (Tanaka et al., 2016; Tanaka et al., 2015; Zhu et al., 2017). Whoever the exact audience may be, these accounts suggest that self-punishment has an interpersonal goal beyond any grappling with one's moral identity that may be going on internally. We draw on a social-psychological model of punishment and justice in order to understand how self-punishment might achieve redemption in an onlooker's eyes.

Punishing Transgressors Restores Justice

Moral transgressions elicit moral outrage and anger (Batson et al., 2007; Haidt, 2003), motivating both victims and third parties to punish the perpetrator (Darley & Pittman, 2003;

Fehr & Gächter, 2002; Van Prooijen, 2010; Vidmar, 2000; Wenzel & Okimoto, 2016). Punishing others can help restore justice and ease the desire for retaliation (Goldberg, Lerner, & Tetlock, 1999), even when the perpetrator suffers harm through mere misfortune (Austin, 1979; Austin, Walster, & Utne, 1976). Punishment is motivated in part because it can address the *symbolic implications* of injustice (Okimoto & Wenzel, 2008): Punishment (1) re-balances status/power; and (2) re-establishes the legitimacy of the shared moral values violated by a moral transgression.

First, punishment can redress the status and power that moral transgressors have usurped from both the victim and society at large. Punishment *derogates* the perpetrator's status/power through humiliation and deprivation—constituting the retaliatory logic that underlies retributive modes of punishment (Wenzel, Okimoto, Feather, & Platow, 2008). The sentiment of retributive justice can be found in the adage "an eye for an eye" and notions of deservingness (Darley, 2002). Consequently, punishing the transgressor restores social equity and reassures everybody that the world is just, as bad things come to those who threaten the social order (Lerner & Miller, 1978).

Since it is the victim who has suffered the greatest loss of status/power as a result of a moral transgression, restoring status/power is most important to victims (Shnabel & Nadler, 2008). Nevertheless, some research indicates that third parties can be concerned with addressing victim needs (Chavez & Bicchieri, 2013; Gromet, Okimoto, Wenzel, & Darley, 2012; Lotz, Okimoto, Schlösser, & Fetchenhauer, 2011) and might be invested in restoring the victim's status/power to some degree.[1] The extent to which an observer is interested in redressing victim needs is likely to be a function of the closeness and similarity of the victim to the observer (Loewenstein & Small, 2007; Van Prooijen, 2010).

[1] Some scholars have argued that a third party's response to mistreatment is similar to the victim's own reaction, only less intense. However, some evidence suggests that their justice motives can diverge (see Skarlicki & Kulik, 2004).

Second, condemning perpetrators through an act of punishment can re-establish the legitimacy of the values violated by the transgression (Okimoto & Wenzel, 2009). This symbolism extends the meaning of justice restoration to a forward-looking, consequentialist view: By communicating the legitimacy of the violated values to both the perpetrator and to others, punishment motivates people to comply with the norms that are necessary for group cooperation (Funk, McGeer, & Gollwitzer, 2014; Morris, 1981). Since the violation of values represents a threat to everyone within the group who shares those values, establishing a value consensus through punishment is of importance to the wider group as well as the victim—going some way to explain why third parties punish perpetrators even when there appears to be no direct reward for the uninvolved party (Fehr & Fischbacher, 2004; Krasnow, Delton, Cosmides, & Tooby, 2016).

Furthermore, there are indications that being the punisher has some unique benefits for third parties. Third party punishment can function as a social signal to maintain the third party's reputation (Kurzban, DeScioli, & O'Brien, 2007). That is, punishing transgressors demonstrates a third party's commitment to shared values and concern for the victim, which leads others to perceive that he or she is trustworthy, group-focused and worthy of respect (Barclay, 2006). Restoring justice through punishment can thus boost third parties' social status, and is accompanied by a psychological reward of feeling morally just (Adams, 2011). On the other hand, third party punishment comes at a cost, as it risks inciting conflict. Indeed, unlike victims, third parties avoid punishing transgressors when they can bolster their social image in another way (Kriss, Weber, & Xiao, 2016).

How Self-Punishment Might Address Symbolic Concerns

When transgressors take it upon themselves to self-punish, does this satisfy us to the same extent as when others inflict the punishment? One recent study provides some evidence in favour of this proposition, at least from the victim's perspective: Following hypothetical

interpersonal transgressions, some victims believed they would be more likely to forgive transgressors when transgressors made self-critical statements rather than self-compassionate statements (Allen, Barton, & Stevenson, 2015). Considering self-punishment within the symbolic injustice framework may explain any main effect of self-punishment on third party forgiveness.

Self-punishment might redress the status/power balance upset by the transgression by repaying the "suffering debt" that is owed as a result of the victim's suffering—just as when others do the punishing on their behalf. In fact, self-criticism has been similarly conceptualised as a submissive cue that can signal to others that an individual is not a threat to the social ranking order (Sloman et al., 1994; Zuroff et al., 1999). Self-punishment may demonstrate that one does not feel superior to others, taking one's rightful place in the social hierarchy by degrading oneself.

However, there is some doubt about whether onlookers would feel that status/power has truly been restored by self-punishment. It may reduce the offender's status/power, but it may not be as effective at increasing the victim's. Though self-punishment might go some way towards affirming the victim's status through a sign of respect, it is perhaps at best an indirect strategy to restore victim status/power—to which third parties may be sensitive. Self-punishment removes the opportunity for victim empowerment through being the one to choose and/or inflict the penalty for wrongdoing. Victims appear to prefer personally delivering revenge, and having a say about the fate of the offender (Gollwitzer, Meder, & Schmitt, 2011; Gollwitzer et al., 2014; Orth, 2003; Strelan, Di Fiore, & Van Prooijen, 2017). Self-punishment may also prevent a restoration of victim status/power in more direct ways such as receiving compensation or an apology (Darley & Pittman, 2003; Okimoto & Wenzel, 2008).

Self-punishment might be relatively more effective at establishing a value consensus. In fact, self-punishment may do so even better than punishment inflicted by others. Externally-imposed punishments can only *dispatch* a message to transgressors that the act was wrong and that they should reform; it does not guarantee that the transgressor truly accepts and internalises this message. Self-punishment, on the other hand, appears to provide this assurance, since the act of punishing oneself implies a denunciation of the immoral act and thus an endorsement of shared values. In this case, observers may be convinced that the transgressor holds the values necessary for group membership and welcome him or her back to the moral community. Of course, when transgressors take it upon themselves to self-punish, third parties do not have the opportunity to demonstrate their own commitment to group values by punishing the transgressor (as suggested by Kurzban et al., 2007). Thus, there is perhaps a missed opportunity for others to affirm shared values. Yet, the person who has perpetrated the wrongdoing is perhaps better positioned to revalidate the values and firmly re-establish the legitimacy of the social order.

In summary, despite some indications that self-punishment can restore the transgressor's public moral image, there are some reasons to question the strength of these effects, specifically in regards to restoring status/power. If third parties are interested in attending to victim needs, they may view self-punishment as undercutting the opportunity for victim empowerment.

Perceived Motivation Matters For Judgments of Sincerity and Forgiveness

A "successful" display of self-punishment can thus be defined as one through which the perpetrator communicates a commitment to values or a restoration of status/power, ultimately earning them forgiveness and acceptance from others. However, third parties could make a number of alternate attributions about the self-punisher's motives that might reduce the success of this strategy. The perpetrator's display might be interpreted as *insincere*—as

not coming from a genuine sense of remorse, which is critical in forgiveness following apologies and confessions (Hareli & Eisikovits, 2006; Schumann, 2012; Weiner, Graham, Peter, & Zmuidinas, 1991). Though costly displays of remorse are generally seen as more sincere than non-costly acts (Ohtsubo et al., 2012), there may still be some factors that influence perceived sincerity of a self-punishment. It is difficult to theorise on every possible feature of a self-punishment that would lead to attributions of (in)sincerity; nevertheless, we offer a couple of candidates.

Observers may doubt self-punishers' sincerity if the suffering is not considered severe enough. The level of pain that is required to meet this threshold may depend on the severity of the initial transgression: A half-hearted display of self-punishment for a serious harm denies full restoration of lost status/power and might call the self-punisher's commitment to values into question. In these cases, observers may perceive a lack of true suffering, or even that transgressors are gaining some pleasure from their behaviour. Rather, the self-punishment may be perceived as an act that serves the punisher more than others—perhaps being used to "cleanse" one's moral conscience, to distract from one's guilt and to avoid the harder work of taking genuine responsibility (i.e., it may be interpreted as an act of moral cleansing rather than moral repair). Thus, an underweight display of self-punishment may convey the impression that the perpetrator is not dealing with the transgression in a sincere manner in line with genuine remorse and moral repair.

Nor is a transgressor safe from a charge of *overdoing* self-punishment. Pain displays that are judged as being disproportionate or excessive relative to a threat can result in rejection by peers by undermining judgments of sincerity (Boothby, Thorn, Overduin, & Ward, 2004). Similarly, continuous displays of pain endurance (e.g., chronic sufferers of pain) have been shown to fatigue an empathic response over time, as observers begin to doubt the sufferer's credibility (Craig, 2009). Furthermore, dyad studies of relationship partners

navigating conflicts have indicated that trait self-criticism (implying frequent self-criticism) may result in negative feedback from partners (Vettese & Mongrain, 2000), and excessive self-blame can damage relationships (Pelucchi, Paleari, Regalia, & Fincham, 2013). Proportionality is therefore possibly important in obtaining forgiveness through self-punishment; excessive or ongoing self-punishment may be counterproductive.

Under the symbolic injustice framework, any backfire effects of excessive self-punishment may operate on two levels. First and foremost, excessive self-punishment may call the self-punisher's values into question: Observers may feel that the self-punisher's focus is becoming increasingly self-focused, and this mismatch gives the impression that the perpetrator is not truly considering the violated values, but rather some other hidden agenda. For instance, excessive self-punishment could suggest that the perpetrator has become fixated on their own torment rather than thinking through what they did wrong. Alternatively, self-punishers may be accused of being unable to realistically assess and understand the values they have violated. Not only is the violated value important, but also the accompanying expectations around how to respond to it appropriately, and others' beliefs about what a failure to follow these expectations suggests about the perpetrator (Worthington & Wade, 1999). By punishing oneself to an extreme degree, self-punishers may inadvertently elicit suspicion and further social ostracism. Therefore, in such cases we might expect self-punishment to *diminish* perceptions of a value consensus.

Second, excessive self-punishment could also influence perceptions of sincerity, though this link is arguably less clear than between sincerity and value consensus. An argument could be made that excessive self-punishment could undermine any positive effects of lowering one's status by tipping the balance too far in the other direction, perhaps implying that the transgressor now needs our help. Ongoing self-criticism can provoke irritation and anger by overwhelming demands on others (Gergen & Wishnov, 1965; Platt,

1977; Potthoff, Holahan, & Joiner, 1995). In the context of transgressions in particular, this may provoke irritation; it may be too much to demand care and protection for one who has so recently transgressed. Instead, the self-punisher may be rejected from one's group in order to reduce obligations towards him or her. Thus, in attempting to re-balance status/power, one has to tread a fine line between degrading oneself as a show of humility, and turning oneself into a victim.

In addition to self-punishment excessiveness, contextual factors might influence perceptions of sincerity. Nelissen and Zeelenberg (2009) argued that self-punishment is only employed when it is not possible to compensate for the harm done directly—a position that may be extrapolated to suggest that self-punishers may not be well received if such alternatives exist. Observers may regard self-punishment with suspicion when it is at the expense of victim repair. In particular, there are times when third parties are interested in more typically restorative measures (i.e., those that directly attend to the victim's needs) rather than (or as well as) retributive ones (Gromet, 2009).

In these cases, observers may become angered by what they may perceive as a self-interested action of self-punishment—especially if it is not backed up by other restorative behaviour. Such an act might be interpreted as avoidant, suggesting that the transgressor is too cowardly to face one's wrongdoing and engage with the victim (i.e., the self-punisher's motive might be perceived as being a desire for moral cleansing rather than for moral repair). These attributions may lead observers to perceive that the transgressor has a hidden agenda for their behaviour, or has violated expectations about how to appropriately respond to the wrongdoing. Thus, such interpretations again appear to have a potentially strong bearing on perceptions of value consensus. In addition, self-punishing at the expense of repair may be seen as inappropriate as it denies the victim the power to accept or reject the perpetrator (Okimoto & Wenzel, 2008). Therefore, self-punishing *at the cost of* other expected responses

may lead others to question self-punishers' motivation, and consequently rob self-punishers of the forgiveness and acceptance they seek.

Overview

We predict that when self-punishment is considered a sincere and proportional response to a wrongdoing, it will communicate self-punishers' recommitment to violated values and relinquishment of their own status/power. However, self-punishment will be more limited in its ability to elevate victim status/power, which may be a concern for third parties in some circumstances. Overall, in so far as observers perceive that self-punishment communicates a value consensus and re-balances status/power, they will be more willing to reconcile with the perpetrator (see Figure 5.1).

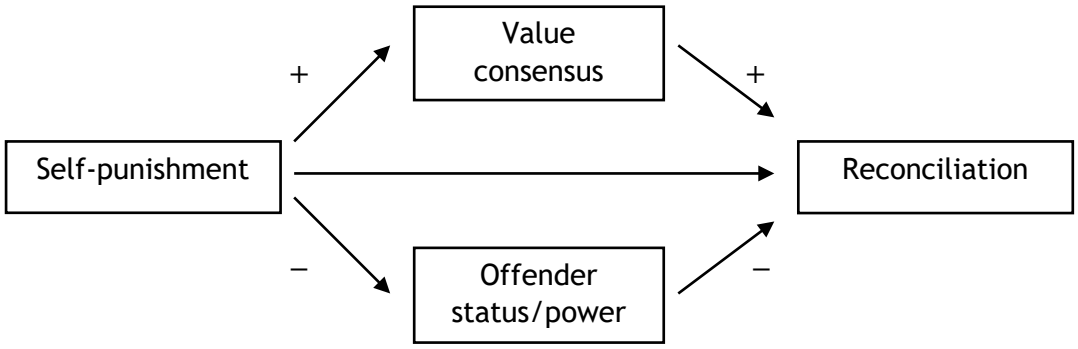


Figure 5.1. Model of third party perceptions of self-punishment: Self-punishment promotes interpersonal reconciliation via a perceived value consensus and status/power.

In Study 5.1, we examine the effect of self-punishment on notions of justice, forgiveness, and social/moral standing from a third party's perspective. We test the hypothesis that self-punishment facilitates interpersonal reconciliation, mediated by an increase in value consensus and reduced offender status/power. We expand the design in Studies 5.2 and 5.3, testing whether excessive self-punishment reduces the perceived sincerity of the behaviour, and thus, reduces its ability to restore the perpetrator's public moral image.

Study 5.1

The purpose of the first study was to provide preliminary evidence that transgressors engaging in self-punishment could earn third party observers' forgiveness and social acceptance, via value consensus and status/power. Further, we predicted that self-punishment would have a stronger effect on value consensus, and thus reconciliation, relative to punishment inflicted by others.

Method

Sample. We based our estimated effect sizes for main effects of self-punishment on previous studies examining third party forgiveness of transgressors following apologies to the victim (Green, Burnette, & Davis, 2008), as well as hypothetical scenario studies of transgressor self-criticism (Allen et al., 2015) and third party punishment (Strelan et al., 2017) on victim perceptions of forgiveness and justice satisfaction. All effect sizes were medium to large.[2] Thus, conservatively we aimed for a medium effect size. G*Power 3 (Faul et al., 2007) suggested a sample size of 51 per condition for 80% power to detect a medium effect of $d = 0.50$ at an alpha level of .05, which provided a guide for this and the following studies.

Members of the Australian public ($N = 150$) were recruited through an online research participation pool and paid a small monetary reward for their participation ($1.50). The sample was 51% female and 49% male, mean age 30.75 years ($SD = 10.78$).

Procedure. Participants read a vignette describing a hypothetical interpersonal transgression in which they were asked to imagine they were a mutual friend of both the perpetrator and victim. They were then randomly assigned to one of three experimental conditions. In the control condition, the perpetrator was described as not having been

[2] Green et al. (2008) report a main effect of apology (to the victim) on third party forgiveness equivalent to a Cohen's *d* of 0.76. Allen et al. (2015) report a partial η^2 of .04 for the difference between self-critical and self-compassionate transgressor statements on forgiveness (Study 2). Strelan et al. (2017) find Cohen's *d*s of 0.42 1.17, and 1.29 for the effect of third party punishment on forgiveness (in Studies 1, 2, and 3, respectively), and *d*s of 0.93, 2.09, and 3.11 on justice satisfaction (in Studies 1, 2, and 3, respectively).

punished. In the self-punishment condition, the perpetrator self-punishes. In a third condition, the victim punishes the perpetrator. This last condition offered an additional point of contrast in order to pull apart the effects of the self-inflicted nature of self-punishment relative to punishment in general.

The nature of the punishment chosen was social isolation, as it could be easily manipulated to be inflicted both by the victim and the perpetrator without influencing the severity of the act (e.g., banging one's head against a wall seems fairly innocuous but for someone to bang another's head against a wall is far more extreme). We excluded self-harm because we judged that participants might have alternate preconceptions about its meaning rather than seeing it as a self-punishment (e.g., as a symptom of clinical psychopathology).

Following the vignette and manipulation, participants completed the outcome measures including justice variables, interpersonal reconciliation variables, and manipulation checks. Last, we explored participants' interpretations of the motives underling the self-punishment.

Materials.

Vignettes. Vignettes described a short story between a transgressor and a victim. The characters' gender was matched to the participant's gender to ensure the story was as relatable as possible. The male version is presented here for illustration:

> *Please imagine that you are friends with two men called Louis and Daniel. You share a similarly close friendship with both men. The following passage contains an imaginary scenario involving these friends. What follows next is your account of the events that happened. Try and immerse yourself in the situation. Try to think of what you would think and feel.*
>
> *Louis and Daniel are 25 year old men who have been close friends for several years. They are currently housemates, having lived together for some time. Unbeknownst to*

Daniel, lately Louis has been secretly stealing from Daniel, taking cash from Daniel's wallet when he leaves it unattended. Over the course of several months, Louis manages to steal about three hundred dollars from Daniel.

Additional information was then provided as the experimental punishment manipulation. Participants in the control condition (n = 48) received the following information: "Louis stops stealing money from Daniel. Over the next few weeks, Louis goes on living much like he did before anything had happened." The self-punishment condition (n = 55) read: "Louis stops stealing money from Daniel. Over the next few weeks, Louis isolates himself, not allowing himself to go out or see his friends." Those in the victim punishment condition (n = 47) were presented with: "Louis stops stealing money from Daniel. Over the next few weeks, Daniel isolates Louis by not inviting him out with their mutual friends."

Outcome measures. All outcome measures were measured on a 7-point rating scale. As this was a preliminary study, some of the scales were shortened from their original format so that we could accommodate a wider range of measures, with the intention of expanding them in subsequent studies were they to prove useful.[3] Scale reliabilities were calculated using Cronbach's alpha, except for scales with three or fewer items (since scales with few items can bias Cronbach's alpha below the normally accepted standard of .7; Cortina, 1993).

Justice variables. A value consensus scale was based on Wenzel and Okimoto (2010) and comprised four items: "I feel Louis accepts moral values widely shared in our community"; "I feel Louis ignores a broadly accepted understanding of what is right and wrong" (reverse coded); "I feel Louis embraces commonly shared beliefs and values"; "I feel Louis and I would agree on principles of decent conduct" (α = .71).

[3] Three additional measures included in the initial battery are not reported here: victim satisfaction (Gromet et al., 2012), agency/patiency (based on Khamitov, Rotman, & Piazza 2016), and revenge (based on McCullough et al 1998). These measures did not yield significant results in the analyses and were subsequently dropped from the survey after Study 5.1.

Two status/power scales were constructed: one assessing perceived victim status/power and one assessing offender status/power (Wenzel & Okimoto, 2010, 2015). The offender scale included four items: “Louis feels superior to others”; “Louis feels he can dominate others”; “Louis feels he has lost others’ respect” (reverse coded); “Louis feels weak relative to others “(reverse coded) (α = .69). The victim scale was also made up of four items: “Daniel feels equal to others”; “Daniel feels he does not let himself be pushed around”; “Daniel feels he has lost others’ respect” (reverse coded); “Daniel feels weak relative to others” (reverse coded). The victim status/power scale was not used in the analyses as it showed poor internal reliability (α = .43).

Justice satisfaction was measured using four items based on Wenzel and Okimoto (2014): “Given everything that has happened, is your sense of fairness satisfied?”; “Are you satisfied with the way the problem was resolved?”; “Do you feel the situation as it now stands is unfair?” (reverse coded); “Considering the events, do you feel a strong sense of injustice?” (reverse coded) (α = .74).

Interpersonal reconciliation. We used four measures to assess observers’ interest in reconciliation with the transgressor, including remorse, trust, moral standing, and social liking/rejection.

Observer perceptions of the transgressor’s remorse were measured with three items: “Louis has expressed remorse about his wrongdoing”; “Louis isn’t really ashamed of his wrongdoing” (reverse coded); “Louis’ actions show that he has taken responsibility for his wrongdoing.” The average item intercorrelation was .40, suggesting good internal reliability (i.e., above .2, Briggs & Cheek, 1986; Cortina , 1993).

Forgiveness was measured using two items that tapped both into one’s personal sense of forgiveness and also whether the victim should forgive the transgressor: “You forgive

Louis for what he did to Daniel" (used in a third party context in Green et al., 2008); "Daniel should forgive Louis for his wrongdoing" ($r = .69$, $p < .001$).

Three items were used to assess trust: "I think that Louis will be a good friend to me in the future"; "Generally speaking, Louis is trustworthy"; "I think Louis would do the same thing again if he had the chance" (average item intercorrelation = .57).

Moral standing was measured using three items from Piazza, Landy, and Goodwin's (2014) scale: "Harming Louis would be morally wrong"; "Louis deserves to be treated with care and compassion"; "I have sympathy for Louis" (average item intercorrelation = .42).

Social liking/rejection was comprised of four items from Coyne (1976): "Would you admit Louis back into your circle of friends?"; "Would you be willing to work with Louis on a job?"; "Would you like to meet Louis in real life?"; "Would you sit next to Louis on a 3-hour bus trip?" ($\alpha = .88$).

Manipulation checks. Two manipulation check items were used to distinguish the experimental conditions from each other: "In the scenario you were given, Louis isolated himself from his friends"; "In the scenario you were given, Daniel isolated Louis from his friends."

Self-punishment interpretation. An additional item was only presented to those in the self-punishment condition, asking participants: "Which of the following describe why you think Louis isolated himself?" Four choices tapped into genuine moral repair and four choices tapped into avoidance or moral cleansing, respectively (choices were randomised): "To better understand his behaviour"; "To learn from what he did"; "To make amends"; "To earn Daniel's forgiveness"; "To run away from his problems"; "To make himself feel better"; "To gain others' pity"; "To avoid Daniel's wrath." Participants could either check or not check as many of these choices as they liked.

Results

Manipulation checks indicated that the manipulation was successful. A one-way analysis of variance (ANOVA) indicated that there was an overall condition effect on the belief that the perpetrator had self-punished, *Welch's F*(2, 66.49) = 235.52, $p < .001$.[4] Planned contrasts revealed that the mean for the self-punishment condition (M =6.87, SD = 0.39) was significantly higher than that of the control (M = 3.33, SD = 1.96) and victim punishment (M = 2.09, SD = 1.74) conditions, as expected (both ps < .001). The second manipulation check item also showed a significant effect of the experimental manipulation, *Welch's F*(2, 92.74) = 689.93, $p < .001$. Planned contrasts confirmed that the effects were as expected: Participants in the victim punishment condition (M = 6.64, SD = 0.76) were more likely to believe that the perpetrator was punished by the victim, relative to both control (M = 1.77, SD = 1.15) and self-punishment (M = 1.29, SD = 0.74) conditions (both ps < .001).

Self-punishment promotes reconciliation, mediated by value consensus and status/power. To assess the effects of the manipulations, one-way ANOVAs were conducted on the dependent variables. See Table 5.1 for cell means, standard deviations and test statistics (see also Appendix B for inter-item correlations). Results indicated that the manipulation had significant effects on values, offender status/power, remorse, justice, trust, and liking. *Tukey* post hoc tests revealed that self-punishers were perceived as more remorseful, trustworthy, and likeable compared to the control condition, and self-punishment also led to greater value consensus and justice restoration, and less offender status/power. Victim-inflicted punishment also led to perpetrators being assigned less status/power, as well as to greater justice restoration relative to the control condition. However, self-punishment had stronger effects beyond victim punishment for status/power.

[4] *Welch's F* is reported where a Levene Test indicated the assumption of homogeneity of variances was violated. Corresponding contrasts/comparisons do not assume equal variances.

Table 5.1

Descriptive and Between-Group Statistics for Dependent Variables (Study 5.1)

	M (*SD*)		
Dependent variable	Control	Self-punishment	Victim punishment
Values	2.22 (1.02)	2.78 (1.04)	2.41 (1.07)
Offender status/power	4.71 (0.92)	3.45 (1.10)	4.02 (1.07)
Remorse	2.50 (1.19)	3.57 (1.08)	2.89 (1.23)
Justice satisfaction	1.89 (0.92)	2.53 (1.11)	2.71 (1.13)
Forgiveness	2.73 (1.36)	3.23 (1.42)	3.05 (1.63)
Trust	1.91 (0.87)	2.67 (1.19)	2.38 (1.18)
Moral standing	4.24 (1.31)	4.37 (1.26)	3.96 (1.32)
Liking/rejection	2.79 (1.32)	3.61 (1.43)	2.86 (1.39)

		Contrast effect size (*d*)		
Dependent variable	ANOVA *F*-test	C vs. SP	C vs. VP	SP vs. VP
Values	$F(2,147) = 3.85$*	0.54*	0.18	-0.35
Offender status/power	$F(2,147) = 19.05$**	-1.24**	-0.69*	0.53*
Remorse	$F(2,147) = 11.23$**	0.94**	0.32	-0.59*
Justice satisfaction	$F(2,147) = 8.04$**	0.63*	0.80**	0.16
Forgiveness	$F(2,147) = 1.50$	0.40	0.21	-0.12
Trust	$F(2,147) = 6.29$*	0.73*	0.45	-0.24
Moral standing	$F(2,147) = 1.27$	0.10	-0.21	-0.32
Liking/rejection	$F(2,147) = 5.71$*	0.60*	0.05	-0.53*

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

Next we tested the symbolic justice model of self-punishment, assessing whether value consensus and status/power mediated the effects of self-punishment on observers' attitudes towards the perpetrator. We ran simple mediation models on the dependent variables to assess each of the mediators separately for their mediation effects. The PROCESS macro for SPSS (Hayes, 2013; model 4) was used with 10,000 bootstraps and bias-corrected confidence intervals (adjusted to 99% CIs for multiple comparisons). Experimental condition was entered as the exogenous variable (set as a contrast between control and self-

punishment), value consensus and offender status/power as mediators, and the seven remaining outcome variables as the dependent variable in the models. See Table 5.2 for results. Analyses revealed significant indirect effects through value consensus on all dependent variables except for moral standing. There were also significant indirect effects through offender status/power on remorse and trust.

Table 5.2

Mediation Model Statistics: Indirect Effects of Self-Punishment Via Value Consensus and Status/Power (Study 5.1)

Model effects	*b*	*SE*	β	$CI_{99\%}$ (*b*)
Mediator (M): value consensus				
IV→M	0.56	0.20	.26	0.02, 1.09
DV: justice				
M→DV	0.41	0.09	.41	0.17, 0.65
IV→DV (direct effect)	0.41	0.19	.19	-0.10, 0.91
IV→M→DV (indirect effect)	0.23	0.10	.11	0.03, 0.53
DV: remorse				
M→DV	0.55	0.10	.47	0.30, 0.80
IV→DV (direct effect)	0.76	0.20	.31	0.23, 1.29
IV→M→DV (indirect effect)	0.31	0.12	.12	0.04, 0.68
DV: forgiveness				
M→DV	0.51	0.13	.38	0.18, 0.84
IV→DV (direct effect)	0.21	0.27	.08	-0.48, 0.91
IV→M→DV (indirect effect)	0.28	0.13	.10	0.03, 0.72
DV: trust				
M→DV	0.51	0.09	.49	0.28, 0.75
IV→DV (direct effect)	0.48	0.19	.21	-0.02, 0.97
IV→M→DV (indirect effect)	0.29	0.12	.26	0.03, 0.67
DV: moral standing				
M→DV	0.24	0.12	.20	-0.08, 0.56
IV→DV (direct effect)	-0.01	0.26	-.003	-0.69, 0.67
IV→M→DV (indirect effect)	0.13	0.10	.05	-0.03, 0.49

Model effects	*b*	*SE*	β	$CI_{99\%}$ (*b*)
DV: liking				
M→DV	0.57	0.12	.42	0.25, 0.89
IV→DV (direct effect)	0.50	0.26	.18	-0.17, 1.18
IV→M→DV (indirect effect)	0.32	0.14	.11	0.03, 0.75
Mediator (M): status/power				
IV→M	-1.26	0.20	-.53	-1.79, -0.73
DV: justice				
M→DV	-0.11	0.10	-.13	-0.38, 0.15
IV→DV (direct effect)	0.49	0.24	.23	-0.13, 1.12
IV→M→DV (indirect effect)	0.14	0.14	.07	-0.21, 0.54
DV: remorse				
M→DV	-0.36	0.10	-.35	-0.64, -0.09
IV→DV (direct effect)	0.61	0.25	.25	-0.05, 1.27
IV→M→DV (indirect effect)	0.46	0.15	.18	0.11, 0.88
DV: forgiveness				
M→DV	-0.39	0.13	-.33	-0.74, -0.05
IV→DV (direct effect)	0.001	0.31	.00	-0.82, 0.82
IV→M→DV (indirect effect)	0.50	0.21	.18	-0.08, 1.09
DV: trust				
M→DV	-0.34	0.10	-.37	-0.60, -0.09
IV→DV (direct effect)	0.33	0.23	.15	-0.28, 0.94
IV→M→DV (indirect effect)	0.44	0.17	.20	0.07, 0.99
DV: moral standing				
M→DV	-0.17	0.12	-.16	-0.50, 0.16
IV→DV (direct effect)	-0.09	0.30	-.04	-0.87, 0.69
IV→M→DV (indirect effect)	0.22	0.19	.09	-0.31, 0.71
DV: liking				
M→DV	-0.10	0.13	-.08	-0.45, 0.25
IV→DV (direct effect)	0.70	0.32	.24	-0.15, 1.54
IV→M→DV (indirect effect)	0.13	0.21	.04	-0.38, 0.74

Note. $N = 103$. Indirect effect estimates bootstrapped with 10,000 samples. For the DVs remorse and trust (for which both mediators are significant), the indirect effects via both values and status/power remain significant when tested simultaneously in a parallel mediation model.

Observers' interpretations of self-punishment. Frequencies for participant interpretations of self-punishment are presented in Table 5.3. Generally speaking, defensive interpretations were endorsed more often than reasons that reflect moral repair. It is interesting to note that for the most part, participants appeared to be rather cynical about self-punishers' motivation, yet, we observed (relatively large) positive main effects of self-punishment. However, the binary response scale of the interpretation variables may have concealed nuances in observers' confidence or degree of endorsement of interpretations.

Table 5.3

Endorsed Interpretations of Self-Punishment (Study 5.1)

	Endorsed		
Interpretation	*n*	%	Category
Run away from problems	48	87	Cleansing
Avoid Daniel's wrath	28	51	Cleansing
Gain pity from others	17	31	Cleansing
Feel better	16	29	Cleansing
Learn from it	10	18	Repair
Understand behaviour	9	16	Repair
Earn Daniel's forgiveness	8	15	Repair
Make amends	7	13	Repair

Interpretation items were exploratory in order to gain a preliminary sense of how observers construed the self-punisher's motives. Nevertheless, we attempted to explore the associations between these variables. In order to conduct an exploratory factor analysis with binary data, we followed Finney and DiStefano's (2013) recommendation to use the robust WLS estimation method in structural equation modelling, using Mplus SEM software (Muthén & Muthén, 1998-2012). The factor analysis revealed a 2-factor structure with some of the item loadings making sense in regards to predictions (e.g., run away, avoid wrath, gain pity, understand behaviour [-] and learn from it [-] loading on one factor); however, our

sample did not meet the sample size requirements for this analysis, and therefore the results may not be valid. Indeed, goodness of fit indices were slightly lower than accepted thresholds (RMSEA > 0.06), some of the loadings exceeded 1.00, and the software generated warnings about empty cells due to the skewed distribution of some of the variables.

Another approach some researchers have used for factor analysis with binary variables is cluster analysis (Henry, Dymnicki, Mohatt, Allen, & Kelly, 2015; Henry, Tolan, & Gorman-Smith, 2005), though this method relies to some extent on judgment about which cluster solution/s are preferred. Variables were subjected to a hierarchal cluster analysis with between-groups linkage. Results indicated that both a 2-cluster and a 5-cluster solution were reasonable. The 2-cluster solution identified that avoid wrath and run away clustered together (which are both cleansing), while the remaining variables belonged to a second cluster. The 5-cluster solution indicated that understand behaviour and learn from it clustered together (both repair); another cluster contained make amends, earn forgiveness, and feel better; and the remaining three items were each in their own clusters. Both of the clustering solutions thus validated some aspects of the theoretical factors, but this was not straightforward.

The results of the two analyses above gave us some confidence that we could proceed with the factors of moral cleansing and moral repair for exploratory purposes. Participants were given a score of zero or one for moral cleansing and moral repair, if at least one item in that set was endorsed. Moral cleansing was negatively correlated with value consensus (r_{pb} = -.32, $p < .05$), forgiveness (r_{pb} = -.28, $p < .05$), moral standing (r_{pb} = -.36, $p < .05$), and liking (r_{pb} = -.30, $p < .05$); though note that the moral cleansing variable was skewed due to most people endorsing the run away item. Moral repair was positively correlated with value consensus (r_{pb} = .32, $p < .05$), justice satisfaction (r_{pb} = .46, $p < .001$), forgiveness (r_{pb} = .27, $p < .05$), and liking (r_{pb} = .41, $p < .05$). Therefore there was some indication that interpretations of the self-punishment varied meaningfully with the outcome variables.

Discussion

Study 5.1 provided preliminary evidence for the claim that self-punishment is a useful interpersonal strategy for transgressors to gain forgiveness and social acceptance from third parties. Observers perceived self-punishers as more remorseful, trustworthy, and likeable than transgressors who did not engage in self-punishment. Furthermore, self-punishment restored justice, and many of the effects on outcome variables were partially or completely accounted for by indirect effects via value consensus and offender status/power. Interestingly, though self-punishment had a larger main effect on status/power than on value consensus, it was more often the change in value consensus that was associated with positive attitudes towards the perpetrator. In other words, third parties' attitudes towards the self-punisher were particularly swayed by the self-punisher's commitment to shared values.

Self-punishment had stronger effects than victim-inflicted punishment on several outcome variables, though going against predictions self-punishment diminished the perpetrator's status/power, but not value consensus (relative to victim punishment). To make sense of the latter finding, it may be that a transgressor's self-degradation is more humiliating from an observer's perspective than seeing the victim inflict revenge. Self-derogation could be seen as strange and ostracising, whereas a victim's punishment is perhaps more expected and consequently less degrading. We were unable to reliably measure *victim* status/power, for which one would certainly expect self-punishment to be inferior to victim punishment. On the other hand, it is puzzling that participants did not perceive that self-punishers endorsed shared values more than transgressors punished by the victim. It seems counterintuitive that being reprimanded by another would lead to more value consensus than if one initiated a punishment on one's own accord. This finding could suggest a role for negative attributions being made about the self-punishment.

Indeed, there were indications that not all participants saw the self-punishment in a positive light. Mediation analyses revealed a significant indirect effect of self-punishment on forgiveness via value consensus, but there was no significant total effect, suggesting a negative effect suppressing the positive indirect effect. We explored one possible reason for such a negative effect: the extent to which observers were cynical about the self-punisher's motivation for the behaviour. Confirming our suspicions, when self-punishment was interpreted as avoidant, it was met with more negative responses, while if the self-punisher was perceived to be confronting one's wrongdoing, they were more successful in regaining their social standing. That is, self-punishment is a more effective pathway to forgiveness when transgressors are perceived as undergoing a process of moral repair and genuinely revising their values, rather than misunderstanding or minimising the wrongdoing.

Regardless, the conflicting findings—participants endorsing cynical interpretations yet generally forgiving the self-punisher—suggests that people are torn in their response to self-punishment. Given this, there may be conditions under which negative interpretations are more likely or salient, and thus undermine the otherwise positive inferences. Study 5.2 will investigate one possible feature that may present such a condition: excessive self-punishment.

Study 5.2

The rationale for Study 5.2 was threefold. First, we sought to replicate the generally positive effects of self-punishment on observers' perceptions via value consensus and status/power (Hypothesis 1), generalised to a different form of self-punishment. Second, we tested a boundary condition for these effects, predicting that self-punishment would have a backfire effect on positive attitudes when the self-punishment was ongoing for what may be considered an excessive period of time (Hypothesis 2), as per the theoretical argument that this might make observers suspicious about the self-punisher's motives. Last, we predicted that positive effects (especially on value consensus) would emerge due to observers

perceiving the self-punishment as motivated by moral repair rather than by moral cleansing (Hypothesis 3).

Method

Participants (N = 150) were recruited online in the same fashion as described in Study 5.1; 50% female, 50% male, mean age 30.01 years (SD = 8.22). Following the procedure from Study 5.1, participants were presented with a hypothetical transgression vignette (identical to Study 5.1). The subsequent experimental conditions were also broadly in line with the previous study for the control and self-punishment condition, however, a third condition was designed to constitute the excessive self-punishment condition. Further, to test whether the Study 1 findings were not dependent on a particular expression of self-punishment we changed the type of self-punishment to overworking oneself.

For the brief self-punishment condition (n = 50), the manipulation text read: "Over the next week or two, Louis throws himself into his work (for no extra pay or recognition), neglecting self-care and working to the point of exhaustion." For the excessive self-punishment condition (n = 49), the text read: "Over the course of many months afterwards, Louis throws himself into his work (for no extra pay or recognition), neglecting self-care and working to the point of exhaustion. Six months later, he is still continually overworking himself." The control condition (n = 51) read: "Louis stops stealing money from Daniel. Over the next few weeks, Louis goes on living much like he did before anything had happened."

Outcome measures for value consensus (α = .77), offender status/power (α = .67), justice satisfaction (α = .72), remorse (average item intercorrelation = .45), forgiveness (r = .78, p < .001), and trust (average item intercorrelation = .52) were identical to Study 5.1.

The victim status/power scale was revised, rewording and removing items that appeared problematic based on analysis from Study 5.1 (as assessed via principal component analysis, inter-item reliabilities, and convergent/divergent validity with other scales). The

revised victim scale items were: “Daniel feels secure in his relationships”; “Daniel does not let himself be pushed around”; “Daniel does not feel respected by others” (reverse coded); “Daniel feels weak relative to others” (reverse coded). Unfortunately the changes did not improve reliability ($\alpha = .40$) therefore the scale was again omitted from further analyses.

Moral standing was expanded to the full five-item scale from Piazza et al. (2014). The additional items improved the average item intercorrelation from .42 to .49 ($\alpha = .83$).

The social liking/rejection scale was also expanded from four to six items, adding: “Would you approve if a close relative were married to Louis?” “Would you ask Louis for advice?” ($\alpha = .92$). The remaining two items from Coyne’s (1976) original eight-item scale were not included because these asked about sharing an apartment and inviting the transgressor to one’s house, which speaks directly to the nature of the transgression (stealing money that is left lying around one’s house). We judged this might be problematic.

Participants in the two self-punishment conditions were asked for their interpretations of the self-punishment. Items were revised based on the results of the factor and cluster analyses in Study 5.1, and also to tie items more closely to the concepts underlying moral repair and moral cleansing. Items were: “Louis is really trying to work through what he did wrong” (repair); “Louis is trying to confront and learn from his wrongdoing” (repair); “Louis’ actions are superficial and not about really dealing with what he did wrong” (cleansing); “Louis is trying to avoid and run away from his wrongdoing” (cleansing). The scale of measurement for each item was also changed from binary to a 7-point rating scale (item intercorrelation for repair $r = .78$, $p < .001$, for cleansing $r = .29$, $p = .004$[5]). Items were combined into mean repair and cleansing indices, which were inversely related ($r = -.36$, $p < .001$).

[5] Though the correlation for cleansing was relatively small, results for the two items were consistent: no significant differences between the brief and ongoing self-punishment conditions; and correlation coefficients with all other scales were in the same direction for both items (except for moral standing, with which one item was positively correlated and the other negatively related, but neither correlation was significant).

Two manipulation check items were used. As in the previous study, all participants were asked a general question about self-punishment: “In the scenario you were given, Louis overworked himself to the point of exhaustion.” In addition, the two self-punishment conditions were asked: “In the scenario you were given, Louis overworked himself for:” (7-point scale from “A week or two” to “Six months or more”).

Results

Manipulations were successful. An ANOVA indicated that there was a significant overall condition effect on the perception that the perpetrator had self-punished, *Welch's F*(2, 95.74) = 149.40, $p < .001$. Planned contrasts confirmed that participants in the brief self-punishment condition ($M = 6.44$, $SD = 0.97$) and the excessive self-punishment condition ($M = 6.49$, $SD = 1.02$) were more likely to identify that the perpetrator had self-punished than those in the control condition ($M = 2.35$, $SD = 1.53$; both ps $< .001$). Furthermore, the two self-punishment conditions differed from each other on the second manipulation check item, a *t*-test indicating that those in the excessive self-punishment condition ($M = 5.94$, $SD = 1.44$) correctly identified that the self-punishment lasted longer relative to those in the brief self-punishment condition ($M = 1.98$, $SD = 1.45$), $t(97) = -13.65$, $p < .001$.

Hypothesis 1: Self-punishment promotes reconciliation, mediated by value consensus and status/power. We conducted one-way ANOVAs across all three experimental conditions on outcome variables. See Table 5.4 for descriptive and test statistics. *Tukey* post hoc tests indicated that the two self-punishment conditions did not differ from one another on any of the outcome variables.

Table 5.4

Descriptive and Between-Group Statistics for Dependent Variables (Study 5.2)

	M (*SD*)		
Dependent variable	Control	Brief self-punishment	Excessive self-punishment
Values	3.08 (1.38)	3.01 (1.32)	3.17 (1.31)
Offender status/power	4.30 (1.28)	3.60 (1.00)	3.63 (1.01)
Justice satisfaction	2.46 (1.08)	2.46 (1.29)	2.78 (1.23)
Remorse	2.93 (1.50)	3.85 (1.44)	3.96 (1.46)
Forgiveness	3.27 (1.71)	3.48 (1.63)	3.79 (1.68)
Trust	2.69 (1.50)	2.59 (1.26)	3.14 (1.25)
Moral standing	3.57 (1.49)	3.99 (1.15)	4.32 (1.33)
Liking/rejection	3.39 (1.79)	3.35 (1.28)	3.44 (1.53)
Moral cleansing	-	4.92 (1.51)	5.05 (1.25)
Moral repair	-	3.73 (1.68)	3.65 (1.77)

		Contrast effect size (*d*)		
Dependent variable	ANOVA *F*-test	C vs. SP1	C vs. SP2	SP1 vs. SP2
Values	$F(2,147) = 0.17$	-0.05	0.07	0.12
Offender status/power	$F(2,147) = 6.40$*	-0.61*	-0.58*	0.03
Justice satisfaction	$F(2,147) = 1.17$	0.00	0.28	0.25
Remorse	$F(2,147) = 7.41$**	0.63*	0.70*	0.08
Forgiveness	$F(2,147) = 1.17$	0.13	0.31	0.19
Trust	$F(2,147) = 2.34$	-0.07	0.33	0.44
Moral standing	$F(2,147) = 4.00$*	0.32	0.53*	0.27
Liking/rejection	$WF(2,96.37) = 0.26$[a]	-0.03	0.03	0.06
Moral cleansing	$t(97) = -0.47$[b]	-	-	0.09
Moral repair	$t(97) = 0.22$[b]	-	-	-0.05

[a]*Welch's F* is reported where a Levene Test indicated variances were not equal. [b] The control condition did not receive these items so an independent *t*-test was run instead.

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

Since there were no significant differences between the two self-punishment conditions on any of the variables, these were pooled into a single self-punishment condition to make use of the entire dataset. We then ran independent samples *t*-tests to explore main

effects of self-punishment on justice and reconciliation measures. Relative to control, seeing the perpetrator engaging in self-punishment increased observers' perceptions that he or she had less status/power, $t(82.51) = 3.32$, $p > .001$, and was more remorseful, $t(148) = -3.85$, $p > .001$, consistent with Study 5.1. Furthermore, self-punishment increased the transgressor's moral standing, $t(148) = -2.54$, $p = .01$. Effects on justice, forgiveness, trust, and liking were not significant.

Next we tested for any indirect effects on reconciliation via offender status/power (simple mediation). We again used the PROCESS macro for SPSS (Hayes, 2013; model 4) with 10,000 bootstraps and bias-corrected confidence intervals (adjusted to a higher threshold of 99% CIs for multiple tests), with the contrast between control and pooled self-punishment as the exogenous variable. There were small but statistically significant indirect effects of self-punishment on all the reconciliation measures via offender status/power (see Table 5.5). In so far as self-punishment diminished offender status/power, it increased perceptions of justice, and judgments that the transgressor was remorseful, trustworthy, deserving of forgiveness, and likeable. Not surprisingly, because self-punishment had no significant effects on the mediator value consensus, none of the indirect effects via value consensus were significant.

Table 5.5

Mediation Model Statistics: Indirect Effects of Self-Punishment Via Value Consensus and Status/Power (Study 5.2)

Model effects	*b*	*SE*	β	$CI_{99\%}$ (*b*)
Mediator (M): value consensus				
IV→M	0.01	0.23	.002	-0.59, 0.60
DV: justice				
M→DV	0.58	0.06	.64	0.43, 0.73
IV→DV (direct effect)	0.15	0.16	.06	-0.26, 0.57
IV→M→DV (indirect effect)	0.002	0.14	.001	-0.34, 0.36

Model effects	*b*	*SE*	β	$CI_{99\%}$ (*b*)
DV: remorse				
M→DV	0.60	0.08	.52	0.40, 0.80
IV→DV (direct effect)	0.97	0.21	.30	0.41, 1.52
IV→M→DV (indirect effect)	0.003	0.14	.001	-0.36, 0.39
DV: forgiveness				
M→DV	0.75	0.08	.60	0.53, 0.97
IV→DV (direct effect)	0.35	0.23	.10	-0.25, 0.96
IV→M→DV (indirect effect)	0.004	0.18	.001	-0.47, 0.46
DV: trust				
M→DV	0.69	0.06	.68	0.53, 0.85
IV→DV (direct effect)	0.17	0.17	.06	-0.28, 0.62
IV→M→DV (indirect effect)	0.004	0.16	0.001	-0.42, 0.42
DV: moral standing				
M→DV	0.40	0.08	.39	0.20, 0.60
IV→DV (direct effect)	0.58	0.21	.20	0.03, 1.13
IV→M→DV (indirect effect)	0.002	0.09	.001	-0.09, 0.08
DV: liking				
M→DV	0.64	0.08	.56	0.44, 0.85
IV→DV (direct effect)	0.06	0.22	.02	-0.52, 0.64
IV→M→DV (indirect effect)	0.003	0.15	.001	-0.41, 0.41
Mediator (M): status/power				
IV→M	-0.68	0.19	-.28	-1.18, -0.19
DV: justice				
M→DV	-0.34	.09	-.33	-0.56, -0.12
IV→DV (direct effect)	-0.08	0.21	-.03	-0.61, 0.46
IV→M→DV (indirect effect)	0.23	0.10	.09	0.04, 0.54
DV: remorse				
M→DV	-0.37	0.11	-0.28	-0.64, -0.10
IV→DV (direct effect)	0.72	0.25	.22	0.06, 1.38
IV→M→DV (indirect effect)	0.25	0.11	.08	0.04, 0.61
DV: forgiveness				
M→DV	-0.32	0.12	-.22	-0.64, -0.001
IV→DV (direct effect)	0.14	0.30	.04	-0.63, 0.91
IV→M→DV (indirect effect)	0.22	0.11	.06	0.02, 0.61

Model effects	*b*	*SE*	β	$CI_{99\%}$ (*b*)
DV: trust				
M→DV	-0.35	0.10	-.30	-0.60, -0.10
IV→DV (direct effect)	-0.07	0.24	-.02	-0.68, 0.55
IV→M→DV (indirect effect)	0.24	0.11	.08	0.04, 0.59
DV: moral standing				
M→DV	-0.23	0.10	-.20	-0.49, 0.02
IV→DV (direct effect)	0.43	0.24	.15	-0.19, 1.04
IV→M→DV (indirect effect)	0.16	0.08	.06	0.01, 0.46
DV: liking				
M→DV	-0.32	0.11	-.24	-0.61, -0.03
IV→DV (direct effect)	-0.15	0.27	-.05	-0.86, 0.55
IV→M→DV (indirect effect)	0.22	0.10	.07	0.03, 0.57

Note. $N = 150$. Indirect effect estimates bootstrapped with 10,000 samples.

Hypothesis 2: Excessive self-punishment can undermine its positive effects. To test our second hypothesis, we re-examined mean scores for outcome variables across the two self-punishment conditions. As already described, the two self-punishment conditions did not differ from one another on any outcome variable (see Table 5.4). Both brief and excessive self-punishment resulted in less offender power/status and more remorse, relative to control. For moral standing, only the excessive self-punishment differed significantly from control, while brief self-punishment did not differ significantly from excessive self-punishment or control. Thus, excessive self-punishment did not appear to have much influence—positive or negative—beyond a brief expression of self-punishment.

Although the manipulation did not provide strong evidence for our prediction, it could have been that the manipulation between the self-punishment conditions was too subtle to influence participants' perceptions of excessiveness. In addition, what is considered "excessive" may be subjective, depending on how participants perceive the severity of the transgression and self-punishment. Therefore, we further investigated whether there was any merit for the backfire hypothesis by exploring within-group data. Specifically, we determined

that the first manipulation check item, which elicited agreement with the statement that the perpetrator had self-punished, could be used as a subjective measure to tap into excessiveness of self-punishment. Though not a perfect (or planned) estimator of excessiveness, it may serve as a proxy for participants' perception of the intensity of the self-punishment.

To model the idea that excessive self-punishment will backfire, we would expect curvilinear relationships between self-punishment intensity and outcome measures. Scores on the manipulation check item were centred then squared to calculate a quadratic term. In the first step of a hierarchical multiple regression analysis, the untransformed (linear) item was entered, predicting the dependent variable. In the second step of the regression analysis, the quadratic term was added as a predictor. Significant quadratic terms in all models—except moral standing—confirmed the predicted relationships (see Table 5.6). Negative coefficients for the quadratic effect imply that the relationship between perceived self-punishment intensity was concave down (an inverted U shape), in line with the hypothesised backfire effect whereby initial positive effects of self-punishment were reversed at the extreme end. The function for remorse as a dependent variable is plotted in Figure 5.2 for illustration.

Table 5.6

Linear and Quadratic Effects of Self-Punishment Intensity on Outcome Variables (Study 5.2)

	Coefficients in model (*B*)			
Outcome variable	Linear	Quadratic	Adjusted R^2	R^2 change (step 2)
Values	-0.16*	-0.12**	.09**	.09**
Status/power	0.01	0.10**	.18**	.09**
Justice	-0.20**	-0.13**	.13**	.14**
Remorse	0.06	-0.11**	.19**	.06**
Forgiveness	-0.03	-0.11*	.09**	.05*
Trust	-0.10	-0.12**	.11**	.09**
Moral standing	0.13	-0.03	.09**	.01
Liking	-0.10	-0.10*	.06*	.06*

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

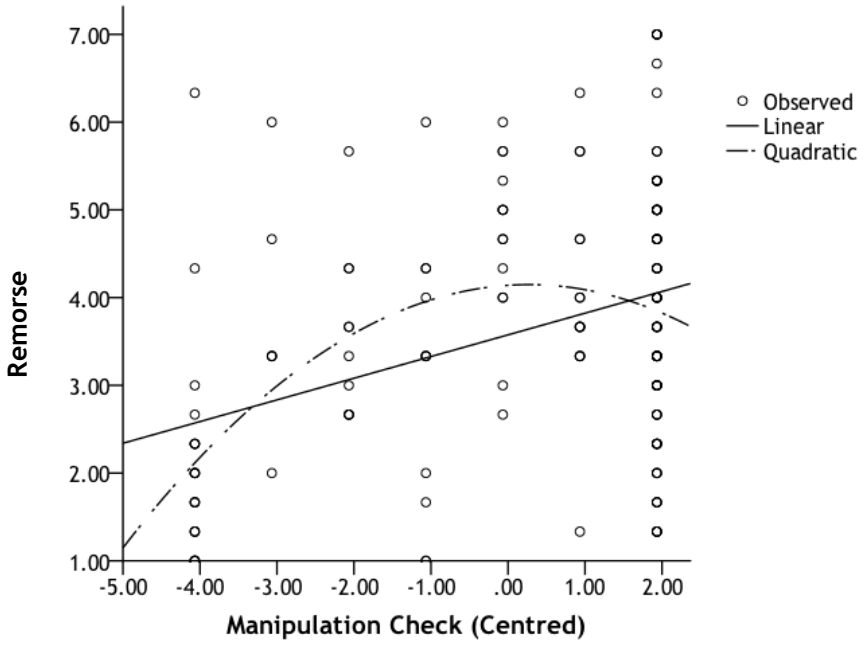


Figure 5.2. Relationships between self-punishment intensity and remorse (Study 5.2).

Hypothesis 3: Positive effects of self-punishment are driven by attributions of perpetrator sincerity. Last, we investigated the hypothesis that self-punishment would be beneficial to the self-punisher insofar as it was interpreted as a process of genuine moral repair rather than moral cleansing. Since there were no main effects of excessive self-punishment on moral repair and moral cleansing, we again turned to within-group data to explore this hypothesis. Correlations presented in Table 5.7 indicate that as observers became less critical of the motivation behind the self-punishment, they adopted more charitable and forgiving attitudes towards the perpetrator. Participants' interpretation of the self-punishment as a process of moral repair—that is, the perception that perpetrators were trying to think through what they had done and learn from the experience—was particularly influential. Moreover, consistent with Study 5.1 and in line with predictions, these interpretations were more strongly related to value consensus than status/power.

Table 5.7

Correlations Between Self-Punishment Interpretations and Dependent Variables (Study 5.2)

Measure	Moral cleansing	Moral repair
Moral cleansing	1	
Moral repair	-.36**	1
Values	-.22*	.68**
Status/power (offender)	.07	-.21*
Justice	-.47**	.73**
Remorse	-.13	.61**
Forgiveness	-.19	.63**
Trust	-.23*	.58**
Moral standing	-.01	.42**
Liking	-.14	.56**

Note. $n = 99$.

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

Discussion

Self-punishment was again a useful strategy for transgressors to restore their public moral image. This was in part attributable to its symbolic value as a resignation of one's status/power—an act of self-degradation. However, the total effects were not as pronounced as in Study 5.1. One notable difference between the two studies was that self-punishment had no effect on perceived value consensus in the current study. Given that a perceived value consensus was critical to observers' attitudes towards transgressors in Study 5.1, a failure to communicate this in Study 5.2 may have limited self-punishment's effects. The reason for this discrepancy is unclear. It may be that findings in the current (or indeed, the previous) study were due to random sampling error: The mean scores on dependent variables in the control conditions of Studies 5.1 and 5.2 (for which the instructions were identical) differed dramatically between the two studies. In Study 5.2, participants in the control condition were much more charitable than those in Study 5.1 across most of the dependent variables, limiting the incremental effect of self-punishment on these variables.

Alternatively, we could speculate that the ability for self-punishment to communicate a value consensus (and thus influence other outcome variables) might depend on the type of self-punishment employed. Perhaps observers intuit that social isolation implies a value consensus more than overworking oneself. In addition, it should be noted that indirect effects were still detectable when total effects were not (i.e., on justice, forgiveness, trust, and liking). This suggests that in Study 5.2 there were negative factors working against the positive pathway via status/power, leading the effects to cancel each other out in some cases. This could be again due to a characteristic of the overwork scenario—some feature of the behaviour that elicits negative reactions. Thus, one should not assume that self-punishment is homogenous across its different expressions.

Results also suggested that cynical interpretations might shape observers' responses to self-punishment, but the overall positive effects of self-punishment appeared relatively resistant to the excessiveness manipulation. Excessive self-punishment did appear to yield diminishing returns, but there was no backfire effect. It is possible that the experimental manipulation was too subtle or not excessive enough (i.e., not far along enough on the theoretical curve) to obtain the backfire effect, suggesting that self-punishment may need to be quite exaggerated before it backfires. In line with this, within-groups data showed that more extreme self-punishment might backfire on the self-punisher.

Study 5.3

The final study had several aims. First, we sought to demonstrate once again self-punishment's positive effects on observers' attitudes, mediated by a perceived value consensus and status/power (Hypothesis 1). To maximise the chances of achieving this effect—thereby facilitating exploration of subsequent hypotheses that rely on it—we returned to the self-isolation form of self-punishment utilised in Study 5.1, which appeared to have stronger effects on attitudes compared to overwork (Study 5.2). In order to balance repeating

this paradigm with the generalisability of our findings, we changed the transgression context instead.

Second, we again tested a boundary condition for these effects: We predicted excessive self-punishment would have a backfire effect on positive attitudes (Hypothesis 2a). Though we did not observe any backfire effects of excessive self-punishment in Study 5.2, we hoped that returning to a social isolation form of self-punishment would be more fruitful, as it had stronger main effects on attitudes. In addition, we tested an alternate boundary condition for self-punishment (as discussed in the introduction): When there is the opportunity for direct victim repair, self-punishment might be seen as a cop-out, an avoidance of constructive but potentially more uncomfortable actions. We predicted that under conditions of repair being possible, self-punishers would be subject to more suspicion than those in the other two self-punishment conditions (Hypothesis 2b).

Further, we hypothesised that effects of self-punishment—its ability to communicate a relinquishment of status/power and in particular, a value consensus—would be underpinned by interpretations of the act as one that is driven by genuine moral repair, rather than moral cleansing (Hypothesis 3).

Method

Participants (N = 200) were recruited online as per the previous protocol. The sample was 52% female and 48% male, mean age was 36.72 years (SD = 12.93).

The hypothetical transgression vignette was changed to the following:

Louis and Daniel are two men who have been close friends for many years. One day they come across a very attractive job advertisement and they think it would be "really cool" to get that job. They both decide to go for it. They sit down together that day and complete their applications. After completing all the forms, they need to send their applications by post. Louis offers to deliver both his and Daniel's applications to

the post office. However, when Louis gets to the post office, suddenly a thought crosses his mind. Louis knows that Daniel is perfect for the role and thinks he will have a better chance at getting the position if Daniel is out of the running. So, on the spur of the moment, instead of posting both applications he only posts his own application and throws Daniel's away. Louis doesn't get the job in the end.

When Daniel calls the company to ask for feedback, he learns that his application was not received. Since the applications were supposedly sent together, he suspects what Louis has done.

There were four experimental conditions: control (n = 50; text as per previous studies), a brief period of self-imposed social isolation (n = 50; text as per Study 5.1), a more excessive period of self-punishment (n = 50; text as per Study 5.2 but modified to self-isolation), and self-punishment despite repair (n = 50). This fourth condition was similar to the excessive self-punishment but also included overtures from the victim to spend time together (and theoretically discuss the wrongdoing): "Over the course of many months afterwards, Louis isolates himself, not allowing himself to go out or see his friends. Daniel continually invites him out, giving Louis many chances to spend time with Daniel. Six months later, Louis is still isolating himself, despite many invitations from Daniel." This condition intentionally contained both an excessive self-punishment and a reluctance to reconcile with the victim, in order to maximise the chances that it would elicit cynical interpretations.

Participants completed the same outcome measures of value consensus (α = .79), offender status/power (α = .70), justice satisfaction (α = .72), remorse (average item intercorrelation = .34), forgiveness (r = .78, p < .001), trust (average item intercorrelation = .52), moral standing (α = .77), liking/rejection (α = .91), moral repair (r = .60, p < .001), and

moral cleansing (r = .46, p < .001) as in Study 5.2. Moral repair and moral cleansing were inversely correlated, r = -.28, p < .001.

The victim status/power scale was once again revised based on analysis from the previous study. The revised scale included: “Daniel feels secure in his relationships”; “Daniel feels like an easy target to push around” (reverse coded); “Daniel feels respected by others”; “Daniel feels weak relative to others” (reverse coded). The third attempt at this scale appeared successful (α = .74).

We included three manipulation checks. The first item assessed whether the perpetrator had engaged in self-punishment as per Studies 4.1 and 4.2 (presented to all participants), and the second item assessed for how long the self-punishment had gone on for, as per Study 5.2 (self-punishment conditions only). A third item asked whether Daniel had invited Louis out, designed to distinguish the repair opportunity condition (all conditions).

Results

Manipulation checks indicated that the experimental manipulations were successful. There was a significant overall condition effect on the perception that the perpetrator had self-punished, *Welch's* $F(3, 99.35) = 88.56$, p < .001. Planned contrasts confirmed that participants in the brief self-punishment (M = 6.70, SD = 0.89), excessive self-punishment (M = 6.84, SD = 0.55), and self-punishment with repair opportunity conditions (M = 6.42, SD = 1.62) were more certain that the perpetrator had self-punished relative to those in the control condition (M = 2.32, SD = 1.88; all ps < .001). For the second check, results indicated that the excessive self-punishment condition (M = 6.42, SD = 1.49) and the self-punishment despite repair condition (M = 6.40, SD = 1.18) both believed that the perpetrator had self-punished for a longer period than those in the brief self-punishment condition (M = 2.64, SD = 1.59; both ps < .001; overall *Welch's* $F(2, 96.25) = 104.50$, p < .001). Last, there was an overall condition effect on the check for repair opportunity, $F(3, 196) = 96.82$, p < .001, such that—

as expected—the repair opportunity condition scored higher (M = 6.16, SD = 1.71) than the control (M = 2.18, SD = 1.56), brief self-punishment (M = 1.76, SD = 1.22), and excessive self-punishment conditions (M = 1.74, SD = 1.62; all ps < .001).

Hypothesis 1: Self-punishment promotes reconciliation, mediated by value consensus and status/power. One-way ANOVAs were conducted with post-hoc tests to determine the nature of the manipulation effects. *Tukey* post-hoc tests revealed that there were no significant differences between the three self-punishment conditions on any outcome variables, except for victim status/power (we will discuss the ANOVA results in more detail in the next sections). Therefore, as in Study 5.2, we pooled the self-punishment conditions into a single self-punishment condition (n = 150).[6] Relative to control, observers of self-punishment felt that there was more justice restored, while perceiving self-punishers to possess more shared values, less status/power, to be more remorse and likeable, and granting them more forgiveness, trust, and moral standing; these were medium to large effects (see Table 5.8).

[6] The only variable on which the self-punishment conditions differed was victim status/power, driven by the elevated scores in the self-punishment despite repair condition. Given this measure was possibly a proxy for victim agency (see results for Hypothesis 2b), and also given the low internal validity of the measure's previous iterations, we do not consider this difference a good reason against pooling the data. Nevertheless, victim status/power is not included in the subsequent analyses.

Table 5.8

T-Test Statistics for Differences Between Control and Pooled Self-Punishment Conditions (Study 5.3)

	M (*SD*)				
Dependent variable	No self-punishment	Self-punishment	*t*	*p*	*g*[a]
Values	1.65 (0.84)	2.85 (1.35)	*t*(136.34) = 7.44	< .001	0.97
Offender status/power	4.21 (1.37)	2.88 (1.31)	*t*(198) = -6.84	< .001	-1.11
Justice	1.66 (0.92)	2.26 (0.99)	*t*(198) = 3.74	< .001	0.62
Remorse	1.61 (0.86)	2.85 (1.35)	*t*(116.67) = 13.81	< .001	1.91
Forgiveness	2.80 (1.58)	3.68 (1.60)	*t*(198) = 3.38	< .001	0.55
Trust	1.73 (0.90)	2.92 (1.32)	*t*(123.49) = 7.12	< .001	0.97
Moral standing	3.12 (1.16)	3.94 (1.21)	*t*(198) = 4.25	< .001	0.69
Liking	2.10 (1.07)	3.16 (1.41)	*t*(109.51) = 5.58	< .001	0.79
Moral cleansing	5.13 (1.75)	5.61 (1.46)	*t*(198) = 1.93	.06	0.31
Moral repair	1.43 (0.75)	2.73 (1.39)	*t*(157.15) = 8.39	< .001	1.03

[a]Hedges' *g* is reported instead of Cohen's *d* as a measure of effect size as it better accounts for unequal cell sizes.

Next we examined whether these effects could be explained by value consensus and offender status/power (see Footnote 6), using the same two-step procedure as per the previous two studies (PROCESS model 4, 10,000 bootstraps, 99% bias-adjusted CIs) with the contrast between control and pooled self-punishment as the exogenous variable. Mediation models indicated partial support for our hypothesis: Insofar as the self-punishment led to a sense that the transgressor shared values with participants (and the broader moral community), it restored justice and increased perceptions of the transgressor as remorseful, trustworthy and likeable, granting them forgiveness and moral standing. In addition, self-punishment also increased perceptions of likeability via status/power (see Table 5.9).

Table 5.9

Mediation Model Statistics: Indirect Effects of Self-Punishment Via Value Consensus and Status/Power (Study 5.3)

Model effects	*b*	*SE*	β	$CI_{99\%}$ (*b*)
Mediator (M): value consensus				
IV→M	0.60	0.10	.39	0.34, 0.86
DV: justice				
M→DV	0.35	0.05	.47	0.22, 0.48
IV→DV (direct effect)	0.09	0.08	.08	-0.11, 0.29
IV→M→DV (indirect effect)	0.21	0.04	.18	0.12, 0.33
DV: remorse				
M→DV	0.43	0.06	.40	0.28, 0.58
IV→DV (direct effect)	0.82	0.09	.49	0.59, 1.05
IV→M→DV (indirect effect)	0.26	0.06	.15	0.13, 0.42
DV: forgiveness				
M→DV	0.28	0.09	.23	0.04, 0.51
IV→DV (direct effect)	0.27	0.14	.15	-0.09, 0.63
IV→M→DV (indirect effect)	0.17	0.06	.09	0.02, 0.34
DV: trust				
M→DV	0.46	0.06	.46	0.30, 0.62
IV→DV (direct effect)	0.32	0.10	.21	0.07, 0.57
IV→M→DV (indirect effect)	0.28	0.06	.18	0.14, 0.45
DV: moral standing				
M→DV	0.22	0.07	.24	0.05, 0.40
IV→DV (direct effect)	0.28	0.10	.20	0.01, 0.55
IV→M→DV (indirect effect)	0.13	0.05	.09	0.02, 0.28
DV: liking				
M→DV	0.35	0.07	.34	0.17, 0.54
IV→DV (direct effect)	0.32	0.11	.20	0.03, 0.61
IV→M→DV (indirect effect)	0.21	0.07	.13	0.07, 0.41
Mediator (M): status/power				
IV→M	-0.67	0.10	-.44	-0.92, -0.41
DV: justice				
M→DV	0.05	0.06	.07	-0.10, 0.20

Model effects	*b*	*SE*	β	$CI_{99\%}$ (*b*)
IV→DV (direct effect)	0.33	0.09	.29	0.10, 0.56
IV→M→DV (indirect effect)	-0.03	0.04	-.03	-0.14, 0.07
DV: remorse				
M→DV	-0.07	0.07	-.07	-0.25, 0.10
IV→DV (direct effect)	1.03	0.10	.61	0.76, 1.30
IV→M→DV (indirect effect)	0.05	0.05	.03	-0.07, 0.19
DV: forgiveness				
M→DV	-0.03	0.09	-.03	-0.28, 0.22
IV→DV (direct effect)	0.42	0.14	.22	0.04, 0.80
IV→M→DV (indirect effect)	0.02	0.07	.01	-0.18, 0.20
DV: trust				
M→DV	-0.14	0.07	-.14	-0.33, 0.05
IV→DV (direct effect)	0.50	0.11	.33	0.21, 0.79
IV→M→DV (indirect effect)	0.10	0.05	.06	-0.02, 0.26
DV: moral standing				
M→DV	-0.12	0.07	-.13	-0.31, 0.06
IV→DV (direct effect)	0.33	0.11	.23	0.05, 0.61
IV→M→DV (indirect effect)	0.08	0.05	.06	-0.04, 0.24
DV: liking				
M→DV	-0.21	0.08	-.20	-0.41, -0.01
IV→DV (direct effect)	0.39	0.12	.24	0.08, 0.70
IV→M→DV (indirect effect)	0.14	0.06	.09	0.01, 0.34

Note. $N = 200$. Indirect effect estimates bootstrapped with 10,000 samples. For the DV likeability (for which both mediators are significant), the indirect effects via both values and status/power remain significant when tested simultaneously in a parallel mediation model.

Hypothesis 2a: Excessive self-punishment can undermine its positive effects.

Descriptive and test statistics across all four experimental conditions are presented in Table 5.10. As already explained, there were no significant differences between the self-punishment conditions on any outcome variables. Relative to control, both brief and ongoing self-punishment conditions led to favourable judgments for the perpetrator (i.e., more value consensus, less offender status/power, more remorse, trust, moral standing, and liking, all *p*s

< .05); however, on these measures the excessive and brief self-punishment conditions did not significantly differ from one another (all *p*s > .05). For justice satisfaction and forgiveness, excessive self-punishment resulted in more justice and more forgiveness relative to control (all *p*s < .05); the brief self-punishment condition did not differ significantly from control or from the excessive self-punishment condition on these two measures (all *p*s > .05). The three conditions did not differ on victim status/power. Taken together, at the condition level the results indicate a diminishing returns pattern, whereby excessive self-punishment did not confer any additional advantage over brief self-punishment. However, there was no evidence for the predicted backfire effect (i.e., excessive self-punishment did not confer any disadvantage over brief self-punishment).

Table 5.10

Descriptive and Between-Group Statistics for Dependent Variables (Study 5.3)

	M (*SD*)			
Dependent variable	Control	Brief self-punish	Excessive self-punish	Self-punish despite repair
Values	1.65 (0.84)	2.71 (1.32)	3.13 (1.29)	2.72 (1.41)
Offender status/power	4.21 (1.37)	3.14 (1.25)	2.87 (1.17)	2.63(0.91)
Victim status/power	3.94 (1.49)	4.13 (1.20)	4.19 (1.24)	4.80 (1.21)
Justice	1.66 (0.92)	2.15 (1.08)	2.35 (0.91)	2.27 (0.98)
Remorse	1.61 (0.86)	3.75 (1.26)	3.76 (1.22)	3.78 (1.14)
Forgiveness	2.80 (1.58)	3.55 (1.66)	3.78 (1.60)	3.71 (1.55)
Trust	1.73 (0.90)	2.73 (1.33)	3.21 (1.32)	2.83 (1.33)
Moral standing	3.12 (1.16)	3.81 (1.16)	4.16 (1.19)	3.86 (1.26)
Liking	2.10 (1.07)	2.90 (1.34)	3.39 (1.44)	3.20 (1.42)
Moral cleansing	5.13 (1.75)	5.62 (1.58)	5.43 (1.41)	5.79 (1.38)
Moral repair	1.43 (0.75)	2.65 (1.32)	2.89 (1.56)	2.65 (1.27)

Dependent variable	ANOVA *F*-test
Values	*WF*(3,106.11) = 19.68**[a]
Offender status/power	*F*(3,196) = 17.32**
Victim status/power	*F*(3,196) = 4.18*
Justice satisfaction	*F*(3,196) = 5.02*
Remorse	*F*(3,196) = 45.55**
Forgiveness	*F*(3,196) = 3.97*
Trust	*WF*(3,107.17) = 17.96**[a]
Moral standing	*F*(3,196) = 6.86**
Liking/rejection	*F*(3,196) = 9.14**
Moral cleansing	*F*(3,196) = 1.69
Moral repair	*F*(3,104.09) = 23.09**[a]

[a]*Welch's F* is reported where a Levene Test indicated variances were not equal.

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

To further probe our hypothesis, we explored whether there was a backfire effect of self-punishment at the within-group level across the whole sample, using linear and quadratic terms of the first manipulation check item (as a proxy for self-punishment intensity) in a

regression framework as described in Study 5.2. All models were significant (see Table 5.11), but predominantly due to the linear effects. For justice, moral cleansing, and moral repair there were significant unique quadratic effects, reflecting the hypothesised initial rise followed by decline (in the inverse direction for moral cleansing, as would be expected).

Table 5.11

Linear and Quadratic Effects of Self-Punishment Manipulation Check on Outcome Variables (Study 5.3)

	Coefficients in model (*B*)			
Outcome variable	Linear (linear only[a])	Quadratic	Adjusted R^2	R^2 change (step 2)
Values	-0.01 (0.17**)	-0.06	.09**	.01
Status/power	-0.37** (-0.26**)	-0.04	.20**	.01
Justice	-0.14 (0.06)	-0.07*	.04*	.03*
Remorse	0.39** (0.35**)	0.02	.29**	.001
Forgiveness	0.12 (0.19**)	-0.03	.06**	.001
Trust	0.13 (0.18**)	-0.02	.09**	.002
Moral standing	0.22* (0.19**)	0.01	.12**	.00
Liking	0.07 (0.20**)	-0.05	.10**	.01
Moral cleansing	0.42** (0.16**)	0.09*	.07**	.02*
Moral repair	-0.05 (0.20**)	-0.09*	.12**	.02*

[a]The value in parentheses represents the linear term coefficient before the quadratic term was added to the models.

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

Hypothesis 2b: Self-punishing at the expense of repair can undermine its positive effects. ANOVA post-hoc tests also indicated that self-punishing at the expense of repair was a better strategy than doing nothing at all (i.e., for the comparison between control and self-punishment despite repair conditions on outcome variables, all *p*s < .05). Examination of mean scores (see Table 5.10) shows that across about half of the variables, there was a pattern whereby the self-punishment despite repair tended to backfire a little on average, relative to the excessive self-punishment condition. However, the three self-punishment conditions did not differ significantly from one another on any outcome (all *p*s > .05). In other words,

though engaging in self-punishment was beneficial for perpetrators, it did not seem to matter whether the self-punishment was brief, excessive, or at the expense of repair. Thus, contrary to predictions, self-punishing at the expense of repair did not backfire on the transgressor in any meaningful sense.

An exception to the above pattern of results was for victim status/power: Those in the self-punishment despite repair condition perceived the victim to have significantly more status/power than those in the control and brief self-punishment condition (all *p*s < .05), and marginally more than those in the excessive self-punishment condition ($p = .09$). Brief self-punishment and excessive self-punishment did not differ significantly from neither control nor from one another on this measure (*p*s > .05). This is counterintuitive, given that in the self-punishment despite repair condition the victim was explicitly denied the chance to receive a direct apology or explanation for the transgression, nor to give the perpetrator any dressing down—opportunities that might normally allow the victim to regain status/power. In light of this, one may suspect that the victim status/power scale was used as a sort of proxy for victim agency or salience, merely recognising that the victim was more active in this scenario.

Hypothesis 3: Positive effects of self-punishment are driven by attributions of perpetrator sincerity. We next examined whether attributions of self-punishers as avoiding or confronting their wrongdoing were instrumental in shaping observers' attitudes. To begin with, we examined differences between conditions on measures of moral cleansing and moral repair (see Table 5.10). Relative to control, all self-punishers (brief, excessive, and despite repair) were more likely to be perceived to be confronting their transgression (i.e., moral repair, *p*s < .05), but the self-punishment conditions did not differ significantly from one another (*p*s > .05). Conditions did not vary on moral cleansing (all *p*s > .05).

Using pooled data, the effects were more straightforward (see Table 5.8). Self-punishers were viewed as both more likely to be confronting their transgression, while also (marginally) more likely to be avoiding their transgression. The more marked effect for moral repair is consistent with the overall positive effects of self-punishment seen in the data (as addressed in Hypothesis 1). These results suggest that participants tended to interpret any kind of self-punishment as an attempt to work through one's wrongdoing, but to some extent also viewed it as an evasive act.

Intercorrelations indicated that these interpretations had a strong bearing on a number of the outcome variables (see Table 5.12). In line with predictions, observers' interpretations of the perpetrator being motivated by moral repair or moral cleansing were more strongly correlated with achieving a value consensus than with redistributing status/power. Moreover, when observers interpreted the perpetrator's behaviour as a process of moral repair, they were more willing to reconcile with him/her.

Table 5.12

Correlations Between Interpretations of Perpetrator Motivation and Dependent Variables (Study 5.3)

Measure	Moral cleansing	Moral repair
Moral cleansing	1	
Moral repair	-.28**	1
Values	-.24*	.46**
Status/power (offender)	-.18*	-.14
Justice	-.32**	.42**
Remorse	-.08	.49**
Forgiveness	.01	.32**
Trust	-.06	.41**
Moral standing	.10	.21*
Liking	.01	.30**

Note. $N = 200$.

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

Next, we conducted mediation analyses to further explore whether these interpretations were instrumental in informing observers' attitudes towards interpersonal reconciliation. The proposed mechanism is as follows: Observers perceive self-punishers to be thinking through and being critical of their wrongdoing (moral repair), which in turn increases the perception that self-punishers are committed to upholding shared moral values, which leads to forgiveness and acceptance. The procedure for the current mediation analyses was similar to that in Hypothesis 1 above, except moral repair and value consensus were entered as sequential mediators (PROCESS model 6; see Table 5.13 and Figure 5.3). Models were significant for five of the six outcome variables, supporting the hypothesis.

Table 5.13

Indirect Effect Estimates of Self-Punishment on Outcome Variables via Moral Repair and Value Consensus (Study 5.3)

	Indirect effect pathways in sequential models								
	Via moral repair only			Via values only			Via both moral repair and values		
Dependent variable	β	*SE*	$CI_{99\%}$	β	*SE*	$CI_{99\%}$	β	*SE*	$CI_{99\%}$
Justice	.10	.03	.03, .19	.09	.03	.03, .19	.06	.02	.02, .11
Remorse	.06	.02	.001, .13	.08	.03	.03, .16	.05	.01	.02, .10
Forgiveness	.09	.03	.01, .19	.04	.02	-.01, .11	.02	.01	-.01, .07
Trust	.06	.03	-.01, .15	.10	.03	.03, .19	.06	.02	.02, .12
Moral standing	.01	.03	-.08, .10	.06	.02	.01, .14	.04	.01	.004, .08
Liking	.03	.04	-.06, .13	.07	.03	.01, .17	.05	.02	.01, .11

Note. $N = 200$. Indirect effect estimates bootstrapped with 10,000 samples.

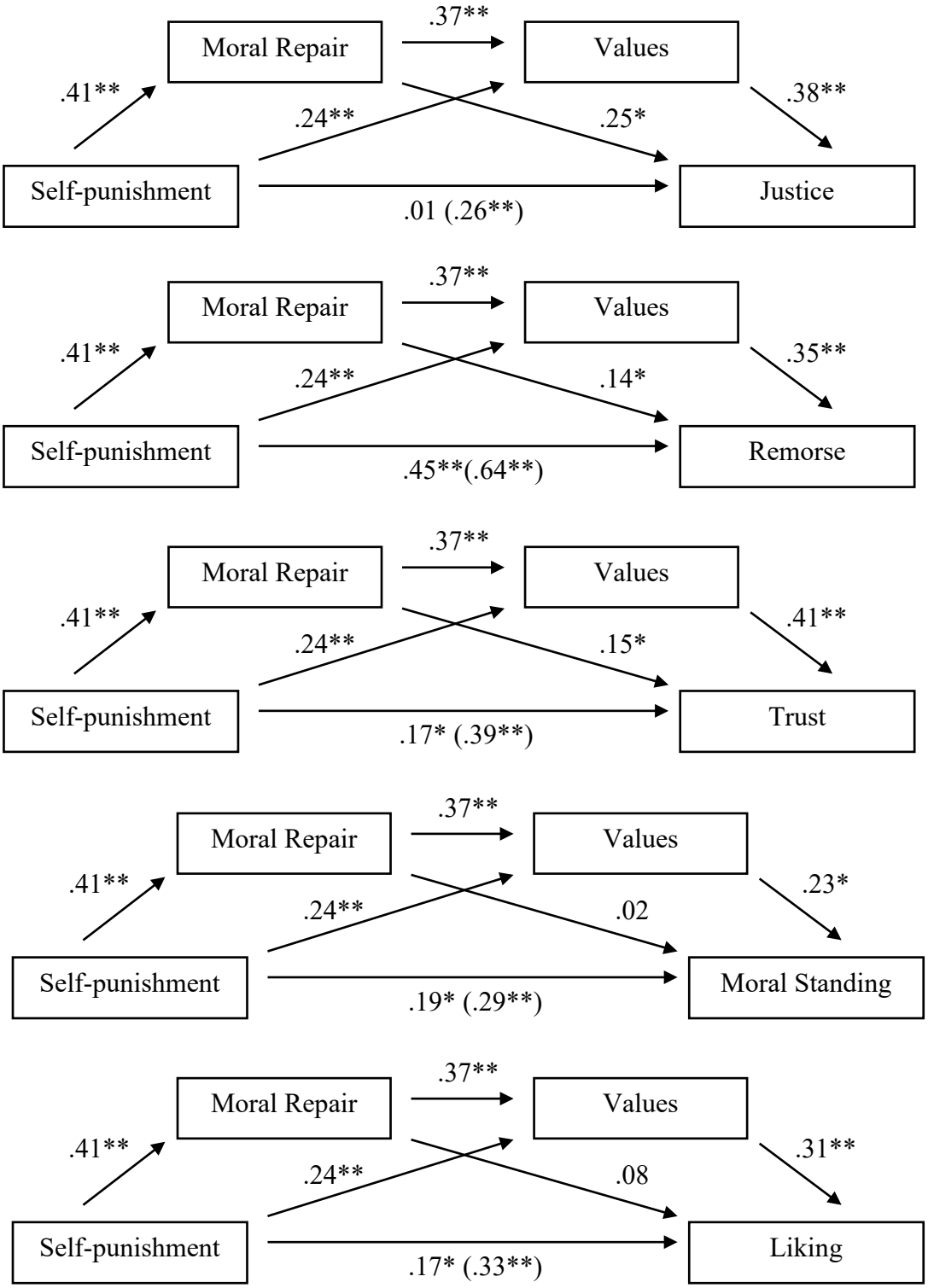


Figure 5.3. Indirect effect models of self-punishment via moral repair and value consensus (Study 5.3). All path values are expressed in standardized coefficients.

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

Discussion

Overall, self-punishment was again a useful strategy for transgressors to regain their peers' support. Observers believed self-punishers were undergoing a process of moral repair,

and rewarded them with forgiveness and acceptance—in part due to the belief that self-punishers were committed to group values. In other words, self-punishment effectively communicated a value consensus, and in doing do, redeemed the transgressor.

These positive effects were surprisingly robust, namely unaffected by manipulated variations in the excessiveness of the self-punishment and the availability of other, more direct (and arguably more constructive) options for repair. Moreover, the self-punishment need not be particularly drastic; though excessive self-punishment was not counterproductive to the transgressor, there were no additional gains to be made by going to greater lengths to prove one's remorse.

Correlational data did provide some evidence of a quadratic relationship between self-punishment intensity and interpretations of the act, suggesting that excessive self-punishment might backfire in some cases, but what is excessive is perhaps more in the eye of the beholder rather than the objective duration of self-punishment. However, this analysis was post-hoc and should be taken with caution. Furthermore, the direction of causality is unclear—it is possible that excessiveness was itself an impression that people formed as a function of the motives they attributed to transgressors; that is, observers perceived the self-punishment either positively or negatively, as avoidant or as reparative, and then these judgments led to the belief that the punishment was excessive or not.

There was evidence for the notion that the effects of self-punishment—its ability to communicate a value consensus in particular—is augmented by charitable interpretations of the self-punisher as an individual who is thinking through and confronting their wrongdoing. Thus, the transgressor's intent and inner psychological processing of the event appear to be factors determining observers' attitudes.

General Discussion

Summary of Findings

Across three experimental studies, self-punishment was a powerful signal that could restore transgressors' moral image in the eyes of third parties. Seeing perpetrators punish themselves addressed observers' justice concerns, through the derogation of the perpetrator's status/power (Studies 5.1 and 5.2), and through communicating the perpetrator's commitment to shared values (Studies 5.1 and 5.3). When this message was communicated to observers, self-punishment was a pathway to interpersonal reconciliation. As a comparison point, seeing transgressors punishing themselves was just as satisfying—if not more so—than seeing victims take their own revenge (Study 5.1).

The positive effects of self-punishment were robust across different transgression contexts and across two possible cases in which the transgressor's sincerity might be called into question. There was little evidence for the predictions that excessive self-punishment (Studies 5.2 and 5.3) and self-punishing despite repair opportunities (Study 5.3) could undermine this interpersonal function. At one level then, the general effectiveness of self-punishment would make this a convenient strategy for perpetrators that would indeed allow them to avoid potentially more arduous or inconvenient forms of victim and relationship repair.

At another level, however, observers were sensitive to the self-punisher's intention: When they believed that self-punishers were thinking through what they did wrong and attempting to learn from their self-punishment (i.e., when they saw it as a process of moral repair), they were also more likely to have the impression that the transgressor held shared moral values, which was critical in determining whether observers saw the act as sincere and eventually excused perpetrators for their misdeeds. Thus, judgments of the intention and psychological process underlying the self-punishment were influential in determining

attitudes, though it is still unclear under what conditions self-punishment leads to these perceptions (if not excessiveness and the opportunity for direct repair). Moreover, though the perception of moral repair was influential, evidence of merely partial mediation suggests it was not necessary for establishing a value consensus or for moral redemption. Perhaps simply engaging in an act of self-punishment, irrespective of motive, may be sufficient to communicate an acknowledgement of the violated values—or perhaps to satisfy alternate notions such as just desert—and thus to facilitate forgiveness.

Implications and Future Directions

First and foremost, the findings provide empirical evidence for the notion that self-punishment can wash away one's sins—at least, in the eyes of third party observers. This complements previous work that has pointed to an interpersonal function of self-punishment from the perpetrator's perspective (Nelissen, 2012; Tanaka et al., 2016; Tanaka et al., 2015; Zhu et al., 2017), and provides experimental evidence for the notion that penalising the self after a wrongdoing can foster forgiveness (Allen et al., 2015). Moreover, much of the total effect of self-punishment can be explained by its ability to address observers' symbolic injustice concerns (Okimoto & Wenzel, 2008): both through relinquishing one's illegitimately-gained status/power, and by establishing a value consensus whereby transgressors accept what they did was wrong and communicate that they are committed to shared moral values.

Though the current findings cannot speak to how self-punishment compares to observers themselves inflicting the punishment, the fact that observers were equally satisfied with victim-inflicted punishment suggests that the messages communicated via self-punishment are powerful. Indeed, if a transgression poses a threat to the legitimacy of shared values, it is the transgressor who has the most power to re-establish the validity of those values by renouncing the violation. Though it is possible that observers can gain reputational

credit by administering punishment (Adams, 2011; Barclay, 2006; Kurzban et al., 2007), this may not always outweigh the potential costs of doing so (Kriss et al., 2016), thus it is not obvious that the reconciliatory effects of self-punishment would be eclipsed by third-party punishment. Of course, this is yet to be empirically determined, and is an avenue for further exploration.

It is also unclear whether *victims* would be equally swayed by seeing their antagonist engage in self-punishment. Research indicates that victims are interested in re-establishing their own sense of power (Orth, 2003; Shnabel & Nadler, 2008) by sending a message to perpetrators through delivering punishment themselves rather than just seeing transgressors suffer (Gollwitzer et al., 2014; Strelan et al., 2017). Victims may also seek empowerment through more restorative actions such as restitution or an apology (Darley & Pittman, 2003; Exline et al., 2007; Okimoto & Wenzel, 2008; Witvliet et al., 2008). Although it is theoretically plausible that self-punishment could augment the victim's status/power to some degree by virtue of it stripping the perpetrator's, this is not necessarily implied. There were some indications to suggest self-punishment could empower the victim, at least from the third party's perspective (Study 5.3), but there were concerns about the validity of this measure. If self-punishment were indeed limited in its ability to restore victim status/power, then victims may be left unsatisfied by an act of self-punishment.

On the other hand, the logical implication of Nelissen's (2012) argument—that self-punishment is directed at the victim, not third parties—is that victims should be convinced by this action. Moreover, research suggests that victims are more motivated than third parties to make more benign attributions about the transgressor and repair the relationship (Green et al., 2008). Thus, one could also make the argument that victims may be *more* receptive than third parties to self-punishment. Future research could elucidate whether victims are more or less convinced by self-punishers' moral transformation; though at the least, the strength of the

current findings imply that we should not limit our thinking about the interpersonal function of self-punishment (or indeed, of any response to wrongdoing) to the victim-offender dyad. Holding perpetrators accountable for their actions is a concern not only for victims, but a joint enterprise for the entire social group of which these two parties are part.

According to the current studies, self-punishers can regain social standing by virtue of self-punishment as an act of justice restoration. However, there appear to be alternate pathways to redemption, as evidenced by some of the partial mediations of value consensus and status/power on reconciliation. What other mechanisms outside the justice framework might account for this residual effect? One should not overlook automatic responses to witnessing suffering, driven by "ancient biological systems" (Craig, 2009, p. 29) that instinctively elicit sympathy by seeing others in pain (Bastian et al., 2014), and through this, forgiveness (Riek & Mania, 2012). From this perspective, self-punishment might lead to a sense of empathy that is somewhat independent from attributions about the transgression and self-punishment act themselves; in fact, it might distract from more reflective justice considerations and narrow observers' focus on the suffering of the self-punisher. In lay terms, self-punishers may be able to "play the victim" without anybody suspecting or caring that they are doing so—the instinctive response to the suffering may be too powerful to be altogether cast aside by more reflective processes. It may be worthwhile to disentangle the contributions of empathy/sympathy and justice-related cognition, for instance in an experimental design in which both a justice concern (e.g., the importance of the value violated) and empathy are manipulated.

Limitations

All three studies used hypothetical scenarios to examine observer responses to self-punishment. This may limit the ecological validity of the findings. Hypothetical scenarios afford experimenters a high degree of control over confounding variables, thus producing a

cleaner test for noisy phenomena. Though there is some concordance between moral judgments about real events and about hypothetical events (Turiel, 2008), there is little doubt that hypothetical scenarios lack the complexity and emotional engagement of real-life scenarios. In particular, the failure of the excessiveness manipulation in Studies 5.2 and 5.3 might have been due to the inability of a hypothetical interaction at a single time point to generate the empathic and cognitive demands that might occur when a person is subject to continuous displays of self-punishment over a period of time.

One could address this limitation by eliciting a transgression and self-punishment experimentally, in either an elaborate confederate-aided set-up, or through a computer game with ostensible players (e.g., an economic game, Tanaka et al., 2016). Though both of these options carry their own doubts about credibility, such paradigms could provide convergent validity for the current findings. In addition, within the hypothetical paradigm, efforts could be made to more robustly test the backfire effect of excessive self-punishment. For instance, ecological validity could be boosted through the introduction of *repeated* hypothetical interactions with feedback (participants could imagine themselves consoling the self-punisher, followed by information that the transgressor is engaging in yet more self-punishment), in which a fatigue element may be better drawn out.

Another limitation of the results is that they may not be generalisable to all types of self-punishment. Indeed, results differed somewhat between social isolation and overwork. It is intriguing to consider whether different types of self-punishment may have varying signalling currencies; perhaps some convey a stronger sense of value consensus (as observed for social isolation), while others may be more readily perceived as status/power restoration (as observed for overwork). If this is the case, then even the same type of self-punishment may not consistently redeem the perpetrator's moral image; rather, extending Okimoto and Wenzel's (2008) argument, the nature of the symbolic concerns elicited by the particular

transgression might determine victim and observer preferences for different types of self-punishment. The symbolic fit of self-punishment and transgression is another avenue for future research.

Conclusion

Self-punishment is a beneficial interpersonal strategy that can persuade observers that justice has been done, by signalling a commitment to shared moral values and a surrender of ill-gotten status/power. In this sense, though the Medieval self-flagellants were arguably misguided in the target of their self-punitive behaviour, they were not entirely mistaken about the consequences of their actions: Self-punishment can indeed elicit interpersonal support and forgiveness—that is, it can provide a moral absolution of sorts.

CHAPTER 6: General Discussion

In this thesis, I have undertaken an empirical analysis of why people punish themselves—a seemingly self-defeating but not altogether uncommon phenomenon (Klonsky, 2011; Nock, 2009; Nyström & Mikkelsen, 2013; Yelsma et al., 2002). The central research question I addressed is: Can self-punishment redeem one's morality? The broad answer is yes, in several ways. I showed that self-punishment resolves the threat to one's personal moral identity through one of two processes: moral repair or moral cleansing. As well as exploring the intrapersonal dimension of self-punishment, I also investigated its interpersonal consequences. I found that self-punishment can neutralise the threat to one's public moral identity by addressing third parties' justice concerns. Understanding these processes allows us to determine when and why self-punishment achieves, in people's minds, moral absolution.

In the following chapter I first review the major insights from my research in the context of the existing literature, noting the unique contributions of my research. I then comment on the implications of my research for theory and practice, and suggest some avenues for further exploration. Finally, I note some limitations of my research, before drawing a final conclusion.

Insights and Contributions

The research presented in this thesis investigated the implications of self-punishment for the self-punisher as well as for third-party observers, and correspondingly speaks to the intrapersonal and interpersonal functions of self-punishment. The findings contribute to the self-punishment literature by highlighting the multiple functions of self-punishment, as well as complement related literature on justice, moral emotions, and moral cognition. Self-punishment may be one of many ways to resolve a threat to moral identity, but it carries unique implications for intrapersonal and interpersonal repair.

Intrapersonal Functions of Self-Punishment: Psychological Defence and Repair

To date there are no detailed psychological explanations of the ways in which self-punishment might provide the sense of "moral purification" frequently intimated by historical and fictional accounts of self-punishment—or if it is even able to do so. Indeed, my qualitative analysis of self-punisher narratives (Study 2) revealed that self-punishment is aimed at addressing the psychological distress elicited by a perceived wrongdoing. In this thesis, however, I demonstrated that the path to moral redemption is not straightforward. I developed an identity-regulating model of self-punishment (Chapter 4) that draws on the notion that maintaining a moral identity—and a positive sense of self more broadly—is key to understanding responses to immoral behaviour (Aquino & Reed, 2002; Blasi, 1993; J. M. Carlsmith & Gross, 1969; Gino et al., 2015; Sachdeva et al., 2009; J. L. Tracy & Robins, 2004). The model delineates two processes that attempt to address one's sense of immorality: *moral cleansing* and *moral repair*.

Moral cleansing builds on recent empirical studies finding that self-punishment can reduce guilt (Bastian et al., 2011; Inbar et al., 2013). Yet, a reduction in guilt is but one aspect of a process of moral cleansing. I extended this effect to describe the cognitive and behavioural outcomes of self-punishment, demonstrating over a series of studies using various methodological paradigms (Studies 4.1, 4.2, and 4.3) that self-punishment can indeed resolve the threat to moral identity by "wiping the slate clean" (S. W. Lee & Schwarz, 2011). In this way, self-punishment as moral cleansing can buffer individuals against negative feedback, restoring their identity; however, this distances self-punishers from their transgression—undermining a sense of empathy towards the victim and reducing the willingness to engage in reparatory action.

The identity-regulating model outlines a second intrapersonal process yet to be articulated in the self-punishment literature: the notion that self-punishment can be an attempt

to directly confront, rather than excuse, one's threatened moral identity. My qualitative analysis of self-punisher accounts (Study 2) indicated that self-punishment provides an opportunity to reflect on the wrongdoing and affirm the wrongness of that act. Though this might be a difficult process, it might allow self-punishers to feel morally restored. This second function of self-punishment complements literature indicating that psychological self-punishment might perpetuate feelings of psychological distress (Dyer et al., 2017; Fisher & Exline, 2006; Whelton & Greenberg, 2005). I found evidence for this second function in experimental studies in the laboratory (Studies 4.1, 4.2, and 4.3), showing that such a process can facilitate moral engagement, victim empathy, and reconciliation.

I proposed that acknowledgment of a moral need determines which self-punishment process is employed. To the extent that transgressors interpreted their wrongdoing as constituting a significant threat to their moral identity, they were more likely to shift away from a moral cleansing to a moral repair mode (Studies 4.1, 4.2, and 4.3). While there was some evidence that objective features of the transgression context (i.e., severity of transgression and relationship to victim) contributed to moral need acknowledgment, these features did not appear to be causally responsible for it—at least, not relationship closeness (Study 4.4). Thus, though there may be some truth in the conjecture that less serious transgressions are more likely to result in simple psychological defences (Tetlock et al., 2000), this might be less to do with objective transgression features—but rather differences in the ways that individuals tend to interpret and respond to moral threats.

The failure to observe the predicted effects in Study 4.4 thus suggest a role for trait-level factors in self-punishment processes. In line with this argument, self-punishers expressed that self-punishment often felt automatic, unconscious, and inevitable; perhaps suggesting there is a self-punitive personality (Chapter 2). Moreover, perpetrator sensitivity (the tendency to see oneself as a perpetrator of injustice) was related to how moral threats

were interpreted and whether moral need was acknowledged (Study 4.4). However, there was only weak empirical evidence that this trait moderated the effects—and meaning—of self-punishment (Study 4.4). Nevertheless, a role for traits would be broadly in line with van Bunderen and Bastian's (2014) study showing differential effects of self-punishment on victim compensation depending on trait victim sensitivity. Therefore my research gives some credence to the role of traits in determining how self-punishment is employed, and warrants further exploration.

Viewing self-punishment as an act of both avoidance and confrontation dovetails with research pointing to the two sides of moral emotions, and how both shame and guilt can lead to reparative and defensive responses (Cibich et al., 2016; Gausel & Leach, 2011; Schmader & Lickel, 2006). Moreover, the moderation by moral need in the present model is consistent with research indicating that acknowledging shame is important in prompting constructive processes in the context of moral transgressions (Ahmed & Braithwaite, 2004; Woodyatt & Wenzel, 2014). My research also mirrors clinical observations of self-critics, which have indicated that self-criticism can be both a form of self-correction—stopping oneself from making mistakes and being alert to errors—as well as an attempt to rid the self of bad aspects (P. Gilbert et al., 2004; P. Gilbert & Procter, 2006).

More broadly, my model parallels research on self-forgiveness, which has been argued to have two pathways similar to those identified in the present research (Woodyatt et al., 2017). In that line of research, affirming the violated values facilitated moral engagement, self-forgiveness, and reconciliation—while affirming alternate qualities (i.e., a process of "pseudo self-forgiveness"; side-stepping the moral threat) did not (Woodyatt & Wenzel, 2014; Woodyatt et al., 2017). Thus, self-punishment—like self-forgiveness—serves more than one aim, and can be understood as one of many strategies transgressors have at their disposal to attempt to resolve the threat to their moral identity.

Interpersonal Functions of Self-Punishment: Restoring Justice

In this thesis I also examined the function of self-punishment from a third party perspective (Chapter 5). Across three experimental studies using hypothetical transgressions, I found that self-punishment could earn third parties' forgiveness by diminishing the perpetrator's status/power, and by communicating a value consensus. These results validate the utility of the symbolic justice framework (Okimoto & Wenzel, 2008) as applied to self-punishment. Moreover, findings are consistent with the argument put forth by researchers that self-punishment has an interpersonal function beyond the intrapersonal experience (Allen et al., 2015; Nelissen, 2012; Tanaka et al., 2016; Tanaka et al., 2015; Zhu et al., 2017).

While self-punishers themselves may be preoccupied with restoring their moral identity, third parties may be more concerned with notions of justice. However, there is an important overlap between these models. In particular, to the extent that self-punishers were considered to be thinking through and confronting their wrongdoing, value consensus was achieved (Study 5.3). This finding implies that self-punishers (at least, those motivated by moral repair) and third parties converge on the belief that revision of, and recommitment to, values is an important part of self-punishment. In this sense, self-punishment might be conceptualised less as a purely retributive measure (as a backwards-facing notion of deservingness; K. M. Carlsmith, 2006), but rather, also as a restorative and forward-looking measure that enhances adherence to shared values and ultimately reintegrates the perpetrator (Morris, 1981). This sets self-punishment apart somewhat from punishment inflicted by others; the latter "is unlikely to be an effective means for re-establishing value consensus" (Wenzel et al., 2008, p. 382). In contrast, through the voluntary act of punishing oneself, self-punishers can show that they accept the wrongness of the act, and in doing so, restore both their private and public moral identity.

Implications, Future Directions, and Limitations

In short, I proposed that self-punishment is an attempt to resolve the threat to one's moral identity, and can facilitate forgiveness from others. That is, self-punishment is seen by transgressors and third parties as a legitimate means to restore moral order. As already discussed, it remains unknown whether victims themselves would be equally satisfied by self-punishment; but there are certainly ways in which self-punishment could address victims' needs (see Chapter 5 General Discussion). It is conceivable that self-punishment could contribute to offenders', victims', and third parties' symbolic healing following a transgression.

Implications for Theory

However, what about the potential harms of self-punishment? Is self-punishment necessarily constructive, or can it be destructive? To some extent, it is both; there is a tension in self-punishment. In this thesis I have generally approached the topic from the perspective of self-punishment as functional: as a way for transgressors to achieve a goal—that of morally redeeming themselves, both in their own eyes and in those of others. This behaviour may have evolved to be functional under certain conditions, but this does not mean that it is always the case. People desire their own punishment, and others welcome it, yet there are times when self-punishment may be dysfunctional, both directly to the self as well as taking a toll on one's relationships. I now consider some of the costs of self-punishment.

First, and most obviously, there are some forms or expressions of self-punishment that go beyond a simple self-reprimand or a short period of social isolation; self-punishment can be extremely destructive in its more extreme forms (e.g., self-harm; Klonsky, 2011; Nock, 2009). Despite some of the benefits implicated in the current research (e.g., moral repair can facilitate reconciliation), one should not forget that self-punishment can have a significant toll on self-punishers' health and well-being. Likewise, self-punishment can directly endanger

and hurt social relations—such as when self-imposed social isolation might prevent the restoration of a fulfilling relationship.

Furthermore, one can speculate about other circumstances in which self-punishment might not be conducive to social harmony. Specifically, there may be interpersonal costs involved in self-punishing beyond the initial self-punishment exercise—even if one has secured self-regard and forgiveness in the short term. I will briefly comment on how both moral cleansing and moral repair might become maladaptive in the long term.

Moral cleansing might provide relief to the transgressor by avoiding distress, but in bypassing processing of the wrongdoing it may be difficult to learn from the wrongdoing and to avoid repeating it in future. And indeed, if observers believe self-punishers are committed to a shared morality when this is not the case, then perhaps it would be unwise for victims to be so quick to "wipe the slate clean." In other words, self-punishment might provide poor predictive information about the self-punisher's values and motivations. Observers may become frustrated at a person who apologises profusely for doing something wrong, only to repeat a similar offence a week later. In this manner, self-punishment could mislead others, sending a "false signal" about future moral behaviour. This could backfire on self-punishers, much like hypocrites who inspire moral outrage (Jordan, Sommers, Bloom, & Rand, 2017).

Another negative implication of moral cleansing is the possibility that self-punishment not only works as a backward-facing cleansing of a past wrong, but it may also prospectively protect the self-punisher from future moral threats (see Cramwinckel et al., 2013; Zhong, Strejcek, & Sivanathan, 2010). In this sense self-punishment could have a moral licensing effect (Merritt, Effron, & Monin, 2010; Monin & Miller, 2001) on future immoral behaviour: Since perpetrators' moral identity has been restored through self-punishment, they may feel licensed to do something immoral. This could involve an insidious spiral of immoral behaviour leading to self-punishment as moral cleansing, but then

justifying new immoral behaviour. Thus, if self-punishment is merely a way to reinstate one's moral image (or "moral credentials"; Monin & Miller, 2001), it may not always facilitate future moral behaviour.

On the other hand, I have argued that moral repair allows transgressors to come to terms with their wrongdoing, and motivates them to make amends and behave better in the future. In this sense, it might be easy to paint moral repair as a functional, prosocial process. But it may not always be an optimal strategy. Moral repair is likely to be functional to the extent that it motivates transgressors to identify and neutralise the threat to their moral standing. If a transgressor is unable to let go of their wrongdoing and becomes trapped in a cycle of self-condemnation (i.e., a state of perpetual threat), this may not be conducive to psychological wellbeing for themselves or for their relationship partners over time. In line with this, an observational study showed that the more that offending partners condemned themselves, the less relationship satisfaction both they and their partners reported (Pelucchi et al., 2013). Empirical designs such as a longitudinal dyad study might be able to capture potentially negative consequences of self-punishment in the long term.

Implications for Practice

My research also has significant implications for clinicians working with self-punitive individuals. First, clinicians may wish to carefully assess whether the source of a client's self-condemnation comes from a genuine moral need, or whether it might be misplaced. For example, researchers have suggested that experiencing trauma is related to guilt, shame, self-criticism and self-harm (Cox, MacPherson, Enns, & McWilliams, 2004; Glassman, Weierich, Hooley, Deliberto, & Nock, 2007; Stotz, Elbert, Müller, & Schauer, 2015). In these cases, the source of the guilt or shame is difficult to resolve; often it is the self who has been victimised rather than being guilty of a transgression in the traditional sense. Though one might feel guilty of some charge, there is no simple way to resolve these feelings of immorality through

self-punishment. How does one go about pinpointing their wrongdoing and the ways they can make things right again? To whom should one make amends? The functional aspects of moral repair may break down in these circumstances. Thus, self-punishment, even as an attempt at moral repair, may become counterproductive and circular as individuals are unable to resolve the perceived threat to their moral identity and move forward.

Clinicians have suggested that individuals carrying unresolvable guilt or shame might benefit from compassion-focused therapies that promote self-acceptance (P. Gilbert, 2010; P. Gilbert & Procter, 2006; Neff, 2003). Rather than focusing on the self as a perpetrator and *acknowledging* a moral need, in these situations—where there is a misplaced perception of a moral threat—it may be more tenable to *challenge* one's guilt, and to target irrational or unhelpful appraisals (e.g., that the individual could have prevented or controlled the perceived failure, that he or she is truly responsible for it, etc.).

However, self-compassion is not without its limits. For one, one must be careful in determining that self-compassion is the optimal strategy; clients may be denying their wrongdoing or the difficult steps they suspect they need to take to remedy the situation. Where there is scope for individuals to acknowledge a wrongdoing, self-compassion therapies may undermine reconciliation and future moral behaviour by short-cutting moral engagement (Woodyatt et al., 2017)—mirroring moral cleansing. The notion that self-compassion can be counterproductive was recently examined in the context of the most serious transgressions of all: criminal offences. Criminal offenders who tended to accept themselves, reserving self-judgment and self-criticism (a core element of mindfulness-based compassion therapies) were at higher risk of reoffending. As argued by the study authors, some scrutiny of one's thoughts, feelings, and behaviours may be adaptive; "components of mindfulness interventions may exacerbate patterns of criminal thinking, such as blaming others for their behavior by allowing people to not address their shortcomings and to

rationalize their behavior" (Tangney, Dobbins, Stuewig, & Schrader, 2017, pp. 6–7). Therefore, clinicians may wish to tread carefully in recommending self-compassion for all self-punitive individuals.

Moreover, some self-punitive individuals may be resistant to self-compassion therapies. In an attempt to eliminate self-punitive attacks, a loving-kindness meditation program was trialled for a group of self-critics, but with limited success (Shahar et al., 2015). These results are perhaps not surprising if self-punishment is a conditioned (and "deserved") coping strategy individuals use to deal with perceived moral failure. What does it mean to individuals to *not* punish themselves? Being encouraged to be self-compassionate could suggest to participants that they do not need to respond to a perceived threat, which is incompatible with motivations underlying self-punishment (especially moral repair). Thus, it may not be enough to ask self-punishers to direct feelings of kindness and warmth to themselves—a demand that some could find frightening (Gilbert & Procter 2006). Rather, clinicians could help clients to recognise the psychological needs that self-punishment serves (e.g., the need to resolve a threat to moral identity), identify the source of the threat, and break down the idea that self-punishment is the best way to deal with distress (in line with Gilbert and colleagues' take on compassionate mind training for self-critics; Gilbert & Procter 2006).

More broadly speaking, Gilbert's approach recognises an important fact: that there are underlying needs and motivations driving self-criticism that should be identified and acknowledged, rather than denied or ignored. Even painful experiences can have functional purposes (Bastian et al., 2014). In line with an evolutionary perspective, my research argues that self-punishment is potentially part of a system that has evolved to deal with moral threats. Self-punishment can avoid the more costly psychological resources of dealing with a wrongdoing extensively (moral cleansing); at other times it can help perpetrators work

through their wrongdoing so as to avoid repeating it in future (moral repair). In the short-term, these actions can earn others' forgiveness. Self-punishment should thus be understood as a behaviour that evolved to serve a social-evolutionary function. The problem—the dysfunctional aspect of self-punishment—arises when it is applied to contexts in which it no longer serves that function effectively or where it brings with it other costs. For example, self-punishment may cease to be functional when individuals become over-sensitized to perceived moral threats, or when self-punishment comes with significant costs to one's wellbeing. Seeing (and appreciating) self-punishment for what it is, rather than labelling it as dysfunctional, may be an important first step for chronic self-punishers in making peace with themselves.

Limitations

The current research is not without its limitations. A key limitation is the ambiguity about how moral need acknowledgment is best operationalised, and what drives need acknowledgment. Specifically, the relationship closeness manipulation was unsuccessful in manipulating moral need (Study 4.4); furthermore, there were inconsistencies between Studies 4.1, 4.2, and 4.3 in terms of which moral need variables were responsible for the moderated effects (see Chapter 4 General Discussion). The inconsistencies in the findings cast some doubt over the interpretation of the results. Though my research provides support for both self-punishment processes of cleansing and repair, it is less clear what exactly is behind these divergent effects—what leads some people to acknowledge a moral need, and how can this be reliably measured? Therefore, the concept of moral need acknowledgment requires clarification if the model is to be more accurate in predicting self-punishment effects. As already discussed (see Chapter 4 General Discussion) there are multiple possible operationalisations of moral need acknowledgment (e.g., acknowledgment versus denial; trait versus state), all of which should be explored if we are to truly understand what motivates

perpetrators to use self-punishment in such disparate ways. A clearer operationalisation would also contribute to the ongoing debate in the theoretical literature on which aspects of moral threat are the key drivers of behavioural responses to such threats (Leary, 2004; Shnabel & Nadler, 2008; J. L. Tracy & Robins, 2004).

Second, it should be noted that all the research contained in this thesis was conducted with Australian community members and students. Although it appears that self-punishment may be ubiquitous from historical and religious accounts, different cultures may view it through a different lens, which might limit the generalisability of the current findings. Cross-cultural researchers have noted the diverse ways identity and morality are conceptualised across cultures, and how this influences individuals' interpretation and experience of guilt and shame (Bedford & Hwang, 2003; Wong & Tsai, 2007). Moreover, the meaning of punishment and preferences for justice procedures vary between cultural contexts (Deater-Deckard, Dodge, & Sorbring, 2005; Y.-T. Lee, Ottati, Bornman, & Yang, 2011; Melossi, 2001). Such differences are likely to influence motivations for self-punishment, its effects, and perceptions of self-punishers. Therefore, the research reported in this thesis can only genuinely represent the sample used to produce it, that is, a predominantly Western perspective of self-punishment.

A final limitation concerns methodology. There are methodological quirks rising from the nature of self-punishment that may bias or obscure research findings—a charge from which my research is not immune. I explored these issues in depth (Chapter 2), and attempted to devise some criteria and tasks that would allow experimental researchers to surmount some of these obstacles (Chapter 3). Certainly, the use of qualitative methodology (Study 2) yielded insights that had been previously overlooked in the self-punishment literature, which helped to guide predictions and interpretation of the experimental effects in Chapter 4. Of course, it is possible that despite my best efforts to develop a neutral self-punishment task,

participants nonetheless interpreted the auditory task as something else entirely. The potential gap between self-punishment accounts inside versus outside the laboratory has profound implications for researchers seeking to explore this topic.

There is a sense that empirical testing for specific effects of self-punishment has preceded substantive observational research that could generate theory and hypotheses. It is clear from my qualitative analysis in Chapter 2 that there is a wealth of data one could draw upon, in terms of exploring the meanings of self-punishment and constructing more detailed theoretical models based on these observations. As an analogue, the self-harm literature has made much use of qualitative explorations of self-harm accounts. For example, Baker and Fortune (2008) challenged the idea of self-harm websites as a source of risk—as might be suggested by merely examining quantitative statistics on the effects of such websites. In their qualitative research they identified that self-harmers constructed these websites as sources of empathy and understanding, as communities, and as a way of coping. Likewise, if researchers were to more deeply interrogate the ways self-punishers construe their behaviour, we may gain additional insights regarding the functionality of self-punishment, as well as potential avenues to address harmful expressions of it (see also Sinclair & Green, 2005). The existing experimental literature on self-punishment would be well complemented—and indeed, strengthened—by additional grounded data that can contextualise experimental effects.

Conclusion

People have long been punishing themselves as atonement for their perceived misdeeds. The current thesis contributes to the literature by offering a novel model of how self-punishment provides this sense of moral absolution. Bridging lay perceptions with experimental findings, the current research proposed that self-punishment can be understood as a process through which transgressors attempt to resolve the threat to their moral identity. Two processes, in fact: Self-punishment can either cleanse or repair one's moral identity.

These processes tie in with observers' beliefs about justice, and this has implications for ongoing relations between offenders and victims. This research has presented important considerations for researchers in the field, and has provided valuable insights for clinicians wishing to help those struggling with problematic expressions of self-punishment. Notwithstanding its limitations, this thesis contributes a rich and timely analysis of how self-punishment resolves the threat to moral identity—and the cognitive, emotional, and interpersonal implications of these processes.

Appendix A

Coding Frameworks for Thematic Analysis (Study 1)

BLOCK 1: Motivation for Own Self-Punishment

Themes	
Final codes	Initial codes

1. Emotion regulation	- Emotion manipulation
	- Avoidance of emotion
	- Distraction
	- Emotion, e.g.:
	  o Guilt/shame
	  o Anger
	  o Frustration

2. Learning opportunity	- Reflect on self or behaviour - Accept responsibility - Learn something - Change attitude or behaviour
3. Normalised	- Automatic - Habit / natural reaction - No choice or alternative - Deservingness - Inevitable
Non classifiable categories	
5. Other Coherent response that does not fit themes	- Other
6. Don't know	- Don't know - Unsure
7. Because transgressed No reason why beyond transgression	- Transgression - Did something wrong
8. Unclear Does not provide a coherent response	- Describes self-punishment - Irrelevant / unclear

Example response coding:

im not sure i just did naturally.

Initial coding: Natural reaction. Final coding: Normalised.

BLOCK 2: Intrapersonal Outcomes of Own Self-Punishment

Themes	
Final codes	Initial codes
1. Increased negativity *Only applies to negative feelings/thoughts where it is clear that it is a result of the SP*	- Increased -ve emotionality - Increased -ve self-view - Rumination
2. Learning opportunity	- Reflect on self or behaviour - Accept responsibility - Learn something - Change attitude or behaviour
3. Felt better (in non moral way)	- Felt better - More +ve emotions/self-view
4. Restoration of morality	- Redemption/atonement - Justice restored - Forgiveness - Better/moral person
5. No change	- No change - No improvement
Non classifiable categories	
6. Other/unclear Coherent response that does not fit themes OR incoherent or irrelevant response	- Other - Unrelated thoughts/feelings - Unclear
7. Negativity only *Note: Response mentions negativity but does not specify that it is a result of the SP*	- Bad person - Guilty - Sad etc.

Example response coding:

[My self-punishment] *made me feel good that I had the willpower to step back from doing something that was clearly hurting others. I was proud of myself. It definitely changed me for the better.*

Initial coding: +ve emotion/self-view (pride), behaviour change, better person.

Final coding: Emotion regulation, Learning, Moral Restoration.

BLOCK 3: Interpersonal Outcomes of Own Self-Punishment

1. Negative outcomes	- Weakened/lost relationships
	- Negative perceptions by others
2. Positive outcomes	- Stronger/gained relationships
	- Support
	- Sympathy
	- Reconciliation
3. No effect	- No effect or change
No change at all to relationships	- Nothing
	- N/A
4. Other/unclear	- Mixed results without clear effect
Does not fit above categories	- Irrelevant response
	- Unclear

Example response coding:

I turned on others as well as myself and relationships were tarnished.

Initial coding: Weakened relationships. Final coding: Negative.

BLOCK 4: General Perceptions of Self-Punishment

Themes	
Final codes	Initial codes

1. Emotion regulation	- Emotion manipulation
	- Avoidance of emotion
	- Distraction/avoidance
	- Emotion, e.g.:
	  o Guilt/shame
	  o Anger
	  o Frustration

2. Learning opportunity	- Reflect on self or behaviour - Accept responsibility - Learn something - Change attitude or behaviour
3. Normalised	- Automatic - Habit / natural reaction - No choice or alternative - Deservingness - Inevitable
4. Restoration of morality	- Redemption/atonement - Justice restored - Forgiveness - Better/moral person
5. For others	- Attention seeking - Support seeking
Non classifiable categories	
6. Other Coherent response that does not fit themes	- Other
7. Don't know	- Don't know - Unsure
8. Unclear Does not provide a coherent response	- Irrelevant - Unclear

Appendix B

Additional Tables

Table B1

Pearson Correlations Between Outcome Variables (Study 4.1)

Measure	1	2	3	4	5	6	7	8
1 Moral identity threat	1							
2 Guilt	.24*	1						
3 Shame	.25*	.83**	1					
4 Responsibility	-.04	.45**	.48**	1				
5 Empathy	-.01	.36*	.31*	.47**	1			
6 Values	-.27*	.27*	.37**	.42**	.33*	1		
7 Moral engagement	-.03	.49**	.52**	.39**	.29*	.47**	1	
8 Reparation	.15	.65**	.71**	.54**	.41**	.68**	.70**	1

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

Table B2

Pearson Correlations Between Outcome Variables (Study 4.2)

Measure	1	2	3	4	5	6	7	8
1 Guilt	1							
2 Shame	.82**	1						
3 Responsibility	.32*	.29*	1					
4 Empathy	.15	.20	.41**	1				
5 Values	.05	.01	.41**	.33*	1			
6 Moral engagement	.26*	.14	.55**	.43**	.44**	1		
7 Moral disengagement	-.19	-.12	-.41**	-.06	-.17	-.24*	1	
8 Reparation	.22*	.13	.47**	.62**	.34**	.60**	-.36**	1

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

Table B3

Pearson Correlations Between Outcome Variables (Study 4.3)

Measure	1	2	3	4	5	6	7	8
1 Guilt	1							
2 Shame	.75**	1						
3 Responsibility	.23*	.19	1					
4 Empathy	.20	.08	.25*	1				
5 Values	.01	.07	.24*	.00	1			
6 Moral engagement	.19	.25*	.33**	.27*	.39**	1		
7 Moral disengagement	-.21*	-.10	-.42**	-.40**	-.13	-.35**	1	
8 Reparation	.34**	.29*	.25*	.48**	.25*	.35**	-.44**	1

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

Table B4

Correlations between Dependent Variables (Study 5.1)

Measure	1	2	3	4	5	6	7	8
1 Values	1							
2 Status/power	-.25*	1						
3 Justice	.36**	-.22*	1					
4 Remorse	.55**	-.44**	.46**	1				
5 Forgiveness	.41**	-.29**	.32**	.39**	1			
6 Trust	.50**	-.38**	.38**	.46**	.60**	1		
7 Moral standing	.18*	-.20*	.11	.19*	.48**	.39**	1	
8 Liking	.50**	-.21*	.28**	.43**	.52**	.63**	.42**	1

Note. $N = 150$.

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

Table B5

Correlations between Dependent Variables (Study 5.2)

Measure	1	2	3	4	5	6	7	8	9	10
1 Values	1									
2 Status/power	-.38**	1								
3 Justice	.64**	-.32**	1							
4 Remorse	.52**	-.34**	.56**	1						
5 Forgiveness	.60**	-.23*	.57**	.58**	1					
6 Trust	.68**	-.29**	.69**	.56**	.74**	1				
7 Moral standing	.39**	-.24*	.30**	.38**	.67**	.57**	1			
8 Liking	.56**	-.22*	.53**	.43**	.73**	.76**	.69**	1		
9 Moral cleansing	-.22*	.07	-.47**	-.13	-.19	-.23*	-.01	-.14	1	
10 Moral repair	.68**	-.21*	.73**	.61**	.63**	.58**	.42**	.56**	-.36**	1

Note. $N = 150$, except rows (9) and (10) for which $n = 99$.

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

Table B6

Correlations Between Dependent Variables (Study 5.3)

Measure	1	2	3	4	5	6	7	8	9	10	11
1 Values	1										
2 Status/power (offender)	-17*	1									
3 Status/power (victim)	.04	-.34**	1								
4 Justice	.50**	-.06	-.03	1							
5 Remorse	.59**	-.33**	.18*	.31**	1						
6 Forgiveness	.28**	-.12	.20*	.24*	.30**	1					
7 Trust	.54**	-.28**	.10	.34**	.47**	.51**	1				
8 Moral standing	.32**	-.23*	.18*	.21*	.33**	.62**	.54**	1			
9 Liking	.42**	-.31**	.18*	.28**	.36**	.59**	.68**	.65**	1		
10 Moral cleansing	-.24*	-.18*	.05	-.32**	-.08	.01	-.06	.10	.01	1	
11 Moral repair	.46**	-.14	.002	.42**	.49**	.32**	.41**	.21*	.30**	-.28**	1

Note. $N = 200$.

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

Appendix C

Moderation Analyses: Interaction Terms for Tested Moderations (Chapter 4)

Independent variable: experimental condition (no self-punishment vs. self-punishment)

Study 4.1

Moderator	Outcome measure	*B*	*SE*	*p*
Moral identity threat	Guilt	0.46	0.26	.080
	Shame	0.67	0.24	.007
	Responsibility	0.65	0.23	.006
	Empathy	0.11	0.17	.536
	Values	0.30	0.14	.035
	Moral engagement	0.05	0.13	.691
	Reparation	0.33	0.14	.024
	Excluded variables			
	Justice restoration	-0.23	0.16	.144
	Avoidance	0.22	0.22	.319

Study 4.2

Moderator	Outcome measure	*B*	*SE*	*p*
Severity	Guilt	0.06	0.15	.672
	Shame	0.24	0.13	.079
	Responsibility	0.20	0.11	.073
	Empathy	0.00	0.14	.981
	Values	0.04	0.08	.620
	Moral engagement	0.18	0.08	.035
	Moral disengagement	-0.20	0.10	.039
	Reparation	0.15	0.13	.257
Relationship	Guilt	-0.15	0.11	.158
	Shame	-0.08	0.10	.398
	Responsibility	0.07	0.08	.356
	Empathy	0.05	0.08	.553
	Values	0.10	0.05	.060
	Moral engagement	0.13	0.05	.017

Moderator	Outcome measure	*B*	*SE*	*p*
	Moral disengagement	-0.12	0.07	.063
	Reparation	0.03	0.06	.612
Pre-guilt	Guilt	-0.03	0.12	.812
	Shame	0.11	0.11	.319
	Responsibility	0.18	0.10	.076
	Empathy	0.11	0.11	.345
	Values	-0.01	0.07	.917
	Moral engagement	0.04	0.07	.594
	Moral disengagement	-0.11	0.08	.171
	Reparation	0.04	0.10	.684
Pre-shame	Guilt	-0.10	0.10	.354
	Shame	0.05	0.09	.570
	Responsibility	0.10	0.08	.233
	Empathy	0.02	0.09	.798
	Values	0.01	0.06	.811
	Moral engagement	0.02	0.06	.759
	Moral disengagement	-0.04	0.07	.522
	Reparation	-0.10	0.08	.214
Moral identity threat	Guilt	-0.06	0.22	.777
	Shame	-0.13	0.19	.500
	Responsibility	0.17	0.17	.314
	Empathy	-0.10	0.20	.632
	Values	-0.03	0.11	.765
	Moral engagement	0.05	0.13	.691
	Moral disengagement	-0.13	0.14	.364
	Reparation	0.06	0.19	.746

Study 4.3

Moderator	Outcome measure	*B*	*SE*	*p*
Severity	Guilt	0.19	0.11	.075
	Shame	0.12	0.11	.277
	Responsibility	0.17	0.08	.031
	Empathy	0.06	0.10	.560

Moderator	Outcome measure	*B*	*SE*	*p*
	Values	-0.05	0.04	.302
	Moral engagement	-0.05	0.07	.469
	Moral disengagement	-0.05	0.08	.533
	Reparation	-0.02	0.09	.852
Relationship	Guilt	0.06	0.08	.481
	Shame	0.12	0.08	.148
	Responsibility	0.09	0.07	.187
	Empathy	-0.09	0.07	.251
	Values	0.03	0.03	.443
	Moral engagement	0.00	0.06	.952
	Moral disengagement	-0.01	0.06	.941
	Reparation	0.04	0.05	.472
Pre-guilt	Guilt	-0.07	0.09	.437
	Shame	-0.08	0.09	.417
	Responsibility	0.09	0.07	.176
	Empathy	0.04	0.08	.652
	Values	-0.04	0.04	.280
	Moral engagement	-0.01	0.06	.879
	Moral disengagement	-0.07	0.06	.308
	Reparation	-0.05	0.08	.511
Pre-shame	Guilt	0.07	0.09	.443
	Shame	-0.11	0.09	.209
	Responsibility	0.09	0.06	.165
	Empathy	0.18	0.07	.019
	Values	-0.05	0.04	.177
	Moral engagement	0.03	0.06	.614
	Moral disengagement	-0.12	0.06	.028
	Reparation	-0.05	0.07	.497
Moral identity threat	Guilt	0.30	0.16	.070
	Shame	0.08	0.16	.616
	Responsibility	0.15	0.13	.266
	Empathy	0.31	0.14	.033
	Values	0.01	0.06	.888

Moderator	Outcome measure	*B*	*SE*	*p*
	Moral engagement	-0.12	0.12	.333
	Moral disengagement	-0.21	0.12	.091
	Reparation	0.11	0.14	.462

Study 4.4

Moderator	Outcome measure	*B*	*SE*	*p*
Perpetrator sensitivity	Guilt	0.26	0.18	.145
	Shame	0.15	0.17	.379
	Responsibility	0.24	0.15	.113
	Empathy	0.11	0.14	.436
	Values	0.22	0.14	.128
	Moral engagement	0.25	0.10	.016
	Moral disengagement	-0.07	0.13	.575
	Reparation	-0.13	0.17	.439

References

Aarts, H., & Dijksterhuis, A. (2000). The automatic activation of goal-directed behaviour: The case of travel habit. *Journal of Environmental Psychology, 20*(1), 75–82. doi:10.1006/jevp.1999.0156

Abramowitz, J. S., Tolin, D. F., & Street, G. P. (2001). Paradoxical effects of thought suppression: A meta-analysis of controlled studies. *Clinical Psychology Review, 21*(5), 683–703. doi:10.1016/S0272-7358(00)00057-X

Adams, G. S. (2011). *Punishers become more deviant* (Unpublished doctoral dissertation). Stanford Graduate School of Business, Stanford, CA.

Ahmed, E., & Braithwaite, V. (2004). "What, me ashamed?" Shame management and school bullying. *Journal of Research in Crime and Delinquency, 41*(3), 269–294. doi:doi:10.1177/0022427804266547

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology, 20*(1), 1–48. doi:10.1080/10463280802613866

Allen, A. B., Barton, J., & Stevenson, O. (2015). Presenting a self-compassionate image after an interpersonal transgression. *Self and Identity, 14*(1), 33–50. doi:10.1080/15298868.2014.946958

Anderson, C. A., & Murphy, C. R. (2003). Violent video games and aggressive behavior in young women. *Aggressive Behavior, 29*(5), 423–429. doi:10.1002/ab.10042

Andrews, P. W., & Thomson, J. A., Jr. (2009). The bright side of being blue: depression as an adaptation for analyzing complex problems. *Psychological Review, 116*(3), 620–654. doi:10.1037/a0016242

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83*(6), 1423–1440. doi:10.1037/0022-3514.83.6.1423

Austin, W. (1979). The concept of desert and its influence on simulated decision makers' sentencing decisions. *Law and Human Behavior, 3*(3), 163–187. doi:10.1007/bf01039789

Austin, W., Walster, E., & Utne, M. K. (1976). Equity and the law: The effect of a harmdoer's "suffering in the act" on liking and assigned punishment. *Advances in Experimental Social Psychology, 9*, 163–190. doi:10.1016/s0065-2601(08)60061-1

Baird, H. M. (2004). *The Huguenots and Henry of Navarre* (Vol. 1). Eugene, OR: Wipf and Stock. (Original work published 1896)

Baker, D., & Fortune, S. (2008). Understanding self-harm and suicide websites. *Crisis, 29*(3), 118–122. doi:10.1027/0227-5910.29.3.118

Bandura, A. (1990). Selective activation and disengagement of moral control. *Journal of Social Issues, 46*(1), 27–46. doi:10.1111/j.1540-4560.1990.tb00270.x

Bandura, A. (1991). Social cognitive theory of moral thought and action. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development: Vol 1. Theory* (pp. 45–103). Hillsdale, NJ: Erlbaum.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*(3), 193–209. doi:10.1207/s15327957pspr0303_3

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325–344. doi:10.1007/s10683-006-7052-9

Bastian, B., Jetten, J., & Fasoli, F. (2011). Cleansing the soul by hurting the flesh: The guilt-reducing effect of pain. *Psychological Science, 22*(3), 334–335. doi:10.1177/0956797610397058

Bastian, B., Jetten, J., Hornsey, M. J., & Leknes, S. (2014). The positive consequences of pain: A biopsychosocial approach. *Personality and Social Psychology Review, 18*(3), 256–279. doi:10.1177/1088868314527831

Batson, C. D., Kennedy, C. L., Nord, L. A., Stocks, E., Fleming, D. Y. A., Marzette, C. M., . . . Zerger, T. (2007). Anger at unfairness: Is it moral outrage? *European Journal of Social Psychology, 37*(6), 1272–1285. doi:10.1002/ejsp.434

Baumeister, R. F. (1988). Masochism as escape from self. *Journal of Sex Research, 25*(1), 28–59. doi:10.1080/00224498809551444

Baumeister, R. F. (1997). *Evil: Inside human violence and cruelty*. New York, NY: W.H. Freeman.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*(3), 497–529. doi:10.1037//0033-2909.117.3.497

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: an interpersonal approach. *Psychological Bulletin, 115*(2), 243–267. doi:10.1037//0033-2909.115.2.243

Beard, E. (2013). *Public penance, public salvation: An exploration of the Black Death's influence on the Flagellant movement*. UCSC History Annals, Santa Cruz, CA.

Beard, M. (1994). The Roman and the foreign: The cult of the 'Great Mother' in Imperial Rome. In N. Thomas & C. Humphrey (Eds.), *Shamanism, history, and the state* (pp. 164–190). Ann Arbor, MI: University of Michigan Press.

Bedford, O., & Hwang, K.-K. (2003). Guilt and shame in Chinese culture: A cross-cultural framework from the perspective of morality and identity. *Journal for the Theory of Social Behaviour, 33*(2), 127–144. doi:10.1111/1468-5914.00210

Bibas, S., & Bierschbach, R. A. (2004). Integrating remorse and apology into criminal procedure. *Yale Law Journal*, 114(1), 85–148. doi:10.2307/4135717

Blasi, A. (1993). The development of identity: Some implications for moral functioning. In G. G. Noam & T. E. Wren (Eds.), *The moral self* (pp. 99–122). Cambridge, MA: MIT Press.

Boothby, J. L., Thorn, B. E., Overduin, L. Y., & Ward, L. C. (2004). Catastrophizing and perceived partner responses to pain. *Pain, 109*(3), 500–506. doi:10.1016/s0304-3959(04)00122-8

Brañas-Garza, P., Bucheli, M., Espinosa, M. P., & García-Muñoz, T. (2013). Moral cleansing and moral licenses: experimental evidence. *Economics and Philosophy, 29*(2), 199–212. doi:10.1017/s0266267113000199

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. doi:10.1191/1478088706qp063oa

Brewer, M. B., & Caporael, L. R. (2006). An evolutionary perspective on social identity: Revisiting groups. In J. S. M. Schaller, & D. & Kenrick (Eds.), *Evolution and social psychology* (pp. 143–161). New York, NY: Psychology Press.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54*(1), 106–148. doi:10.1111/j.1467-6494.1986.tb00391.x

Brown, S. A., Williams, K., & Collins, A. (2007). Past and recent deliberate self-harm: Emotion and coping strategy differences. *Journal of Clinical Psychology, 63*(9), 791–803. doi:10.1002/jclp.20380

Callan, M. J., Kay, A. C., & Dawtry, R. J. (2014). Making sense of misfortune: Deservingness, self-esteem, and patterns of self-defeat. *Journal of Personality and Social Psychology, 107*(1), 142–162. doi:10.1037/a0036640

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. doi:10.1037/h0046016

Carlsmith, J. M., & Gross, A. E. (1969). Some effects of guilt on compliance. *Journal of Personality and Social Psychology, 11*(3), 232–239. doi:10.1037/h0027039

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology, 42*(4), 437–451. doi:10.1016/j.jesp.2005.06.007

Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology, 95*(6), 1316–1324. doi:10.1037/a0012165

Carveth, D. L. (2006). Self-punishment as guilt evasion: Theoretical issues. *Canadian Journal of Psychoanalysis, 14*(2), 176–198. Retrieved from https://academicjournals.ca/index.php/cjp-rcp/

Cassin, S. E., & von Ranson, K. M. (2005). Personality and eating disorders: A decade in review. *Clinical Psychology Review, 25*(7), 895–916. doi:10.1016/j.cpr.2005.04.012

Chang, E. C., Watkins, A., & Banks, K. H. (2004). How adaptive and maladaptive perfectionism relate to positive and negative psychological functioning: Testing a stress-mediation model in black and white female college students. *Journal of Counseling Psychology, 51*(1), 93–102. doi:10.1037/0022-0167.51.1.93

Chapman, A. L., Gratz, K. L., & Brown, M. Z. (2006). Solving the puzzle of deliberate self-harm: The experiential avoidance model. *Behaviour Research and Therapy, 44*(3), 371–394. doi:10.1016/j.brat.2005.03.005

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology, 39*, 268–277. doi:10.1016/j.joep.2013.09.004

Cibich, M., Woodyatt, L., & Wenzel, M. (2016). Moving beyond "shame is bad": How a functional emotion can become problematic. *Social and Personality Psychology Compass, 10*(9), 471–483. doi:10.1111/spc3.12263

Cohen, T. R., Wolf, S. T., Panter, A. T., & Insko, C. A. (2011). Introducing the GASP scale: a new measure of guilt and shame proneness. *Journal of Personality and Social Psychology, 100*(5), 947–966. doi:10.1037/a0022641

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. doi:10.1037//0021-9010.78.1.98

Cox, B. J., MacPherson, P. S. R., Enns, M. W., & McWilliams, L. A. (2004). Neuroticism and self-criticism associated with posttraumatic stress disorder in a nationally representative sample. *Behaviour Research and Therapy, 42*(1), 105–114. doi:10.1016/S0005-7967(03)00105-0

Coyne, J. C. (1976). Depression and the response of others. *Journal of Abnormal Psychology, 85*(2), 186–193. doi:10.1037//0021-843x.85.2.186

Craig, K. D. (2009). The social communication model of pain. *Canadian Psychology/Psychologie Canadienne, 50*(1), 22–32. doi:10.1097/j.pain.0000000000000185

Cramwinckel, F. M., van Dijk, E., Scheepers, D., & van den Bos, K. (2013). The threat of moral refusers for one's self-concept and the protective function of physical cleansing. *Journal of Experimental Social Psychology, 49*(6), 1049–1058. doi:10.1016/j.jesp.2013.07.009

Creswell, J. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Cutcliffe, J. R. (1998). Hope, counselling and complicated bereavement reactions. *Journal of Advanced Nursing, 28*(4), 754–761. doi:10.1046/j.1365-2648.1998.00724.x

Darley, J. M. (2002). Just punishments: Research on retributional justice. In M. Ross & D. T. Miller (Ed.), *The justice motive in everyday life* (pp. 314–333). New York, NY: Cambridge University Press.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review, 7*(4), 324–336. doi:10.1207/s15327957pspr0704_05

De Hooge, I. E., Breugelmans, S. M., & Zeelenberg, M. (2008). Not so ugly after all: when shame acts as a commitment device. *Journal of Personality and Social Psychology, 95*(4), 933–943. doi:10.1037/a0011991

De Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2010). Restore and protect motivations following shame. *Cognition and Emotion, 24*(1), 111–127. doi:10.1080/02699930802584466

De Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2011). A functionalist account of shame-induced behaviour. *Cognition and Emotion, 25*(5), 939–946. doi:10.1080/02699931.2010.516909

Deater-Deckard, K., Dodge, K. A., & Sorbring, E. (2005). Cultural differences in the effects of physical punishment. In M. Rutter & M. Tienda (Eds.), *Ethnicity and causal mechanisms* (pp. 204–226). New York, NY: Cambridge University Press.

DeWall, C. N., Twenge, J. M., Koole, S. L., Baumeister, R. F., Marquez, A., & Reid, M. W. (2011). Automatic emotion regulation after social exclusion: Tuning to positivity. *Emotion, 11*(3), 623–636. doi:10.1037/a0023534

Dyer, K. F., Dorahy, M. J., Corry, M., Black, R., Matheson, L., Coles, H., . . . Middleton, W. (2017). Comparing shame in clinical and nonclinical populations: Preliminary findings. *Psychological Trauma: Theory, Research, Practice, and Policy, 9*(2), 173–180. doi:10.1037/tra0000158

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*(4), 479–504. doi:10.1521/jscp.2007.26.4.479

Exline, J. J., Root, B. L., Yadavalli, S., Martin, A. M., & Fisher, M. L. (2011). Reparative behaviors and self-forgiveness: Effects of a laboratory-based exercise. *Self and Identity, 10*(1), 101–126. doi:10.1080/15298861003669565

Fagius, J., & Wahren, L. K. (1981). Variability of sensory threshold determination in clinical use. *Journal of the Neurological Sciences, 51*(1), 11–27. doi:10.1016/0022-510x(81)90056-3

Fairburn, C. G., Shafran, R., & Cooper, Z. (1999). A cognitive behavioural theory of anorexia nervosa. *Behaviour Research and Therapy, 37*(1), 1–13. doi:10.1017/s1352465800018348

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. doi:10.4324/9780203127698

Favazza, A. R. (1996). *Bodies under siege: Self-mutilation and body modification in culture and psychiatry*. Baltimore, MD: John Hopkins University Press.

Fayard, J. V., Bassi, A. K., Bernstein, D. M., & Roberts, B. W. (2009). Is cleanliness next to godliness? Dispelling old wives' tales: Failure to replicate Zhong and Liljenquist (2006). *Journal of Articles in Support of the Null Hypothesis, 6*(2), 21–30. Retrieved from http://www.jasnh.com/

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63–87. doi:10.2139/ssrn.495443

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137–140. doi:10.1038/415137a

Finney, S. J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte, NC: Information Age.

Fisher, M. L., & Exline, J. J. (2006). Self-forgiveness versus excusing: The roles of remorse, effort, and acceptance of responsibility. *Self and Identity, 5*(2), 127–146. doi:10.1080/15298860600586123

Freud, S. (1957). Some character-types met with in psycho-analytic work. In J. Strachey (Ed. & Trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 14, pp. 311–333). London, United Kingdom: Hogarth Press. (Original work published 1916)

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin, 40*(8), 986–997. doi:10.1177/0146167214533130

Gausel, N., & Leach, C. W. (2011). Concern for self-image and social image in the management of moral failure: Rethinking shame. *European Journal of Social Psychology, 41*(4), 468–478. doi:10.1002/ejsp.803

Gausel, N., Leach, C. W., Vignoles, V. L., & Brown, R. (2012). Defend or repair? Explaining responses to in-group moral failure by disentangling feelings of shame, rejection, and inferiority. *Journal of Personality and Social Psychology, 102*(5), 941–960. doi:10.1037/a0027233

Gergen, K. J., & Wishnov, B. (1965). Others' self-evaluations and interaction anticipation as determinants of self-presentation. *Journal of Personality and Social Psychology, 2*(3), 348–358. doi:10.1037/h0022385

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998). Immune neglect: a source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 75(3), 617–638. doi:10.1037//0022-3514.75.3.617

Gilbert, P. (2000). Varieties of submissive behavior as forms of social defense: Their evolution and role in depression. In L. Sloman & P. Gilbert (Eds.), *Subordination and defeat: An evolutionary approach to mood disorders* (pp. 3–46). Hillsdale, NJ: Lawrence Erlbaum.

Gilbert, P. (2010). An introduction to compassion focused therapy in cognitive behavior therapy. *International Journal of Cognitive Therapy, 3*(2), 97–112. doi:10.1521/ijct.2010.3.2.97

Gilbert, P., Clarke, M., Hempel, S., Miles, J. N., & Irons, C. (2004). Criticizing and reassuring oneself: An exploration of forms, styles and reasons in female students. *British Journal of Clinical Psychology, 43*(1), 31–50. doi:10.1348/014466504772812959

Gilbert, P., & Procter, S. (2006). Compassionate mind training for people with high shame and self-criticism: Overview and pilot study of a group therapy approach. *Clinical Psychology and Psychotherapy, 13*(6), 353–379. doi:10.1002/cpp.507

Gino, F., Kouchaki, M., & Galinsky, A. D. (2015). The moral virtue of authenticity: How inauthenticity produces feelings of immorality and impurity. *Psychological Science, 26*(7), 983–996. doi:10.1177/0956797615575277

Glaser, B. G., & Strauss, A. L. (2012). The discovery of grounded theory: Strategies for qualitative research. New Brunswick, NJ: Aldine Transaction. (Original work published 1967)

Glassman, L. H., Weierich, M. R., Hooley, J. M., Deliberto, T. L., & Nock, M. K. (2007). Child maltreatment, non-suicidal self-injury, and the mediating role of self-criticism. *Behaviour Research and Therapy, 45*(10), 2483–2490. doi:10.1016/j.brat.2007.04.002

Goffman, E. (1955). On face-work: An analysis of ritual elements in social interaction. *Psychiatry, 18*(3), 213–231. doi:10.1080/00332747.1955.11023008

Gold, G. J., & Weiner, B. (2000). Remorse, confession, group identity, and expectancies about repeating a transgression. *Basic and Applied Social Psychology, 22*(4), 291–300. doi:10.1207/15324830051035992

Goldberg, J. H., Lerner, J. S., & Tetlock, P. E. (1999). Rage and reason: The psychology of the intuitive prosecutor. *European Journal of Social Psychology, 29*(5–6), 781–795. doi:10.1002/(sici)1099-0992(199908/09)29:5/6<781::aid-ejsp960>3.0.co;2-3

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology, 41*(3), 364–374. doi:10.1002/ejsp.782

Gollwitzer, M., Skitka, L. J., Wisneski, D., Sjöström, A., Liberman, P., Nazir, S. J., & Bushman, B. J. (2014). Vicarious revenge and the death of Osama bin Laden. *Personality and Social Psychology Bulletin, 40*(5), 604–616. doi:10.1093/acprof:osobl/9780199738663.003.0001

Gray, K., & Wegner, D. M. (2010). Torture and judgments of guilt. *Journal of Experimental Social Psychology, 46*(1), 233–235. doi:10.1016/j.jesp.2009.10.003

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*(3), 407–418. doi:10.1177/0146167207311534

Gromet, D. M. (2009). *Restoration and retribution: People's negotiation of multiple responses to wrongdoing* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (Order No. 3356714)

Gromet, D. M., Okimoto, T. G., Wenzel, M., & Darley, J. M. (2012). A victim-centered approach to justice? Victim satisfaction effects on third-party punishments. *Law and Human Behavior, 36*(5), 375–389. doi:10.1037/h0093922

Hagen, E. H., Watson, P. J., & Hammerstein, P. (2008). Gestures of despair and hope: A view on deliberate self-harm from economics and evolutionary biology. *Biological Theory, 3*(2), 123–138. doi:10.1162/biot.2008.3.2.123

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.) *Handbook of affective sciences* (pp. 852–870). Oxford, UK: Oxford University Press.

Hammersley, M. (2002). Ethnography and realism. In A. M. Huberman & M. B. Miles (Eds.), *The qualitative researcher's companion* (pp. 65–80). Thousand Oaks, CA: Sage.

Hareli, S., & Eisikovits, Z. (2006). The role of communicating social emotions accompanying apologies in forgiveness. *Motivation and Emotion, 30*(3), 189–197. doi:10.1007/s11031-006-9025-x

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford.

Henry, D. B., Dymnicki, A. B., Mohatt, N., Allen, J., & Kelly, J. G. (2015). Clustering methods with qualitative data: A mixed-methods approach for prevention research with small samples. *Prevention Science, 16*(7), 1007–1016. doi:10.1007/s11121-015-0561-z

Henry, D. B., Tolan, P. H., & Gorman-Smith, D. (2005). Cluster analysis in family psychology research. *Journal of Family Psychology, 19*(1), 121–132. doi:10.1037/0893-3200.19.1.121

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist, 52*(12), 1280–1300. doi:9780199765829.003.0010

Hiler, J. L. (2015). *The role of motives and decision rules in restaurant tipping* (Doctoral dissertation). Retrieved from http://digitalcommons.lsu.edu/

Holstein, J. A., & Gubrium, J. F. (1995). *The active interview*. Thousand Oaks, CA: Sage.

Inbar, Y., Pizarro, D. A., Gilovich, T., & Ariely, D. (2013). Moral masochism: On the connection between guilt and self-punishment. *Emotion, 13*(1), 14–18. doi:10.1037/a0029749

Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. In J. Suis (Ed.), *Psychological perspectives on the self* (Vol. 1, pp. 231–261). Hillsdale, NJ: Erlbaum.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science, 28*(3), 356–368. doi:doi:10.1177/0956797616685771

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136–153. doi:10.1037/0033-295X.93.2.136

Keltner, D., & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition and Emotion, 13*(5), 505–521. doi:10.1080/026999399379168

Kim, H.-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry and Endodontics, 38*(1), 52–54. doi:10.5395/rde.2013.38.1.52

Klonsky, E. D. (2007). The functions of deliberate self-injury: A review of the evidence. *Clinical Psychology Review, 27*(2), 226–239. doi:10.1016/j.cpr.2006.08.002

Klonsky, E. D. (2011). Non-suicidal self-injury in United States adults: prevalence, sociodemographics, topography and functions. *Psychological Medicine, 41*(9), 1981–1986. doi:10.1017/s0033291710002497

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science, 27*(3), 405–418. doi:10.1177/0956797615624469

Kriss, P. H., Weber, R. A., & Xiao, E. (2016). Turning a blind eye, but not the other cheek: On the robustness of costly punishment. *Journal of Economic Behavior and Organization*, 128, 159–177. doi:10.2139/ssrn.2549689

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*(2), 75–84. doi:j.evolhumbehav.2006.06.001

Leach, C. W., & Cidam, A. (2015). When is shame linked to constructive approach orientation? A meta-analysis. *Journal of Personality and Social Psychology, 109*(6), 983–1002. doi:10.1037/pspa0000037

Leary, M. R. (2004). Digging deeper: The fundamental nature of "self-conscious" emotions. *Psychological Inquiry, 15*(2), 129–131. Retrieved from http://www.tandfonline.com/toc/hpli20/current

Leary, M. R., & Baumeister, R. F. (2000). The nature and function of self-esteem: Sociometer theory. *Advances in Experimental Social Psychology, 32*, 1–62. doi:10.1016/s0065-2601(00)80003-9

Leary, M. R., Raimi, K. T., Jongman-Sereno, K. P., & Diebels, K. J. (2015). Distinguishing intrapsychic from interpersonal motives in psychological theory and research. *Perspectives on Psychological Science, 10*(4), 497–517. doi:10.1177/1745691615583132

Lee, S. W., & Schwarz, N. (2011). Wiping the slate clean: Psychological consequences of physical cleansing. *Current Directions in Psychological Science, 20*(5), 307–311. doi:10.1177/0963721411422694

Lee, Y.-T., Ottati, V., Bornman, E., & Yang, S. (2011). A cross-cultural investigation of beliefs about justice in China, USA and South Africa. *International Journal of Intercultural Relations, 35*(4), 511–521. doi:10.1016/j.ijintrel.2011.01.001

Lerner, M. J., & Miller, D. T. (1978). Just world research and the attribution process: Looking back and ahead. *Psychological Bulletin, 85*(5), 1030–1051. doi:10.1037//0033-2909.85.5.1030

Lewis, H. B. (1971). *Shame and guilt in neurosis*. New York, NY: International Universities Press.

Loewenstein, G., & Small, D. A. (2007). The Scarecrow and the Tin Man: The vicissitudes of human sympathy and caring. *Review of General Psychology, 11*(2), 112–126. doi:10.1037/1089-2680.11.2.112

Lotz, S., Okimoto, T. G., Schlösser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of Experimental Social Psychology, 47*(2), 477–480. doi:10.1016/j.jesp.2010.10.004

Luyten, P., Fontaine, J. R., & Corveleyn, J. (2002). Does the Test of Self-Conscious Affect (TOSCA) measure maladaptive aspects of guilt and adaptive aspects of shame? An empirical investigation. *Personality and Individual Differences, 33*(8), 1373–1387. doi:10.1016/s0191-8869(02)00197-6

Madill, A., Jordan, A., & Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *British Journal of Psychology, 91*(1), 1–20. doi:10.1348/000712600161646

Mallett, K. A., Bachrach, R. L., & Turrisi, R. (2009). Examining the unique influence of interpersonal and intrapersonal drinking perceptions on alcohol consumption among college students. *Journal of Studies on Alcohol and Drugs, 70*(2), 178–185. doi:10.15288/jsad.2009.70.178

Melossi, D. (2001). The cultural embeddedness of social control: Reflections on the comparison of Italian and North-American cultures concerning punishment. *Theoretical Criminology, 5*(4), 403–424. doi:10.1177/1362480601005004001

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass, 4*(5), 344–357. doi:10.1111/j.1751-9004.2010.00263.x

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology, 81*(1), 33–43. doi:10.1037//0022-3514.8I.I.33

Morris, H. (1981). A paternalistic theory of punishment. *American Philosophical Quarterly, 18*(4), 263–271. Retrieved from http://www.press.uillinois.edu/journals/apq.html

Murphy, J. G. (2007). Remorse, apology, and mercy. *Ohio State Journal of Criminal Law, 4*(2), 423–453. doi:10.1093/acprof:osobl/9780199764396.003.0007

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nagin, D. S. (1998). Criminal deterrence research at the outset of the twenty-first century. *Crime and Justice, 23,* 1–42. doi:10.1086/449268

Narramore, S. B. (2002). *No condemnation: Rethinking guilt motivation in counseling, preaching, and parenting*. Eugene, OR: Wipf and Stock.

Neff, K. (2003). Self-compassion: An alternative conceptualization of a healthy attitude toward oneself. *Self and Identity, 2*(2), 85–101. doi:10.1080/15298860309032

Nelissen, R. M. (2012). Guilt-induced self-punishment as a sign of remorse. *Social Psychological and Personality Science, 3*(2), 139–144. doi:10.1177/1948550611411520

Nelissen, R. M., & Zeelenberg, M. (2009). When guilt evokes self-punishment: Evidence for the existence of a Dobby Effect. *Emotion, 9*(1), 118–122. doi:10.1037/a0014540

Nielsen, C. S., Staud, R., & Price, D. D. (2009). Individual differences in pain sensitivity: measurement, causation, and consequences. *The Journal of Pain, 10*(3), 231–237. doi:10.1016/j.jpain.2008.09.010

Nock, M. K. (2009). Why do people hurt themselves? New insights into the nature and functions of self-injury. *Current Directions in Psychological Science, 18*(2), 78–83. doi:10.1111/j.1467-8721.2009.01613.x

Nock, M. K., & Prinstein, M. J. (2004). A functional approach to the assessment of self-mutilative behavior. *Journal of Consulting and Clinical Psychology, 72*(5), 885–890. doi:10.1037/0022-006x.72.5.885

Nyström, M. B. T., & Mikkelsen, F. (2013). Psychopathy-related personality traits and shame management strategies in adolescents. *Journal of Interpersonal Violence, 28*(3), 519–537. doi:10.1177/0886260512455512

O'Keefe, D. J. (2000). Guilt and social influence. *Annals of the International Communication Association, 23*(1), 67–101. doi:10.1080/23808985.2000.11678970

Ohtsubo, Y., Matsunaga, M., Komiya, A., Tanaka, H., Mifune, N., & Yagi, A. (2014). Oxytocin receptor gene (OXTR) polymorphism and self-punishment after an unintentional transgression. *Personality and Individual Differences, 69*, 182–186. doi:10.1016/j.paid.2014.05.033

Ohtsubo, Y., Watanabe, E., Kim, J., Kulas, J. T., Muluk, H., Nazar, G., . . . Zhang, J. (2012). Are costly apologies universally perceived as being sincere? A test of the costly apology-

perceived sincerity relationship in seven countries. *Journal of Evolutionary Psychology, 10*(4), 187–204. doi:10.1556/jep.10.2012.4.3

Okimoto, T. G., & Wenzel, M. (2008). The symbolic meaning of transgressions: Towards a unifying framework of justice restoration. *Advances in Group Processes, 25*, 291–326. doi:10.1016/s0882-6145(08)25004-6

Okimoto, T. G., & Wenzel, M. (2009). Punishment as restoration of group and offender values following a transgression: Value consensus through symbolic labelling and offender reform. *European Journal of Social Psychology, 39*(3), 346–367. doi:10.1002/ejsp.537

Orth, U. (2003). Punishment goals of crime victims. *Law and Human Behavior, 27*(2), 173–186. doi:10.1023/a:1022547213760

Parker, I. (1994). Reflexive research and the grounding of analysis: Social psychology and the psy-complex. *Journal of Community and Applied Social Psychology, 4*(4), 239–252. doi:10.1002/casp.2450040404

Paternoster, R. (2010). How much do we really know about criminal deterrence? *The Journal of Criminal Law and Criminology, 100*(3), 765–824. doi:10.2307/25766109

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Pelucchi, S., Paleari, F. G., Regalia, C., & Fincham, F. D. (2013). Self-forgiveness in romantic relationships: It matters to both of us. *Journal of Family Psychology, 27*(4), 541–549. doi:doi.org/10.1037/a0032897

Piazza, J., Landy, J. F., & Goodwin, G. P. (2014). Cruel nature: Harmfulness as an important, overlooked dimension in judgments of moral standing. *Cognition, 131*(1), 108–124. doi:10.1037/e513702014-029

Pierre, J.-M., Hutchinson, E., & Abdulrazak, H. (2007). The Shi'a Remembrance of Muharram: An explanation of the Days of Ashura and Arba'een. Military Review, 87(2), 61–69. Retrieved from https://www.questia.com/library/p5876/military-review

Platt, A. M. (1977). *The child savers: The invention of delinquency* (2 ed.). Chicago, IL: University of Chicago Press.

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management, 12*(4), 531–544. doi:10.1177/014920638601200408

Potter, J., & Hepburn, A. (2005). Qualitative interviews in psychology: Problems and possibilities. *Qualitative Research in Psychology, 2*(4), 281–307. doi:10.1191/1478088705qp045oa

Potthoff, J. G., Holahan, C. J., & Joiner, T. E. (1995). Reassurance seeking, stress generation, and depressive symptoms: An integrative model. *Journal of Personality and Social Psychology, 68*(4), 664–670. doi:10.1037/0022-3514.68.4.664

Powers, T. A., & Zuroff, D. C. (1988). Interpersonal consequences of overt self-criticism: A comparison with neutral and self-enhancing presentations of self. *Journal of Personality and Social Psychology, 54*(6), 1054–1062. doi:10.1037/0022-3514.54.6.1054

Rhodes, S., Bowie, D., & Hergenrather, K. (2003). Collecting behavioural data using the world wide web: considerations for researchers. *Journal of Epidemiology and Community Health, 57*(1), 68–73. doi:10.1136/jech.57.1.68

Riek, B. M., & Mania, E. W. (2012). The antecedents and consequences of interpersonal forgiveness: A meta-analytic review. *Personal Relationships, 19*(2), 304–325. doi:10.1111/j.1475-6811.2011.01363.x

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale.

Personality and Social Psychology Bulletin, 27(2), 151–161. doi:10.1177/0146167201272002

Rothschild, Z. K., Landau, M. J., Keefer, L. A., & Sullivan, D. (2015). Another's punishment cleanses the self: Evidence for a moral cleansing function of punishing transgressors. *Motivation and Emotion, 39*(5), 722–741. doi:10.1007/s11031-015-9487-9

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners the paradox of moral self-regulation. *Psychological Science, 20*(4), 523–528. doi:10.1111/j.1467-9280.2009.02326.x

Scheff, T. J. (1994). *Bloody revenge: Emotions, nationalism, and war*. Boulder, CO: Westview Press.

Schmader, T., & Lickel, B. (2006). The approach and avoidance function of guilt and shame emotions: Comparing reactions to self-caused and other-caused wrongdoing. *Motivation and Emotion, 30*(1), 42–55. doi:10.1007/s11031-006-9006-0

Schmitt, M., Baumert, A., Gollwitzer, M., & Maes, J. (2010). The justice sensitivity inventory: Factorial validity, location in the personality facet space, demographic pattern, and normative data. *Social Justice Research, 23*(2), 211–238. doi:10.1007/s11211-010-0115-2

Schmitt, M., Gollwitzer, M., Maes, J., & Arbach, D. (2005). Justice sensitivity. *European Journal of Psychological Assessment, 21*(3), 202–211. doi:10.1027/1015-5759.21.3.202

Schumann, K. (2012). Does love mean never having to say you're sorry? Associations between relationship satisfaction, perceived apology sincerity, and forgiveness. *Journal of Social and Personal Relationships, 29*(7), 997–1010. doi:10.1177/0265407512448277

Sedikides, C., & Luke, M. (2007). On when self-enhancement and self-criticism function adaptively and maladaptively. In E. C. Chang (Ed.), *Self-criticism and self-enhancement:*

Theory, research, and clinical implications (pp. 181–198). Washington, DC: APA Books.

Shahar, B., Szsepsenwol, O., Zilcha-Mano, S., Haim, N., Zamir, O., Levi-Yeshuvi, S., & Levit-Binnun, N. (2015). A wait-list randomized controlled trial of loving-kindness meditation programme for self-criticism. *Clinical Psychology and Psychotherapy, 22*(4), 346–356. doi:10.1002/cpp.1893

Shnabel, N., & Nadler, A. (2008). A needs-based model of reconciliation: satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation. *Journal of Personality and Social Psychology, 94*(1), 116. doi:10.1037/0022-3514.94.1.116

SimanTov-Nachlieli, I., Shnabel, N., & Nadler, A. (2013). Individuals' and groups' motivation to restore their impaired identity dimensions following conflicts. *Social Psychology, 44*(2), 129–137. doi:10.1027/1864-9335/a000148

Sinclair, J., & Green, J. (2005). Understanding resolution of deliberate self harm: qualitative interview study of patients' experiences. *British Medical Journal, 330*(7500), 1112–1115. doi:10.1136/bmj.38441.503333.8F

Skarlicki, D. P., Folger, R., & Gee, J. (2004). When social accounts backfire: The exacerbating effects of a polite message or an apology on reactions to an unfair outcome. *Journal of Applied Social Psychology, 34*(2), 322–341. doi:10.1111/j.1559-1816.2004.tb02550.x

Skarlicki, D. P., & Kulik, C. T. (2004). Third-party reactions to employee (mis) treatment: A justice perspective. *Research in Organizational Behavior, 26*, 183–229. doi:10.1016/s0191-3085(04)26005-1

Slade, P. (1982). Towards a functional analysis of anorexia nervosa and bulimia nervosa. *British Journal of Clinical Psychology, 21*(3), 167–179. doi:10.1111/j.2044-8260.1982.tb00549.x

Slepian, M. L., & Bastian, B. (2017). Truth or punishment: Secrecy and punishing the self. *Personality and Social Psychology Bulletin*. Advance online publication. doi:doi:10.1177/0146167217717245

Sloman, L., Price, J., Gilbert, P., & Gardner, R. (1994). Adaptive function of depression: Psychotherapeutic implications. *American Journal of Psychotherapy, 48*(3), 401–416. Retrieved from http://www.ajp.org/

Stanciu, C. (2015). *Investigating the proximal causes of non-suicidal self-inflicted injury (NSSII) using semantic priming* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (Order No. 1727757141)

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in Experimental Social Psychology, 21*, 261–302. doi:10.1016/s0065-2601(08)60229-4

Stiles, W. B. (1993). Quality control in qualitative research. *Clinical Psychology Review, 13*(6), 593–618. doi:10.1016/0272-7358(93)90048-Q

Stoeber, J., Harris, R. A., & Moon, P. S. (2007). Perfectionism and the experience of pride, shame, and guilt: Comparing healthy perfectionists, unhealthy perfectionists, and non-perfectionists. *Personality and Individual Differences, 43*(1), 131–141. doi:10.1016/j.paid.2006.11.012

Stone, J., Wiegand, A. W., Cooper, J., & Aronson, E. (1997). When exemplification fails: hypocrisy and the motive for self-integrity. *Journal of Personality and Social Psychology, 72*(1), 54–65. doi:10.1037/0022-3514.72.1.54

Stotz, S. J., Elbert, T., Müller, V., & Schauer, M. (2015). The relationship between trauma, shame, and guilt: Findings from a community-based study of refugee minors in Germany. *European Journal of Psychotraumatology, 6*(1), 25863. doi:10.3402/ejpt.v6.25863

Strelan, P., Di Fiore, C., & Van Prooijen, J.-W. (2017). The empowering effect of punishment on forgiveness. *European Journal of Social Psychology*, *47*(4), 472–487. doi:10.1002/ejsp.2254

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*(4), 983–992. doi:10.1016/j.jesp.2008.02.006

Sykes, G. M., & Matza, D. (1957). Techniques of neutralization: A theory of delinquency. *American Sociological Review, 22*(6), 664–670. doi:10.2307/2089195

Tanaka, H., Ohtsuki, H., & Ohtsubo, Y. (2016). The price of being seen to be just: an intention signalling strategy for indirect reciprocity. *Proceedings of the Royal Society Biological Sciences, 283*(1835), 20160694. doi:10.1098/rspb.2016.0694

Tanaka, H., Yagi, A., Komiya, A., Mifune, N., & Ohtsubo, Y. (2015). Shame-prone people are more likely to punish themselves: A test of the reputation-maintenance explanation for self-punishment. *Evolutionary Behavioral Sciences, 9*(1), 1–7. doi:10.1037/ebs0000016

Tangney, J. P. (2002). Perfectionism and the self-conscious emotions: Shame, guilt, embarrassment, and pride. In G. L. Flett & P. L. Hewitt (Eds.), *Perfectionism: Theory, research, and treatment* (pp. 199–215). Washington, DC: American Psychological Association.

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, NY: Routledge.

Tangney, J. P., Dobbins, A. E., Stuewig, J. B., & Schrader, S. W. (2017). Is there a dark side to mindfulness? Relation of mindfulness to criminogenic cognitions. *Personality and Social Psychology Bulletin*. Advance online publication. doi:10.1177/0146167217717243

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*(1), 345–372. doi:10.1146/annurev.psych.56.091103.070145

Tesser, A. (2001). On the plasticity of self-defense. *Current Directions in Psychological Science, 10*(2), 66–69. doi:10.1111/1467-8721.00117

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology, 78*(5), 853–870. doi:10.1037//0022-3514.78.5.853

Thomaes, S., Bushman, B. J., Stegge, H., & Olthof, T. (2008). Trumping shame by blasts of noise: Narcissism, self-Esteem, shame, and aggression in young adolescents. *Child Development, 79*(6), 1792–1801. doi:10.1111/j.1467-8624.2008.01226.x

Tracy, J. L., & Robins, R. W. (2004). Putting the self into self-conscious emotions: A theoretical model. *Psychological Inquiry, 15*(2), 103–125. doi:10.1207/s15327965pli1502_01

Tracy, S. J. (2010). Qualitative quality: Eight “big-tent” criteria for excellent qualitative research. *Qualitative Inquiry, 16*(10), 837–851. doi:10.1177/1077800410383121

Turiel, E. (2008). Thought about actions in social domains: Morality, social conventions, and social interactions. *Cognitive Development, 23*(1), 136–154. doi:10.1016/j.cogdev.2007.04.001

van Bunderen, L., & Bastian, B. (2014). "I have paid my dues": When physical pain reduces interpersonal justice motivations. *Motivation and Emotion, 38*(4), 540–546. doi:10.1007/s11031-014-9403-8

Van Prooijen, J.-W. (2010). Retributive versus compensatory justice: Observers' preference for punishing in response to criminal offenses. *European Journal of Social Psychology, 40*(1), 72–85. doi:10.1002/ejsp.611

Vettese, L. C., & Mongrain, M. (2000). Communication about the self and partner in the relationships of dependents and self-critics. *Cognitive Therapy and Research, 24*(6), 609–626. Retrieved from https://link.springer.com/journal/10608

Vidmar, N. (2000). Retribution and revenge. In J. Sanders & V. L. Hamilton (Eds.), *Handbook of justice research in law* (pp. 31–63). New York, NY: Kluwer Academic / Plenum.

Wadman, R., Clarke, D., Sayal, K., Vostanis, P., Armstrong, M., Harroe, C., . . . Townsend, E. (2016). An interpretative phenomenological analysis of the experience of self-harm repetition and recovery in young adults. *Journal of Health Psychology*. Advance online publication. doi:10.1177/1359105316631405

Wallace, J., & Sadalla, E. (1966). Behavioral consequences of transgression: I. The effects of social recognition. *Journal of Experimental Research in Personality, 1*(3), 187–194.

Wallington, S. A. (1973). Consequences of transgression: self-punishment and depression. Journal of Personality and Social Psychology, 28(1), 1–7. doi:10.1037/h0035576

Watanabe, E., & Ohtsubo, Y. (2012). Costly apology and self-punishment after an unintentional transgression. *Journal of Evolutionary Psychology, 10*(3), 87–105. doi:10.1556/jep.10.2012.3.1

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063. doi:10.1037/0022-3514.54.6.1063

Watson, P. J., & Andrews, P. W. (2002). Toward a revised evolutionary adaptationist analysis of depression: The social navigation hypothesis. *Journal of Affective Disorders, 72*(1), 1–14. doi:10.1016/s0165-0327(01)00459-1

Weiner, B., Graham, S., Peter, O., & Zmuidinas, M. (1991). Public confession and forgiveness. *Journal of Personality, 5*9(2), 281–312. doi:10.1111/j.1467-6494.1991.tb00777.x

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*(3), 401–417. doi:10.1002/ejsp.629

Wenzel, M., & Okimoto, T. G. (2012). The varying meaning of forgiveness: Relationship closeness moderates how forgiveness affects feelings of justice. *European Journal of Social Psychology, 42*(4), 420–431. doi:10.1002/ejsp.1850

Wenzel, M., & Okimoto, T. G. (2014). On the relationship between justice and forgiveness: Are all forms of justice made equal? *British Journal of Social Psychology, 53*(3), 463–483. doi:10.1111/bjso.12040

Wenzel, M., & Okimoto, T. G. (2015). “We forgive”: A group’s act of forgiveness and its restorative effects on members’ feelings of justice and sentiments towards the offender group. *Group Processes and Intergroup Relations, 18*(5), 655–675. doi:10.1177/1368430215586274

Wenzel, M., & Okimoto, T. G. (2016). Retributive justice. In C. Sabbagh & M. Schmitt (Eds.), *Handbook of social justice theory and research* (pp. 237–256). New York, NY: Springer.

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and restorative justice. *Law and Human Behavior, 32*(5), 375–389. doi:10.1007/s10979-007-9116-6

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2010). Justice through consensus: Shared identity and the preference for a restorative notion of justice. *European Journal of Social Psychology, 40*(6), 909–930. doi:10.1002/ejsp.657

Wenzel, M., Woodyatt, L., & Hedrick, K. (2012). No genuine self-forgiveness without accepting responsibility: Value reaffirmation as a key to maintaining positive self-regard. *European Journal of Social Psychology, 42*(5), 617–627. doi:10.1002/ejsp.1873

Wenzel, M., Woodyatt, L., & McLean, B. (2017). Implicit guilt following wrongdoing: Acceptance threat causes defensiveness, which value affirmation helps to overcome. Unpublished manuscript, School of Psychology, Flinders University, Adelaide, Australia.

Whelton, W. J., & Greenberg, L. S. (2005). Emotion in self-criticism. *Personality and Individual Differences, 38*(7), 1583–1595. doi:10.1016/j.paid.2004.09.024

Witvliet, C. V., Worthington, 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*(1), 10–25. doi:10.1016/j.jesp.2007.01.009

Wong, Y., & Tsai, J. (2007). Cultural models of shame and guilt. In J. L. Tracy, R. W. Robins, & J. P. Tangney (Eds.), *The self-conscious emotions: Theory and research* (pp. 209–223). New York, NY: Guilford.

Woodyatt, L., & Wenzel, M. (2013a). The psychological immune response in the face of transgressions: Pseudo self-forgiveness and threat to belonging. *Journal of Experimental Social Psychology, 49*(6), 951–958. doi:10.1016/j.jesp.2013.05.016

Woodyatt, L., & Wenzel, M. (2013b). Self-forgiveness and restoration of an offender following an interpersonal transgression. *Journal of Social and Clinical Psychology, 32*(2), 225–259. doi:10.1521/jscp.2013.32.2.225

Woodyatt, L., & Wenzel, M. (2014). A needs-based perspective on self-forgiveness: Addressing threat to moral identity as a means of encouraging interpersonal and intrapersonal restoration. *Journal of Experimental Social Psychology, 50*, 125–135. doi:10.1016/j.jesp.2013.09.012

Woodyatt, L., Wenzel, M., & Ferber, M. (2017). Two pathways to self-forgiveness: A hedonic path via self-compassion and a eudaimonic path via the reaffirmation of violated values. *British Journal of Social Psychology*, Advance online publication. doi:10.1111/bjso.12194

Worthington, E. L., Jr., & Wade, N. G. (1999). The psychology of unforgiveness and forgiveness and implications for clinical practice. *Journal of Social and Clinical Psychology, 18*(4), 385–418. doi:10.1521/jscp.1999.18.4.385

Wurmser, L. (1974). Psychoanalytic considerations of the etiology of compulsive drug use. *Journal of the American Psychoanalytic Association, 22*(4), 820–843. doi:10.1177/000306517402200407

Yardley, L. (2000). Dilemmas in qualitative health research. *Psychology and Health, 15*(2), 215–228. doi:10.1080/08870440008400302

Yelsma, P., Brown, N. M., & Elison, J. (2002). Shame-Focused coping styles and their associations with self-esteem. *Psychological Reports, 90*(3), 1179–1189. doi:doi:10.1177/003329410209000320.2

Young, A. M., Boyd, C., & Hubbell, A. (2000). Prostitution, drug use, and coping with psychological distress. *Journal of Drug Issues, 30*(4), 789–800. doi:10.1177/002204260003000407

Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science, 313*(5792), 1451–1452. doi:10.1126/science.1130726

Zhong, C.-B., Liljenquist, K. A., & Cain, D. M. (2009). Psychological perspectives on ethical behavior and decision making. In D. De Cremer (Ed.), *Moral self-regulation* (pp. 75–89). Charlotte, NC: Information Age.

Zhong, C.-B., Strejcek, B., & Sivanathan, N. (2010). A clean self can render harsh moral judgment. *Journal of Experimental Social Psychology, 46*(5), 859–862. doi:http://dx.doi.org/10.1016/j.jesp.2010.04.003

Zhu, R., Jin, T., Shen, X., Zhang, S., Mai, X., & Liu, C. (2017). Relational utility affects self-punishment in direct reciprocity and indirect reciprocity situations. *Social Psychology, 48*(1), 19–27. doi:10.1027/1864-9335/a000291

Zuroff, D. C., Moskowitz, D., & Côté, S. (1999). Dependency, self-criticism, interpersonal behaviour and affect: Evolutionary perspectives. *British Journal of Clinical Psychology, 38*(3), 231–250. doi:10.1348/014466599162827