It's Different When I Do It: Self-Protection Affects Construals of Negative Behaviors

A dissertation presented to
the faculty of
the College of Arts and Sciences of Ohio University

In partial fulfillment
of the requirements for the degree
Doctor of Philosophy

Gregory S. Preuss

August 2011

© 2011 Gregory S. Preuss. All Rights Reserved.
This dissertation titled
It's Different When I Do It: Self-Protection Affects Construals of Negative Behaviors

by

GREGORY S. PREUSS

has been approved for
the Department of Psychology
and the College of Arts and Sciences by

__________________________________________

Mark D. Alicke
Professor of Psychology

__________________________________________

Benjamin M. Ogles
Dean, College of Arts and Sciences
ABSTRACT

PREUSS, GREGORY S., Ph.D., August 2011, Psychology

It's Different When I Do It: Self-Protection Affects Construals of Negative Behaviors

Director of Dissertation: Mark D. Alicke

Self-protection refers to the people's tendency to minimize their shortcomings in order to avoid feeling bad about themselves. The present research examines how self-protection motivations affect people's construals regarding their potentially immoral behaviors. Previous research on this topic has focused on the excuses, rationalizations, and self-serving attributions that people employ to argue that their behaviors are not as bad as they appear. Less is known about how people construe their behaviors that they know are bad. To inform this gap in the literature, the present studies sought a more nuanced view of people's beliefs about their own bad behaviors by examining how these beliefs become integrated into their self-concepts.

The purpose of these three studies was to provide experimental evidence that the magnitude of self-protection varies as a function of levels of construal or abstraction. To this end, each study involved an actor-observer paradigm in which participants provided concrete construals of identical behaviors committed by themselves and others and abstract construals regarding the degree to which these behaviors were linked to higher-level self-concepts. In Study 1, regardless of level of abstraction, participants judged themselves more leniently than target persons. When participants were granted access to target person's justifications for their negative behaviors (Study 2 and Study 3), however, they viewed their own behaviors as less negative than those of others (low-level
construals). Across all three studies, participants predicted that they would be less likely than target persons to perform the negative behaviors again (mid-level construals). When judging themselves relative to target persons in Study 3, participants did not rate themselves less negatively on trait judgments and global attributions (high-level construals). All three studies provide converging evidence that the self-other discrepancies are at least partially accounted for by motivational concerns. Additionally, the results of Study 2 and Study 3 suggest that the function of self-protection is allow people to maintain favorable self-images while simultaneously acknowledging the negativity of their past behaviors. Discussion focuses on potential limitations as well as directions for future research.

Approved: _____________________________________________________________

Mark D. Alicke

Professor of Psychology
ACKNOWLEDGMENTS

I would like to thank my committee members, Keith Markman, Justin Weeks, Matthew Vess, and Laura Black, for the precious time and invaluable feedback they have provided to improve the state of my current research. Likewise, I would like to thank my dissertation director, advisor, and friend Mark Alicke for providing me with sage advice while having unwavering confidence in my ability to make astute decisions.

I would also to extend a special thank you to my parents, Don and Linda, and my grandparents, Mimi and Granddaddy, for the unconditional love they have given me. Without their support, I would not be where I am today. Nothing in life feels better than making them proud of me.

I would also like to extend my deepest thanks to my favorite professor from the University of Kentucky, Richard Smith. Over the last ten years, he has selflessly made time in his busy schedule to talk with me and provide advice during my journey. It has been and will continue to be my sincere pleasure to “pay it forward” by making time in my schedule to mentor my own students and have a positive influence on their lives.

Last but not least, I wish to thank the many students who I have had the pleasure of teaching and mentoring over the years, especially my research assistants (“Team Preuss”), for continually reminding me that my “place on the wheel” is in a college classroom. They are my number one source of inspiration, and I want to thank them from the bottom of my heart.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>3</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>5</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>8</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>9</td>
</tr>
<tr>
<td>Overview of Studies</td>
<td>13</td>
</tr>
<tr>
<td>STUDY 1</td>
<td>17</td>
</tr>
<tr>
<td>Method</td>
<td>19</td>
</tr>
<tr>
<td>Participants</td>
<td>19</td>
</tr>
<tr>
<td>Procedure</td>
<td>19</td>
</tr>
<tr>
<td>Results</td>
<td>20</td>
</tr>
<tr>
<td>Discussion</td>
<td>25</td>
</tr>
<tr>
<td>STUDY 2</td>
<td>27</td>
</tr>
<tr>
<td>Method</td>
<td>29</td>
</tr>
<tr>
<td>Participants</td>
<td>29</td>
</tr>
<tr>
<td>Procedure</td>
<td>29</td>
</tr>
<tr>
<td>Results</td>
<td>31</td>
</tr>
<tr>
<td>Self-Other Comparisons</td>
<td>32</td>
</tr>
<tr>
<td>Self-Esteem and Narcissism</td>
<td>33</td>
</tr>
<tr>
<td>STUDY 3</td>
<td>37</td>
</tr>
<tr>
<td>Participants</td>
<td>38</td>
</tr>
<tr>
<td>Procedure</td>
<td>39</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>44</td>
</tr>
<tr>
<td>First Reality Criterion: Actors-Only Analysis</td>
<td>44</td>
</tr>
<tr>
<td>Additional Reality Criteria: Yoked Observer Analysis</td>
<td>56</td>
</tr>
<tr>
<td>Summary of All Study 3 Analyses</td>
<td>65</td>
</tr>
<tr>
<td>GENERAL DISCUSSION</td>
<td>68</td>
</tr>
<tr>
<td>Before You Judge Me, Let Me Explain</td>
<td>70</td>
</tr>
<tr>
<td>Future Directions</td>
<td>84</td>
</tr>
</tbody>
</table>
Concluding Thoughts.................................................................................................... 88
REFERENCES.................................................................................................................. 89
LIST OF TABLES

Table 1: Percentage of Study 1 Participants Who Engaged in Each Behavior .............. 21
Table 2: Mean Ratings of Self and Others in Study 1 .............................................. 24
Table 3: Percentage of Study 2 Participants Who Engaged in Each Behavior ............. 32
Table 4: Mean Ratings of Self and Others in Study 2 .............................................. 33
Table 5: Percentage of Study 3 Participants Who Engaged in Each Behavior ............. 44
Table 6: Mean Rating of Self and Others in Study 3 ............................................... 47
Table 7: Actors' and Yoked Observers' Mean Subjective Likelihood Estimates that Actors and Target Persons Would Perform Each Negative Behavior Again ...................... 52
Table 8: Situation-Behavior Patterns: Mean Ratings for Actors and Yoked Observers 54
Table 9: Mean Ratings of Actors by Self and Yoked Observers in Study 3 ................. 58
Table 10: Mean Ratings of Target Persons by Actors and Yoked Observers in Study 3 . 63
INTRODUCTION

"Your honor, I am not the bad person or beast I've been made out to be. I have been talked about and ridiculed on a day to day basis by people who really don't know Michael Vick the human being."

– Michael Vick ("I Am Not the Beast," 2007)

"I'm not a bad guy! I work hard, and I love my kids. So why should I spend half my Sunday hearing about how I'm going to Hell?"

– Matt Groening, the Simpsons, U.S. cartoonist and satirist

When professional football player Michael Vick was sentenced in 2007 to 23 months in prison for financing a dog fighting conspiracy, he became a national symbol for animal cruelty. The crimes to which he pled guilty included shooting, electrocuting, and drowning dogs that refused to fight. Although Vick acknowledged that his actions were wrong, he viewed himself much differently than the people who demonized him. Rather than perceiving himself as a bad person, Vick described himself as a "humble, soft-spoken, and caring" father who regretted the negative effect that his actions had on his five-year-old and two-year-old children who viewed news stories that depicted their father as a terrible human being ("I Am Not the Beast", 2007). Unlike the media and public who castigated him, Vick did not feel that people should draw a globally negative impression of him regarding the kind of person that he was on the basis of his crimes. Using reasoning analogous to that used in the opening quote by Matt Groening, Vick felt others should judge him based on his good behaviors rather than his bad behaviors. Both Vick and Groening maintained that their good behaviors represented the "real me."
By claiming that he was not the bad person or beast that people made him out to be, Michael Vick sought to manage not only his public image (Schlenker, 1980), but also his private self-image. Just as Vick wanted to avoid being universally despised by the American public, he also wanted to avoid despising himself. After all, every time Michael Vick looks in the mirror for the rest of his life, he will do so with the knowledge that he committed heinous crimes that most people think only a morally depraved individual could even consider committing. It should come as no surprise that he wishes to regard himself in a manner that is based on something other than his criminal actions.

Although the behaviors in the opening vignette are extreme, the motivations that underlie the opening quotations from Michael Vick and Matt Groening are common. People have a strong desire to view themselves as good, competent, and moral. The fact that an overwhelming majority of people in normal populations has chronically favorable self-views (Schmitt & Allik, 2005) suggests that many people are quite successful at achieving this goal. One of the strategies that people employ to pursue favorable self-conceptions is self-enhancement, which refers to the tendency to exaggerate one's virtues.

Under most circumstances, self-enhancement tendencies are pervasive. Nevertheless, there are times when people do bad things. During these times, self-protection, which refers to the tendency to minimize one's shortcomings, becomes an urgent priority (Alicke & Sedikides, 2009; Sedikides, in press; Sedikides & Gregg, 2008). As illustrated by the opening quotations, people experience discomfort when self-relevant events suggest that they are not good people. To counteract threats to their typically favorable self-images and restore psychological equilibrium, people engage in self-
protection (Sedikides, in press). As postulated by Freud, Festinger, and numerous other psychologists, people have a plethora of mental tools that can be used to ignore or manipulate negative self-relevant information in a manner that minimizes threats to desired self-images (Festinger, 1957; Freud, 1936; Greenwald, 1980; Kunda, 1990; Steele, 1988; Taylor & Brown, 1988).

When reflecting on their wrongdoings, how do people perceive themselves and their actions? For most individuals, the default strategy is to provide self-serving attributions, excuses (Alicke & Sedikides, 2009; Snyder, 1989; Snyder, Higgins, & Stucky, 1983), or rationalizations (Tsang, 2002) to justify behaviors that are inconsistent with their typically favorable self-conceptions (Aronson, 1968, 1969, 1992, 1998; Aronson, Chase, Helmreich, & Ruhnke, 1974; Greenwald & Ronis, 1978; Harmon-Jones & Mills, 1998; Thibodeau & Aronson, 1992). In many cases, not only are excuses designed to bolster public claims that certain behaviors are not as bad as they appear to be, but they are also designed to preserve people's private self-images as good and decent human beings.

Although denying the negativity of their own apparent misdeeds is often an option, there are times when doing so stretches the bounds of believability to oneself and others. Sometimes people engage in behaviors that they know are bad. During these times, in order to avoid appearing delusional, people adopt more reasonable strategies (Sedikides & Gregg, 2003, 2008) that allow them to maintain favorable self-images at the global level while acknowledging the negativity of their past behaviors. For example, Goffman (1971) argued that apologies allow a person to split the "bad" part of the self
from the much larger part of the self that is composed of their "good" characteristics. On a related note, a self-protection strategy known as "temporal bracketing" enables people to split their "bad self" from their "good self" by claiming that their past negative behaviors reflect the "old me." By claiming that their past behaviors reflect the "old me", people attempt to separate themselves from the negative implications of their past wrongdoings (Baumeister, Stillwell, & Wotman, 1990).

As suggested by research on apologies and temporal bracketing, some self-protection strategies allow people to acknowledge the negativity of past behaviors while simultaneously maintaining a positive self-image at the global level. The utilization of these strategies reveals the multifaceted nature of the self. As elucidated by the opening example involving Michael Vick, the multi-faceted self can be utilized to acknowledge the nefarious nature of one's misdeeds without reaching globally negative conclusions about the self. At the level of the concrete behavioral act, Vick acknowledged that it is wrong to drown and electrocute dogs. With judgments at greater levels of abstraction, however, Vick did not allow the negative self-relevant thoughts and feelings associated with these misdeeds to spread to the totality of his self-concept.

Michael Vick is not the first person, nor will he be the last, to urge people to refrain from judging him on the basis of his past negative behaviors. This brings us to the central questions posed by the present research. When people do bad things that they know are bad, how does this negative self-relevant information become incorporated into their self-concepts? How do people continue to feel good about themselves when they commit behaviors that violate their own sense of morality? To answer these questions,
the current research goes beyond previous approaches that have focused on the excuses and self-serving attributions people make for their negative behaviors. Rather, the current research proposes a more nuanced view of people’s beliefs about their own bad behaviors, and in particular, how these beliefs relate to self-concepts.

The purpose of this research is to provide experimental evidence that the magnitude of self-protection varies as a function of levels of construal or abstraction. At the concrete level, participants will indicate how negative they perceive their past potentially immoral behaviors to be. At the abstract level, participants will indicate the degree to which their behaviors reveal negative information about their dispositions. My thesis is that the multifaceted nature of the self allows people to acknowledge the negativity of their concrete behavioral acts while simultaneously feeling good about themselves at the global level.

Overview of Studies

To explore the possibility that people can feel good about themselves on the global level while simultaneously acknowledging the negativity of their past behaviors, each of the three studies in the present research will utilize an actor-observer design. To measure lower-level construals that focus on the concrete behavioral acts, participants will be asked to indicate how negative they perceive their past misdeeds to be. In the case of some behaviors, such as those in the opening vignette involving Michael Vick, the evidence is so overwhelming that it becomes difficult for the perpetrator to convince even himself that the behaviors are not negative. Just as reality can constrain self-enhancement (Sedikides & Gregg, 2003, 2008), reality can limit the degree to which people engage in
self-protection. As suggested by the opening vignette, however, denying the nefariousness of a past misdeed is not the only way to protect one's self-image.

When reality constraints make self-protective construals of concrete behaviors a less viable option, people engage in more judicious forms of self-protection (Alicke & Sedikides, 2009). Specifically, people maintain favorable self-images by strategically shifting the meaning of their behaviors (Sedikides & Gregg, 2003) when defining self-concepts at more general and abstract levels (Liberman, Trope, & Stephan, 2007). Does the commission of a negative behavior necessarily imply that the perpetrator is a bad person? Should the person who engaged in the negative behavior form negative inferences about their own dispositions? Because of the abstract nature of these questions, people have a good deal of latitude in how they make these dispositional self-inferences and global self-judgments. The flexibility that people have in making these abstract judgments provides them with more wiggle room to massage the data in a manner that helps them maintain positive self-images (Critcher, Helzer, & Dunning, 2011).

What yardstick do researchers use to determine whether a self-judgment is an example of self-protection? In order to provide empirical evidence of self-protection, self-judgments regarding a negative behavior must be compared to a reality criterion (Alicke & Sedikides, 2009). The researcher's task is to capture the difference between self-related judgments and reality. Self-protection is empirically demonstrated when people evaluate themselves less unfavorably than a reality criterion. As is the case with studies that seek to measure self-enhancement, the researcher's choice of a suitable
criterion often sparks debate over measurement issues (Krueger & Wright, 2010). To circumvent this challenge, the current set of studies will utilize multiple criteria to measure the presence (or absence) of self-protection biases. The case for self-protection will be strengthened if similar pattern of results are found across multiple reality criteria.

The first criterion, which will be utilized in all three studies, involves within-participants comparisons between actor participants' construals regarding their own socially undesirable behaviors and their construals of target persons who engaged in the exact same behaviors. In these comparisons, actors' construals of target persons represent the accuracy criterion. If a person draws less negative inferences about the self than the other person, then the self-other discrepancies will be interpreted as evidence of self-protection.

To provide a solid basis for a motivational account, procedural steps will be taken in Studies 2 and 3 to rule out an alternative explanation for self-other discrepancies that makes no mention of goal-oriented processes. According to this nonmotivational account, when judging behaviors committed by the self and others, people have more information about their own thoughts, motives, and intentions than those of others (Jones & Nisbett, 1972; Kruger & Gilovich, 2004). In Study 2, participants will have access to others' justifications for engaging in various negative behaviors. In Study 3, participants will have access to an even more extensive array of contextual information surrounding the negative behaviors of others. By taking these steps to further equate the contextual information surrounding the behaviors of self and others, the case for a motivational account for self-other discrepancies will become more compelling.
In cases where the actors render less unfavorable judgments for themselves than others, the possibility exists that their own behaviors are, in actuality, less negative than those of others. To address this possibility, a neutral observer will be yoked to each actor in Study 3. Actors’ construals will be compared to those of yoked observers. As described in Study 3, these comparisons will yield additional criteria by which to examine the possible presence of self-protection effects. If similar patterns of results are found across multiple criteria, nonmotivational accounts for the results will be rendered less tenable. The actor-observer comparisons involving these additional criteria will be discussed in greater detail in the results section of Study 3.
STUDY 1

The purpose of Study 1 is to add to the existing literature by exploring the extent to which self-protection motives influence construals of one's past negative behaviors. An actor-observer design similar to the one used by Alicke, Vredenburg, Hiatt, and Govorun (2001) will be used. In this novel two-phase paradigm, the same participants served in the actor role in the first phase and the observer role in the second phase. In the first phase, participants in a pretesting session provided construals of their past negative behaviors. For example, participants who admitted that they had cheated on a significant other were asked to indicate how bad the behavior was, how representative the behavior was of them, and the likelihood that they would perform the action again in the future. Later in the academic quarter, participants provided construals of target persons who had engaged in the exact same behaviors that they had previously admitted committing.

This methodology provided the opportunity to assess whether construals of the potentially immoral behaviors of self and others differed when the behaviors were equated. I predicted that participants' construals regarding their own past socially undesirable behaviors would be less negative than their construals of other people who had engaged in the same exact behaviors as themselves. Furthermore, I predicted that self-protection motivations would at least partially account for these self-other discrepancies.

For all of the foregoing comparisons, in the absence of motivational concerns, one would expect these construals to be identical. However, if self-protection motivations are involved in the construals of people’s past socially undesirable behaviors, these
motivations could manifest themselves in one or more of the following ways. With regard to the first comparison mentioned above, participants could maintain that although the behaviors that they and target persons committed are identical, their behaviors are less negative than those of others.

At this point, one caveat is in order. Even if participants recognize that their socially undesirable behaviors are just as bad as those of others, it does not, in and of itself, preclude the possibility that they might find other ways to minimize the nefariousness of their deeds. Even if they acknowledge the negativity of their past socially undesirable behaviors, participants might have additional avenues to idiosyncratically construe their behaviors in a manner that serves the need to preserve their self-images as moral and virtuous individuals. For example, participants may serve their self-protection motives via their construals of what their previous negative behaviors reflect about the kind of person they are now and the kind of person they will be in the future.

With regard to the kind of person they are now, evidence for self-protection motivations will be demonstrated if participants indicate that while their own socially undesirable behaviors reveal little about their own personalities, these same behaviors are at least somewhat representative of the personalities of other people. With regard to the kind of person they will be in the future, evidence for self-protection motivations will be demonstrated if participants believe that they are less likely than others to perform these socially undesirable actions again.
Method

Participants

Participants were 165 undergraduates (42 males, 123 females) who received course credit for participating. Their ages ranged from 18 to 22 years. Data were collected in small groups ranging from two to 20 participants.

Procedure

Phase 1: Pretesting Session

The pretesting session was conducted on-line. Participants were asked whether they had previously engaged in any of the following ten behaviors: lied to a significant other, cheated on a significant other, had unprotected sex, slept with a friend’s significant other, stole something, cheated on a test, lied to their parents, drove under the influence of alcohol, intentionally hurt someone’s feelings, or engaged in vandalism. The order in which these behaviors appeared in the on-line questionnaire was randomized. For the behaviors in which they had previously engaged, participants answered three 7-point Likert scale questions. These three questions allowed participants to provided construals at various levels of abstraction. In a question that involved lower-level construals, participants indicated how bad the behavior was (1 = not bad at all, 7 = extremely bad). In two questions that corresponded to higher-level construals, participants indicated how representative the behavior was of them (1 = not at all representative of who you I am as a person, 7 = extremely representative of who I am as a person) and the likelihood that they would perform the action again in the future (1 = extremely unlikely, 7 = extremely likely).
Phase 2: experimental session

During the last four weeks of the academic quarter, participants returned for the final part of the experiment, which was described as an experiment on “judging behaviors.” In contrast to phase 1 when participants answered questions about their own past behaviors, in phase 2, participants answered questions about the socially undesirable behaviors of others. The order in which the 10 behaviors appeared in the questionnaire packet was randomized. For each of the ten behaviors, participants indicated how bad the behavior was, how representative the behavior was of the other person, and how likely the other person was to perform the action again in the future.

During the experimental session, participants were not asked to indicate whether they had personally ever engaged in any of the ten behaviors. Of course, data from the pretesting session allowed me to ascertain whether participants had ever engaged in any of the negative behaviors. Participants’ student identification numbers were used to link their responses during the pretesting session to their responses during the experimental phase.

Results

Data from the four participants who reported during the pretesting session that they had never engaged in any of the ten behaviors were excluded from all analyses. Of the remaining 161 participants, the number of behaviors for which they reported engaging ranged from one to nine \( (M = 4.36, SD = 2.17) \). Table 1 displays data from the pretesting session indicating the percentage of the sample that reported engaging in each behavior.
Table 1

**Percentage of Study 1 Participants Who Engaged in Each Behavior**

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lied to Parents</td>
<td>96.9</td>
</tr>
<tr>
<td>Cheated on Test</td>
<td>67.1</td>
</tr>
<tr>
<td>Lied to Significant Other</td>
<td>65.2</td>
</tr>
<tr>
<td>Intentionally Hurt Someone's Feelings</td>
<td>51.6</td>
</tr>
<tr>
<td>Had Unprotected Sex</td>
<td>50.3</td>
</tr>
<tr>
<td>Stole Something</td>
<td>42.9</td>
</tr>
<tr>
<td>Drove Under the Influence of Alcohol</td>
<td>27.3</td>
</tr>
<tr>
<td>Cheated on Significant Other</td>
<td>21.1</td>
</tr>
<tr>
<td>Engaged in Vandalism</td>
<td>20.5</td>
</tr>
<tr>
<td>Slept with Significant Other of a Friend</td>
<td>3.7</td>
</tr>
</tbody>
</table>

During the experimental session, participants answered three questions about each of the ten behaviors. To remind the reader, one of these questions concerned the perceived negativity of each of the socially undesirable behaviors. The other two questions addressed how representative each behavior was of the person who committed and the estimated likelihood that the person would perform the action again in the future.

The primary focus of Study 1 was to compare participants’ judgments regarding their own socially undesirable behaviors (measured during phase 1) to their judgments of other people who engaged in the exact same behaviors (measured during phase 2).

During phase 1, participants responded to the three aforementioned questions about each of the socially undesirable behaviors they had committed. An average for each of the three questions was computed by aggregating participants’ responses across the negative behaviors that they had committed. For example, if a participant reported
engaging in three of the socially undesirable behaviors, his or her answers to the questions about how likely they were to perform each action again were summed and then divided by three.

Although participants answered questions about ten socially undesirable behaviors of other people during phase 2, their responses to questions pertaining to behaviors that they had personally committed themselves were the only responses that were of theoretical interest. Participants’ judgments about others that corresponded to behaviors that they had not personally committed were not of theoretical interest because participants did not answer questions about behaviors that they had never committed during phase 1. Continuing with the example from the previous paragraph, if a participant reported during phase 1 that he or she had engaged in three of the socially undesirable behaviors, his or her answers about how likely the target persons were to perform each of these actions were summed and then divided by three.

The phase 1 data yielded three averages, one for each of the three dependent measures regarding the socially undesirable behaviors that the participants had personally committed. Likewise, the phase 2 data yielded three averages, one for each of the three dependent measures pertaining to participants’ judgments of other people who had committed the exact same behaviors as they had. In the correlated $t$-tests described below, each participant’s average response to each question regarding his or her own negative behaviors was compared to his or her average response to a similarly-worded question about other people who engaged in the exact same negative behaviors. Participants’ average self-judgments (phase 1) were subtracted from their average
judgments of other people (phase 2). Statistically significant differences were interpreted as evidence that self-protection motives affect the construal of people's own potentially immoral behaviors.

Three correlated $t$-test were conducted, one for each of the three questions. The first question asked participants to indicate how bad they perceived certain socially undesirable behaviors to be. During phase 1, participants answered this question immediately after acknowledging that they had previously engaged in the socially undesirable behavior. During phase 2, participants answered this question without being asked whether they had ever engaged in the corresponding negative behavior. Both composite scores were comprised of the negativity ratings for the socially undesirable behaviors that participants had previously acknowledged committing. As predicted, even though the behaviors being judged in each phase were identical, the results of a correlated $t$-test indicated that on average, participants perceived their own socially undesirable behaviors to be less negative ($M = 3.84$) during phase 1 than during phase 2 ($M = 4.80$), $t(160) = -7.54, p < .001$.

The second correlated $t$-test compared how representative participants felt their own negative behaviors were of them as a person to how representative they felt these same behaviors were of other people who committed them. As predicted, the results indicated that participants felt these negative behaviors were more representative of other people ($M = 3.78$) than themselves ($M = 1.96$), $t(158) = -17.01, p < .001$.

The third question asked participants to estimate the likelihood that a person who engaged in a negative behavior would perform the same action again in the future. As
predicted, the results of the correlated $t$-test indicated that on average, participants felt that they would be less likely to perform these actions again ($M = 2.80$) than other people who had previously engaged in the same behaviors ($M = 5.71$), $t(157) = -27.62, p < .001$ (See Table 2).

Table 2

*Mean Rating of Self and Others in Study 1*

<table>
<thead>
<tr>
<th>Question</th>
<th>$M$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self</td>
<td>Others</td>
<td>$t$</td>
<td>$P$</td>
</tr>
<tr>
<td>How bad is this behavior?</td>
<td>3.84</td>
<td>4.80</td>
<td>-7.54</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>How representative is this behavior of you (this person)?</td>
<td>1.96</td>
<td>3.78</td>
<td>-17.01</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>How likely are you (this person) to repeat this behavior?</td>
<td>2.80</td>
<td>5.71</td>
<td>-27.62</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

*Note.* Ratings could range from 1 to 7. Higher values indicate more negative evaluations.

As previously mentioned, the number of negative behaviors that participants reported committing ranged from one to nine. My decision to create an aggregate score for each person raises the potential concern that the magnitude of the actor-observer differences might be related to the number of negative behaviors that the participant reported committing. To explore this possibility, I computed a difference score between participants’ pretesting session and experimental phase average ratings of how bad their negative behaviors were. This difference scores was not correlated to the total number of socially undesirable behaviors that participants had committed, $r = .02, p = .835$. 
Next, I computed a difference score between the average ratings for how representative participants perceived their negative behaviors to be and how representative they perceived these same behaviors to be of other people. This difference score was positively correlated to the total number of socially undesirable behaviors that participants had committed, $r = .27, p < .001$. As the number of socially undesirable behaviors committed by participants increased the size of the difference score regarding the representativeness of the behaviors increased.

Next, I computed a difference score between the average ratings for how likely participants felt it was that they would perform the negative behaviors in the future and how likely they felt it was that other people would perform these same behaviors in the future. This difference scores was not correlated to the total number of socially undesirable behaviors that participants had committed, $r = -.05, p = .553$.

With the exception of judgments regarding how representative the negative behaviors were of the person who committed them, I tentatively conclude that the decision to create an aggregate score for each participant on each of the dependent variables does not pose serious problems in the interpretation of the results. Similar analysis were conducted in Study 2 to further explore the possibility that the variation in the number of socially undesirable committed by participants might confound the results.

Discussion

The primary focus of Study 1 was to compare participants' judgments regarding their own socially undesirable to their judgments of other people who engaged in the exact same behaviors. Three comparisons revealed that participants construed their
socially undesirable behaviors in self-serving ways for both low and high-level construals. With regard to low-level construals, participants believed their own socially undesirable behaviors were less negative than those of others. With regard to higher-level construals, although participants felt that their past negative behaviors were not at all representative of their own personalities, they felt these exact same behaviors were at least somewhat representative of the personalities of other people who committed them. This self-other difference is consistent with research that has suggested that people think they know others better than others know them (Pronin, Kruger, Savitsky, & Ross, 2001). Also with regard to higher-level construals, participants felt they were much less likely than others to perform the negative behaviors again in the future. These optimistic predictions are consistent with research on the holier than thou phenomenon (Balcetis, Dunning, & Miller, 2008; Epley & Dunning, 2000, 2006).

The results of Study 1 raise important questions. To what extent can the self-other differences in Study be accounted for by self-protection motives? To what extent were these self-other differences due to the fact that actor participants had more information about their own thoughts, motives, and intentions than those of others (Jones & Nisbett, 1972; Kruger & Gilovich, 2004)? It is to these questions that we turn in Study 2.
STUDY 2

Although the self-other differences in Study 1 were in the predicted direction, they may not have been due to self-protection motivations. Rather, a plausible cognitive account for the results based on mundane availability biases exists. According to this account, participants had access to a greater amount of contextual information surrounding their own negative behaviors than those of others. That is, when looking back at their own socially undesirable behaviors, participants were probably able to remember the reasons that they engaged in these behaviors. Doing so would have allowed them to use these reasons to justify their negative behaviors in order to convince themselves that they were not bad people. When judging the socially undesirable behaviors of others, however, participants did not know anything about other people's reasons or justification for their negative behaviors.

In light of research that has shown that people judge themselves on the basis of their intentions while judging others on the basis of their concrete behaviors (Kruger & Gilovich, 2004), an important procedural change was adopted in Study 2 to further equate the contextual information that participants had about their own potentially immoral behaviors and those of others. During the pretesting session in Study 2, whenever participants admitted that they had engaged in particular negative behaviors, they were asked to explain why they engaged in the behavior. When participants were in the observer role in the experimental phase that occurred later in the quarter, they were given a randomly selected participant’s justification for engaging in each behavior. Assuming there is no difference in the quality of the justifications reported by randomly paired
participants, statistically significant actor-observer differences similar to those found in Study 1 would provide even stronger evidence that self-protection motives best account for these self-other asymmetries.

An additional aim of Study 2 was to examine whether the self-other differences that were found in Study 1 might be more pronounced for individuals who hold high opinions of themselves. To examine this possibility, all of the participants in Study 2 completed measures of self-esteem and narcissism. The addition of these two measures marked a second change to the procedures of Study 1. Although a meta-analysis indicated that self-esteem and narcissism are positively correlated ($r = .29, k = 11, n = 2,963, p < .001$) (Campbell, 2001), these constructs are conceptually distinct in theoretically important ways. On the one hand, high self-esteem individuals and narcissists are similar in that they both rate themselves as better than average on agentic traits (e.g., intellectual skills, extraversion). On the other hand, a distinction between high self-esteem individuals and narcissists can be found when these individuals rate themselves on communal traits (e.g., agreeableness, morality). Although people who score high in self-esteem rate themselves as better than average on communal traits, people who score high on narcissism do not rate themselves as better than average on communal traits. For example, narcissists rate themselves as more intelligent, but not more moral, than average (Campbell, Rudich, & Sedikides, 2002).

Although the current research does not focus on traits per se, it does focus on behaviors that are germane to morality. Based on the research of Campbell et al. (2002), I predicted that the self-other differences that I expected to replicate in Study 2 would be
moderated by individual differences in self-esteem. That is, I hypothesized that the magnitude of the self-protection biases would be more pronounced for individuals with higher levels of self-esteem. However, I did not expect these self-other differences would be moderated by individual differences in narcissism because previous research (e.g., Campbell et al., 2002) has demonstrated that narcissists do not rate themselves as better than average in moral contexts.

Method

Participants

Participants were 100 undergraduates (43 males, 57 females) who received course credit for participating. The mean age of participants in the sample was 19.20 (SD = 1.49). Data were collected in small groups ranging from two to 20 participants.

Procedure

With the exception of the methodological changes described below, the procedures and measures in Study 2 were identical to those in Study 1. The first change was that the number of socially undesirable behaviors for which participants made judgments was reduced from 10 to six. The six negative behaviors that were used in Study 2 were those that were most commonly acknowledged among participants in Study 1. These behaviors involved the following possible admissions: having lied to a significant other, having lied to one’s parents, having cheated on a test, having intentionally hurt someone’s feelings, having stolen something, and having engaged in unprotected sex.
The second and third procedural changes involved the pretesting session (phase 1). In the second procedural change, each time participants acknowledged that they had committed one of these six aforementioned behaviors, they were prompted to explain why they had done so. In the third procedural change, participants completed the 10-item Rosenberg Self-Esteem Scale (RSE; 1965) and the 16-item pair measure narcissism (NPI-16; Ames, Rose, & Anderson, 2006). For each of the 16 pairs of statements on the measure of narcissism, participants were instructed to select the statement that came closest to describing their feelings and beliefs about themselves. The statement in each pair that was indicative of people with grandiose self-images was given a value of one, and the 16 items were summed (potential range 0-16). To decrease the potential of restricted range on the self-esteem scale, the number of response options on the ten-item Rosenberg Self-Esteem Scale was increased to six (1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = slightly agree, 5 = agree, 6 = strongly agree) (potential range 10-60).

The fourth change involved the experimental session (phase 2). In contrast to Study 1 where participants were not privy to the reasons why each target person had committed one of the various socially undesirable behaviors, participants read a quotation from a randomly selected participant from the pretesting session in which the target person explained his or her reason for engaging in the behavior. On the cover page to the questionnaire packet regarding the questions about the six behaviors, participants were explicitly informed that they would be making judgments about six different Ohio University students who had participated in a previous study. This sentence was included
in the instructions to ensure that participants did not mistakenly assume that they were making judgments about a single person who had committed all six of the socially undesirable behaviors. At the bottom of each page, participants were asked to indicate how valid they felt the target person’s reason for engaging in the behavior was on a 7-point Likert scale (1 = not valid at all, 7 = extremely valid).

In the final procedural change, Study 2 also took steps to compare the perceived validity of the reasons given by participants and the target persons. To this end, for each of the six socially undesirable behaviors described in the experimental session (phase 2), participants were asked to indicate how valid they felt the target person’s reason for engaging in the behavior was on a 7-point Likert scale (1 = not valid at all, 7 = extremely valid). To measure the perceived validity of participants’ pretesting session (phase 1) reasons for engaging in the socially undesirable behaviors, three independent coders who were blind to the purpose of the study and the hypotheses rated how valid they felt participants’ reasons for engaging in these behaviors were on the same 7-point Likert scale.

Results

Data from the nine participants who reported during the pretesting session that they had never engaged in any of the six behaviors were excluded from all analyses. Of the remaining 91 participants, the number of behaviors for which they reported engaging ranged from one to six ($M = 3.17, SD = 1.59$). Table 3 displays data from the pretesting session indicating the percentage of the sample that reported engaging in each behavior.
Table 3

*Percentage of Study 2 Participants Who Engaged in Each Behavior*

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lied to Parents</td>
<td>91.3</td>
</tr>
<tr>
<td>Cheated on Test</td>
<td>51.1</td>
</tr>
<tr>
<td>Lied to Significant Other</td>
<td>42.4</td>
</tr>
<tr>
<td>Intentionally Hurt Someone's Feelings</td>
<td>43.5</td>
</tr>
<tr>
<td>Had Unprotected Sex</td>
<td>53.3</td>
</tr>
<tr>
<td>Stole Something</td>
<td>35.9</td>
</tr>
</tbody>
</table>

*Self-Other Comparisons*

As was the case in Study 1, a correlated *t*-test was conducted for each of the three questions. To remind the reader, the composite scores that were compared in the three correlated *t*-tests were comprised of participants’ answers to questions pertaining to the socially undesirable behaviors that participants had previously committed. In contrast to Study 1, there was not a self-other discrepancy in the perceived negativity of the socially undesirable behaviors. Participants’ construals of the negativity of their own misdeeds (*M* = 4.22) were not significantly different than the perceived negativity of others' misdeeds (*M* = 4.50), *t*(90) = -1.43, *p* = .156.

Although a self-other difference was not found for the question regarding the concrete behaviors, the self-other differences found in Study 1 regarding the abstract implications of the behaviors were replicated in Study 2. Giving participants access to the reasons that target persons engaged in the same socially undesirable behaviors as themselves did not reduce their proclivity to conclude that these actions were more representative of other people (*M* = 3.97) than themselves (*M* = 2.45), *t*(90) = -7.96, *p* <
.001. Likewise, the increased amount of contextual information that participants had about target persons did not reduce their tendency to predict that they ($M = 4.71$) were less likely than others ($M = 5.63$) to repeat the socially undesirable behaviors in the future, $t(90) = -4.05, p < .001$ (See Table 4).

Table 4

<table>
<thead>
<tr>
<th>Question</th>
<th>Self</th>
<th>Others</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>How bad is this behavior?</td>
<td>4.22</td>
<td>4.50</td>
<td>-1.43</td>
<td>.156</td>
</tr>
<tr>
<td>How representative is this behavior of you (this person)?</td>
<td>2.45</td>
<td>3.97</td>
<td>-7.96</td>
<td>.001</td>
</tr>
<tr>
<td>How likely are you (this person) to repeat this behavior?</td>
<td>4.71</td>
<td>5.63</td>
<td>-4.05</td>
<td>.001</td>
</tr>
</tbody>
</table>

*Note.* Ratings could range from 1 to 7. Higher values indicate more negative evaluations.

**Self-Esteem and Narcissism**

Means and standard deviations for the variables of interest were as follows: RSE ($M = 46.61, SD = 7.27, \alpha = .83$) and NPI-16 ($M = 5.3, SD = 3.16, \alpha = .71$). The RSE and NPI-16 were correlated, $r = .24, p = .033$.

Next, difference scores were computed for each of the three key dependent measures. These difference scores were computed such that positive scores indicated that participants judged their own socially undesirable behaviors more leniently than they judged identical behaviors committed by others. As predicted, self-esteem moderated
some of the self-other asymmetries that were represented by the difference scores. High self-esteem individuals were more likely to contend that the socially undesirable behaviors that they previously admitted committing were more representative of target persons than themselves, $r = .21, p = .051$. Additionally, high self-esteem individuals were slightly more likely to claim that their own socially undesirable behaviors were less negative than those of target persons, despite the fact that these behaviors were identical, $r = .18, p = .083$. However, individual difference in self-esteem did not moderate the self-other asymmetry regarding the likelihood that participants or target person would commit the negative behaviors again in the future, $r = .15, p = .153$.

The findings regarding individual differences in narcissism were also consistent with predictions. Narcissism scores were not related to any of the three self-other difference scores, all $p$s > .10. These results are consistent with research that has demonstrated that narcissists do not rate themselves as better than average in moral contexts (e.g., Campbell et al., 2002).

Discussion

In contrast to Study 1, evidence for self-protection effects was not found in Study 2 for construals involving low-level construals regarding the specific socially undesirable behaviors. Evidence for self-protection effects was found, however, for construals involving higher-level construals pertaining to the characteristics of the person who committed the behaviors. The reasons that there were not self-other differences in the perceived negativity of the socially undesirable behaviors are unclear, and some possible
explanations for this inconsistency between the results of the first two studies are discussed below.

The lack of consistent findings on the dependent measure that focused on the perceived negativity of the behavior might be due to one or more of the differences between the procedures in the two studies. It may be the case that giving participants access to the reasons that target persons engaged in the behaviors tempered the tendency to imagine the behaviors of others in the worst possible light. However, compared to Study 1, the mean negativity ratings of the behaviors were lower not only for behaviors committed by others, but also for behaviors committed by the self. This suggests that the reason that the self-other difference did not replicate could be related to the procedural change by which the number of socially undesirable behaviors was reduced from 10 to the six behaviors that were most commonly reported in Study 1. Perhaps the four behaviors that participants were least likely to report committing in Study 1 were also more negative than the six behaviors that were used in both Study 1 and Study 2. My tentative conclusion is that the self-other difference regarding the negativity of the socially undesirable behaviors failed to replicate due to a combination of the two aforementioned procedural changes.

Although participants did not construe their socially undesirable behaviors to be less negative than those of others, the self-other differences on the other two dependent measures indicated that they continued to feel that these negative behaviors revealed less about themselves than other people. Consistent with the results of Study 1, although participants did not think that their own negative behaviors were representative of their
current personality at all, they did think that these same behaviors were at least somewhat representative of the people whom they judged. The magnitude of this self-other difference was larger for participants who had higher levels of self-esteem. Similarly, participants indicated that they were less likely than other people to perform these socially undesirable behaviors again in the future. Taken as a whole, the results of Study 2 suggest that even though people knew that their past behaviors were bad, they still viewed themselves as less bad than other bad people.

To summarize, in order to further equate the amount of contextual information that participants had about themselves and target persons in Study 2, participants were provided with information about the reasons that the target persons engaged in the socially undesirable behaviors. The purpose of this important procedural change was to address the alternative account that self-other differences found in Study 1 were due to mundane availability biases that give people greater access to their own thoughts and feelings than those of others (Jones & Nisbett, 1972). The results of Study 2 rendered this alternative account less tenable. Rather, in comparison to the results of Study 1, the results of Study 2 provided stronger evidence that when construing identical behaviors committed by the self and others, participants' tendency to draw less negative inferences about themselves than others can be at least partially accounted for by self-protection motives.
STUDY 3

Study 3 had two primary purposes. First, the third study addressed a potential limitation of the second study. In Study 2, an important procedural step was taken to increase the amount of contextual information that participants had about the target persons. Specifically, participants were granted access to the reasons that the target persons provided during the pretesting session to justify their potentially immoral behaviors. The complete battery of questions for the pretesting session in the psychology department consisted of hundreds of items. With so many questions to answer, the justifications that many of the target persons provided for their behaviors during the pretesting session were fairly brief. Due to this limitation of the pretesting session, it may a stretch to argue that participants had as much information about others' behaviors as they did about their own behaviors.

To address this limitation, some important procedural changes were made in Study 3. The purpose of the procedural changes was to increase the amount of contextual information that participants had about the behaviors of the target persons. By giving participants access to more details about the thoughts and motives of the target persons, the contextual information about the behaviors of the self and others was more equated in Study 3 than was the case in Study 2.

The second purpose of Study 3 was to provide a more nuanced view of how people's beliefs about their negative behaviors relate to their self-concepts. To this end, Study 3 examined construals at a greater number of levels of abstraction than was the case in the previous two studies. As was the case in the first two studies, participants
provided construals regarding the negativity of the concrete behavioral acts and behavioral predictions regarding the likelihood that they would perform the negative behaviors again in the future. Additional questions were added to Study 3 that involved construals at intermediate and high levels of abstraction. In the construals at the highest level of abstraction, which are heretofore referred to as global attributions, participants indicated that extent to which the behaviors suggested that the perpetrator was a bad person. In other abstract construals, participants evaluated themselves and others on traits such as dishonest, mean, dangerous, and untrustworthy.

The construals at intermediate levels of abstraction involved situation-behavior patterns (Mischel & Shoda, 1995). Consider the example of someone who cheated on a test. At the abstract level, the relevant trait is the degree to which the behavior suggests that the perpetrator is dishonest across contexts. At the intermediate level, the relevant situation-behavior pattern is the degree to which the behavior suggests that the perpetrator is a bad student. Because this type of construal focuses on a specific context (academics), it is more localized than a decontextualized trait judgment.

Method

Participants

Participants were 107 undergraduates (42 males, 65 females) who served in the actor role and received course credit for participating. The mean age of these actor participants in the sample was 19.25 (SD = 3.09). There were also 99 undergraduates (38 male, 61 female) who served in the yoked observer role and received course credit for
participating. The mean age of theses observer participants in the sample was 19.04 (SD = 1.19). Data were collected in small groups ranging from two to 20 participants.

Procedure

Stimulus Procedure

An initial set of 62 participants (21 males, 41 females) of approximately the same age as the other participants (M = 19.13, SD = 1.71) answered questions about their past potentially immoral behaviors that would later be viewed by actors and observers. These initial participants did not view the responses of any other participants or complete any response measures.

Actor Procedure

Several modifications were made to the actor-observer methodology that was used in Study 2. Rather than collecting information in a pretesting session regarding participants' justifications for engaging in various negative behaviors, this information was collected in a single experimental session. One reason for this change was to provide greater anonymity to participants who answered questions about their own socially undesirable behaviors. In the previous two studies, participants' were aware that their responses could be tied to their student identification numbers. Knowing that their responses were not completely anonymous may have led some participants to be less than forthcoming about the details surrounding their previous misdeeds. To remedy this concern, participants' responses were completely anonymous, and their data were returned to the experimenter in sealed envelopes and placed in a cardboard box containing additional sealed envelopes.
In another procedural change, participants were prompted to answer a series of questions regarding the negative behaviors that they had committed. In the consent form, participants were informed that other participants would be able to read their anonymous responses. At the top of each page, participants were asked to indicate whether they had ever committed the behavior of interest (e.g., cheated on a test, drove under the influence of alcohol or other substances, intentionally hurt someone's feelings, stole something, and cheated on a significant other). In the event that participants had committed the behavior on multiple occasions, they were instructed to write about the most severe instance.

Next, participants answered a series of open-ended and multiple-choice question about the behaviors that they had previously committed. In these questions, participants provided an array of contextual information regarding their past behaviors. All of these questions were designed to provide observers with abundant information about the actors' behaviors. Actor participants elaborated on exactly what they did, when the behavior occurred, their reasons for these behaviors, and the degree to which they regretted their actions. With previous participants' detailed handwritten responses in hand, participants serving in the observer role were granted access to a full menu of actors' potential self-protection strategies. The ways in which the questions tapped into these potential avenues of self-protection are described below.

First, participants were asked to provide specific details about what they did and how they did it. For example, participants who cheated on a test indicated the subject in which they cheated and the method of cheating that they employed (e.g., looked at someone else's answers, brought a cheat sheet to the test, etc.). Similarly, participants
who drove under the influence of alcohol or illegal drugs indicated how much of the substance they consumed and the number of miles that they drove. The inclusion of the questions about the specific details regarding each behavior ensured that observers did not inadvertently imagine behaviors that were more severe than the behaviors that were actually committed by the actors.

In the previous studies, observers did not have complete information about when the behavior occurred. This raises the possibility that actors recalled events in the distant past while observers judged the behaviors under the assumption that the behaviors occurred more recently than was actually the case. To remedy this potential limitation of the previous two studies, participants in Study 3 indicated when they committed each behavior by selecting from one of the following multiple-choice options: (1) less than three months ago, (2) less than one year ago (but more than three months), (3) between one and three years ago, or (4) over three years ago. This question was included to help observers make more informed judgments regarding the likelihood that target persons would commit each behavior again in the future.

Next, as was the case in Study 2, participants indicated the reasons that they engaged in each behavior. Finally, to delve further into how participants felt about their past misdeeds, they were prompted to explain why they did or did not regret each behavior. The rationale for including these two questions was to provide observers with as much information as possible about the thoughts that actors might use to prevent negative self-referent information regarding the concrete behaviors from spreading to higher level self-concepts.
After providing this contextual information regarding their own behaviors, participants answered seven Likert-scale questions. Two of these questions focused on the concrete behaviors that they committed. These questions were as follows:

1) How bad is this behavior? (1 = not bad at all, 7 = extremely bad)
2) How justified is this behavior? (1 = not justified at all, 7 = extremely justified)

Following these two questions, participants answered additional questions, each of which was at a higher level of abstraction than the one that preceded it. For example, someone who admitted having previously cheated on a test answered the following questions:

3) How likely would it be for you to cheat on a test again in the future? (1 = not at all likely, 7 = extremely likely)
4) To what extent does this behavior indicate that you are a bad student? (1 = not at all, 7 = to a great extent)
5) To what extent does this behavior indicate that you are dishonest? (1 = not at all, 7 = to a great extent)
6) To what extent does this behavior indicate that you are a bad person? (1 = not at all, 7 = to a great extent)

In the next phase of the experiment, participants viewed a randomly selected target person's responses to the open-ended questions about each of the five behaviors. In written instructions similar to those used in Study 2, participants were explicitly informed that they would be making judgments about five different Ohio University students.
Viewing the handwritten responses regarding these behaviors gave participants access to a greater amount of contextual information than was the case in the first two studies. Then, participants answered questions about each randomly selected target person that were similar to the Likert-scale questions that they answered about themselves.

As was the case in the first two studies, actors' construals of target persons represented the reality standard. Based on the assumption that the favorableness of the behavior descriptions offered by actors and target persons did not differ (Alicke et al., 2001, Study 4), when actors evaluated their own behaviors and dispositions less unfavorably than this reality standard, these self-other discrepancies were interpreted as evidence for self-protection motivations. Although this rationale is compelling, the issue represents an empirical question. Thus, to further strengthen the motivational account that I advanced in Study 3, a neutral observer was yoked to each actor. The procedure for yoked observers is described in the next section.

Observer Procedure

Each observer was yoked to a single actor and evaluated not only the behaviors of that actor, but also the behaviors of the target persons who were judged by that actor. As was the case with the actor procedure, the procedure for yoked observers consisted of two phases. In the first phase, each yoked observer read the handwritten contextual information pertaining to the behaviors committed by the actor to whom he or she was yoked. In the second phase, each yoked observer read the handwritten contextual information pertaining to the behaviors committed by the same five target persons who were evaluated by the actor to whom he or she was yoked. In both phases of the observer
procedure, the behaviors in the questionnaire packet were presented in an identical order to the packet that was completed by the actor to whom he or she was yoked.

Results and Discussion

Data from one participant who reported during the pretesting session that she had never engaged in any of the six behaviors were excluded from all analyses. Data from eight participants were also excluded from the analyses because a page was missing from the questionnaire packets. Of the remaining 99 participants, the number of behaviors for which they reported engaging ranged from one to five (M = 2.71, SD = 1.15). Table 5 displays data indicating the percentage of the sample that reported engaging in each behavior.

Table 5

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intentionally Hurt Someone's Feelings</td>
<td>64.6</td>
</tr>
<tr>
<td>Stole Something</td>
<td>46.5</td>
</tr>
<tr>
<td>Cheated on a Test</td>
<td>72.7</td>
</tr>
<tr>
<td>Drove Under the Influence of Alcohol or Other Illegal Substances</td>
<td>53.5</td>
</tr>
<tr>
<td>Cheated on a Significant Other</td>
<td>31.3</td>
</tr>
</tbody>
</table>

First Reality Criterion: Actors-Only Analysis

In the first part of the analysis, a strategy similar to the one used in the first two studies was employed. For each question, a correlated t-test was used to compare actors'
judgments regarding their own behaviors to their judgments regarding identical behaviors committed by others. Only the comparisons for the behaviors that participants acknowledged committing were analyzed. In contrast to the previous two studies, I decided not to aggregate across behaviors. Instead, the self-other differences were examined for each individual behavior.

As was the case in the first two studies, actors' construals of target persons represented the reality standard. Based on the assumption that the favorableness of the behavior descriptions offered by actors and target persons did not differ (Alicke et al., 2001, Study 4), when actors evaluated their own behaviors and dispositions less unfavorably than this reality standard, self-other discrepancies were interpreted as evidence for self-protection motivations.

Low-Level, Mid-Level, and High-Level Construals

Study 3 examined a larger number of levels of abstraction than the previous two studies. By doing so, a more nuanced understanding of the manner in which self-protection affected construals of negative behaviors emerged. The following summary of the results of Study 3 begins with a discussion of low and high levels of abstraction at which self-protection effects were usually not found. Next, the intriguing self-protection effects that were consistently found at intermediate levels of abstraction are discussed.

To remind the reader, the questions pertaining to each behavior were ordered such that they began with concrete construals regarding the behavior. The questions that followed these concrete construals were at a higher level of abstraction and generality than the one that preceded it.
Based on the results of Study 2, for the two questions that pertained to the specific behavior (e.g., how bad is this behavior; how justified is this behavior), I predicted that the self-other differences would be small or nonexistent. This hypothesis was supported for four of the five behaviors (See Table 6). At the level of the concrete behavioral act, participants readily acknowledged that their potentially immoral behaviors were equally as negative as those of their peers. Furthermore, participants admitted that they did not have better reasons for their actions than their peers.

Not only did participants readily acknowledge that their past behaviors were just as negative as those of others, for one of the behaviors, participants judged their own actions to be more negative than those of others. When construing instances in which they intentionally hurt others' feelings, actors rated their own behaviors as more negative ($M = 4.10$) than those of others ($M = 3.48$), $t(63) = 3.21$, $p = .002$.

Whereas the first two questions pertained to a concrete behavior, the remaining questions pertained to the dispositional inferences drawn about the person who committed the behavior. Based on the rationale set forth by researchers on self-protection (Alicke & Sedikides, 2009) and social self-analysis (Alicke, Guenther, & Zell, in press), I predicted statistically significant self-other discrepancies would begin to emerge as the construals proceeded to greater levels of abstraction. Generally speaking, this hypothesis was supported (see Table 6). However, there were boundaries to these effects. For both traits judgments and global attributions, which are located at high levels of abstraction, self-protection effects were nonexistent for three of the five behaviors. For these behaviors, not only were participants reluctant to integrate their misdeeds into their
<table>
<thead>
<tr>
<th></th>
<th>Hurt Someone's Feelings</th>
<th>Stole Something</th>
<th>Cheated on Test</th>
<th>Drove Under Influence</th>
<th>Cheated on Significant Other</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>How bad is behavior?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>4.10**</td>
<td>4.22</td>
<td>3.70</td>
<td>4.44</td>
<td>4.87</td>
</tr>
<tr>
<td>Others</td>
<td>3.48</td>
<td>3.98</td>
<td>3.40</td>
<td>4.63</td>
<td>4.54</td>
</tr>
<tr>
<td><strong>How justified?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>3.62*</td>
<td>2.46</td>
<td>2.89</td>
<td>2.76</td>
<td>2.68</td>
</tr>
<tr>
<td>Others</td>
<td>3.16</td>
<td>2.26</td>
<td>2.83</td>
<td>2.41</td>
<td>2.65</td>
</tr>
<tr>
<td><strong>How likely do it again?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>3.15***</td>
<td>2.12***</td>
<td>2.61***</td>
<td>2.45***</td>
<td>1.86 ***</td>
</tr>
<tr>
<td>Others</td>
<td>4.19</td>
<td>3.87</td>
<td>4.43</td>
<td>4.54</td>
<td>4.03</td>
</tr>
<tr>
<td><strong>Situation-Behavior Pattern</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>2.26**</td>
<td>2.96</td>
<td>2.96</td>
<td>2.74*</td>
<td>2.74*</td>
</tr>
<tr>
<td>Others</td>
<td>2.85</td>
<td>3.59</td>
<td>3.59</td>
<td>3.58</td>
<td>3.58</td>
</tr>
<tr>
<td><strong>Trait</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>2.89</td>
<td>2.98</td>
<td>3.36</td>
<td>2.46</td>
<td>2.73**</td>
</tr>
<tr>
<td>Others</td>
<td>3.03</td>
<td>3.30</td>
<td>3.16</td>
<td>2.87</td>
<td>3.84</td>
</tr>
<tr>
<td><strong>Bad person?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>2.14</td>
<td>2.24</td>
<td>1.88</td>
<td>1.67***</td>
<td>2.03</td>
</tr>
<tr>
<td>Others</td>
<td>2.28</td>
<td>2.50</td>
<td>1.83</td>
<td>2.06</td>
<td>2.55</td>
</tr>
</tbody>
</table>
Note. Ratings could range from 1 to 7. For "How justified?" question, higher values indicate more positive ratings. For all other questions, higher numbers indicate more negative evaluations. Standard Deviations are in parentheses. Situation-behavior patterns: cheated on test = bad student; drove under the influence = reckless driver; cheated on a significant other = bad boyfriend/girlfriend. No situation-behavior patterns were included for the following behaviors: intentionally hurt someone's feelings; stole something. Trait: Intentionally hurt someone's feelings = mean; stole something = dishonest; cheated on a test = dishonest; drove under the influence = dangerous; cheated on a significant other = untrustworthy. *p < .05. **p < .01. ***p < .001.
global self-concepts, but they were also reluctant to draw globally negative conclusions about target persons on the basis of a single behavior.

For two of the five behaviors, some evidence of self-protection was found for high-level construals. In the first of these two behaviors, when asked about the extent to which cheating on a significant other indicated that the person who did so was untrustworthy, participants evaluated themselves less unfavorably ($M = 2.73$) than others ($M = 3.84$), $t(30) = -3.01, p = .005$. Also, when asked about the extent to which this behavior indicated that the perpetrator was a bad person, participants judged themselves marginally more leniently ($M = 2.03$) than others ($M = 2.55$), $t(31) = -1.94, p = .062$.

In the second of these two behaviors, for participants who had driven under the influence of alcohol or illegal substances, evidence of self-protection was found for global attributions. When asked about the extent to which driving under the influence indicated that the person who did so was a bad person, participants evaluated themselves less unfavorably ($M = 1.67$) than others ($M = 2.06$), $t(53) = -3.04, p < .001$. However, participants did not evaluate others as significantly more dangerous ($M = 2.46$) than themselves ($M = 2.87$) on the basis of this behavior, $t(53) = -1.55, p = .128$.

To summarize the results of the actors-only within-participants analysis presented thus far, for low-level and high-level construals, self-protection effects were nonexistent for a majority of the behaviors. With regard to the concrete construals of the specific behaviors committed by both the self and others, participants did not construe their own behaviors to be less negative or more justified than those of others. Similarly, self-protection effects were not found for many of the abstract construals regarding trait judgments and global dispositional attributions. Notable exceptions to this trend were
found at high levels of abstraction for the two behaviors that participants perceived to be the most negative: driving under the influence and cheating on a significant other. Acknowledging that they personally committed very negative behaviors (e.g., driving under the influence, cheating on a significant other) appears to have produced substantial levels of self-threat. The high levels of self-threat experienced by participants seem to have initiated the motivation to protect their self-images by preventing the spread of this negative self-referent event to higher-level concepts (Alicke, et al., in press).

The "sweet spot" of self-protection

While self-protection effects were typically absent for low-level and high-level construals, robust self-protection effects emerged for mid-level construals. Behavioral predictions and contextualized situation-behavior patterns were classified as mid-level construals. In one of the key theoretical contributions of the current research, the results of Study 3 revealed that construals falling at intermediate levels of abstraction lie within a "sweet spot" for self-protection.

Behavioral Predictions. One of the questions at this intermediate level of abstraction asked participants to indicate the likelihood that they and target persons would commit the same potentially immoral behavior again in the future. The claim that behavioral predictions for the self (Wakslak, Nussbaum, Liberman, & Trope, 2008) and others (Nussbaum, Trope, & Liberman, 2003) at time points in the distant future involve abstract construals is well-established. Nevertheless, it is important to acknowledge a nontrivial distinction between the question regarding behavioral predictions and the other questions that involved either mid-level or high-level construals. In terms of the precise
wording, each of the questions regarding situation-behavior patterns, dispositional attributions, or global attributions *explicitly* involved a link between a behavior and an abstract judgment. Specifically, each of these questions began with the phrase, “To what extent does this behavior indicate that …” In contrast, each question involving a behavioral prediction did not begin with the aforementioned phrase. The key point to emphasize here is that the level of abstraction was not the only way that the questions regarding behavioral predictions differed from the other questions.

Replicating the results of the first two studies, for all five of the behaviors, participants predicted that they would be far less likely than target persons to perform the action again in the future (See Table 7). By claiming that they were unlikely to repeat their misdeeds in the future, participants protected their self-image by maintaining that their behaviors represented a mere aberration from their typically meritorious actions. At the same time, participants were quite willing to infer that identical misdeeds were part of each target person's disposition.

The results on these behavioral predictions provided compelling evidence for self-protection motivations because the self-other difference for every behavior was replicated in a design in which target persons had ample opportunity to provide mitigating circumstances surrounding their previous potentially immoral behaviors. That is, the self-other differences on behavioral predictions replicated even when target persons provided a substantial amount of detail regarding exactly what they did, when and where the behaviors occurred, the reasons for their behaviors, and whether they regretted engaging in these behaviors. The finding that self-protection effects persisted under conditions in
which participants had access to abundant contextual information regarding the behaviors of target persons renders a cognitive account based solely on information disparity between the self and comparisons targets less tenable (for similar procedures and conclusions, see Weinstein & Lachendro, Study 2).

Table 7

*Actors' and Yoked Observers' Mean Subjective Likelihood Estimates that Actors and Target Persons Would Perform Each Negative Behavior Again*

<table>
<thead>
<tr>
<th>Intentionally Hurt Someone's Feelings</th>
<th>M Actors</th>
<th>Self vs. Targets</th>
<th>M Yoked Observers</th>
<th>Actor vs. Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self</td>
<td>3.15</td>
<td>-4.12 .001</td>
<td>4.54 4.35</td>
<td>.80 .429</td>
</tr>
<tr>
<td>Targets</td>
<td>4.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stole Something</td>
<td>2.12</td>
<td>-5.85 .001</td>
<td>4.65 4.80</td>
<td>-0.45 .656</td>
</tr>
<tr>
<td>Self</td>
<td>3.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targets</td>
<td>5.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cheated on a Test</td>
<td>2.61</td>
<td>-7.01 .001</td>
<td>5.13 5.04</td>
<td>.49 .630</td>
</tr>
<tr>
<td>Drove Under the Influence</td>
<td>2.45</td>
<td>-6.97 .001</td>
<td>5.15 5.15</td>
<td>.00 1.000</td>
</tr>
<tr>
<td>Self</td>
<td>4.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targets</td>
<td>4.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cheated on a Significant Other</td>
<td>1.86</td>
<td>-5.67 .001</td>
<td>5.23 5.13</td>
<td>.23 .823</td>
</tr>
<tr>
<td>Drove Under the Influence</td>
<td>4.03</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Ratings could range from 1 to 7. Higher numbers correspond to greater subjective likelihood. Actors indicated judgments about themselves and five target persons. Each yoked observer indicated judgments about the actor to whom he or she was yoked and the five target persons who were judged by that actor.
Given that actors predicted that they would be far less likely than target persons to repeat their misdeeds, participants seem to have done a remarkable job convincing themselves that their negative behaviors were a thing of the past and did not represent the "real me." Did yoked observers share actors' optimistic behavioral predictions? The answer is a resounding "no." In contrast to actors, for all five behaviors, yoked observers behavioral predictions for actors and target persons did not differ from one another (See Table 7).

**Situation-behavior patterns.** For three of the five behaviors, an additional question at a slightly lower level of generality than abstract trait judgments was added. This question addressed a specific situation-behavior pattern (Mischel & Shoda, 1995) rather than a more general trait judgment. These intermediate construals fell within the "sweet spot" of self-protection. The results described in the following paragraphs illustrate the subtle, but important distinction between construals at intermediate levels of abstraction and more general trait judgments regarding three behaviors: cheating on a test, driving under the influence of alcohol or illegal substances, and cheating on a significant other.

At the generalized level of trait judgments, when asked about the extent to which cheating on a test indicated that the person who did so is dishonest; participants provided similar responses for themselves ($M = 3.36$) and target persons ($M = 3.16$), $t(72) = 1.0$, $p < .322$ (See Table 6). At the more localized level of the situation-behavior patterns (Mischel & Shoda, 1995), however, when asked about the extent to which cheating on a test indicated that the person is a bad student, participants engaged in self-protection by judging themselves more leniently ($M = 2.26$) than target persons ($M = 2.85$), $t(71) = -$
3.39, \( p = .001 \) (See Table 8). That is, participants were more willing to associate cheating on a test with being a bad student for others than themselves.

Table 8

*Situation-Behavior Patterns: Mean Ratings for Actors and Yoked Observers*

<table>
<thead>
<tr>
<th></th>
<th>( M ) Actors</th>
<th>Self vs. Targets</th>
<th>( M ) Yoked Observers</th>
<th>Actor vs. Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self</td>
<td>Targets</td>
<td>( t )</td>
<td>( p )</td>
</tr>
<tr>
<td>Cheated on a Test</td>
<td>2.26</td>
<td>2.85</td>
<td>-3.39</td>
<td>.001</td>
</tr>
<tr>
<td>(Bad Student)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drove Under the Influence</td>
<td>2.96</td>
<td>3.59</td>
<td>-2.52</td>
<td>.015</td>
</tr>
<tr>
<td>(Reckless Driver)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cheated on a Significant Other</td>
<td>2.74</td>
<td>3.58</td>
<td>-2.26</td>
<td>.031</td>
</tr>
<tr>
<td>(Bad Boyfriend/ Girlfriend)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Ratings could range from 1 to 7. Higher numbers indicate more negative evaluations. Each yoked observer indicated judgments about the actor to whom he or she was yoked and the five target persons who were judged by that actor.

Similarly, at the generalized level of trait judgments, when actors were asked to indicate the extent to which driving under the influence of alcohol or illegal drugs indicated that they were a dangerous person; participants provided responses for themselves (\( M = 2.46 \)) and target persons (\( M = 2.87 \)), that were not statistically different from one another, \( t(53) = -1.55, p < .128 \) (See Table 6). At a more localized level,
however, when asked to indicate the extent to which this behavior indicated that the person was a reckless driver; participants engaged in self-protection by judging themselves more leniently ($M = 2.96$) than target persons ($M = 3.59$), $t(53) = -2.52, p = .015$ (See Table 8). Stated another way, participants were more willing to associate driving under the influence with being a reckless driver for others than themselves.

A set of even more robust self-protection effects emerged for the abstract construals of participants who had cheated on a significant other. As previously noted, for ratings of both the self and target persons, this behavior was perceived to be the most negative of the five behaviors. Compared to their judgments of target persons ($M = 3.84$), participants were more reluctant to describe themselves ($M = 2.73$) as untrustworthy on the basis of their past infidelity $t(30) = -3.01, p < .005$ (See Table 6). A similar self-other difference was found at the more localized situation-behavior pattern of being a bad boyfriend or girlfriend. When asked to indicate the extent to which a past occasion when they cheated on a significant other "indicates that you are a bad boyfriend or girlfriend," participants judged themselves ($M = 2.74$) more leniently than target persons ($M = 3.58$) who engaged in the exact same behavior, $t(30) = -2.26, p = .031$.

In summary, at the concrete level of their behavioral acts, participants did not construe their behaviors to be less negative or more justified than those of others. Similar results were typically found for trait judgments and global attributions. In most cases, participants did not evaluate themselves more leniently than their peers on construals at high levels of abstraction.
For construals at intermediate levels of abstraction, however, robust self-protection effects were found. For all five behaviors, participants predicted that other people would be far more likely than themselves to perform the same actions again. Based on these self-other discrepancies, it appears that although participants did not feel that their past negative behaviors were part of their own dispositions, they did feel that these same behaviors reflected the dispositions of target persons. A similar pattern of results also emerged for situation-behavior patterns. Based on these results, I conclude that behavioral predictions and situation-behavior patterns fall within the "sweet spot" of self-protection.

*Additional Reality Criteria: Yoked Observer Analysis*

The initial rationale for incorporating neutral and impartial yoked observers into the design for Study 3 was to provide a second reality criterion for self-protection. This section discusses some analyses that shed additional light on the self-protection effects that were found in Study 3. First, actors' self-construals were compared to observers' construals of the actor to whom they were yoked. Because yoked observers had access to abundant contextual information surrounding actors' behaviors, evidence for self-protection was demonstrated when actors evaluated themselves more leniently than yoked observers did. Second, actors' construals of target persons were compared to observers' construals of these same target persons. The purpose of this second set of comparisons was to explore the possibility that the salience of actors' own negative behaviors might affect the leniency or harshness of their ratings of other people who committed similar negative behaviors.
Actors' Behaviors

The bulk of the analyses discussed in this paper involved comparisons between actors' self-construals and their construals of other people who engaged in the same negative behaviors as themselves. The design of Study 3 provided an additional standard of comparison: yoked observers' construals of the actor to whom they were yoked. Yoked observers had access to a plethora of information that actors could potentially introduce to minimize the nefarious of their past wrongdoings. With rich contextual detail about actors' behaviors, yoked observers were in a position to make well-informed judgments not only about the concrete behaviors, but also about the degree to which these behaviors were linked to the dispositional characteristics of the person who committed these behaviors. Despite the procedural steps that minimized the information disparity between actors and yoked observers, actors consistently judged themselves more leniently than yoked observers did. These data, which are described below, provided additional evidence to support my argument that the self-other discrepancies that were found in these studies can be at least partially accounted for by motivational concerns.

For each question, a correlated $t$-test was conducted to compare actors' self-construals to the yoked observers' construals of actors. The means and standard deviations that were compared in these correlated $t$-tests are displayed in Table 9. When using the judgments of yoked observers as a reality standard, an even greater number of self-serving construals were found. Consistent with the results of the actors-only analyses, robust self-protective construals emerged at intermediate levels of abstraction.
### Means Ratings of Actors by Self and Yoked Observers in Study 3

<table>
<thead>
<tr>
<th></th>
<th>Hurt Someone's Feelings</th>
<th>Stole Something</th>
<th>Cheated on Test</th>
<th>Drove Under Influence</th>
<th>Cheated on Significant Other</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>How bad is behavior?</strong></td>
<td>Self 4.10</td>
<td>Self 4.22</td>
<td>Self 3.70</td>
<td>Self 4.44*</td>
<td>Self 4.87*</td>
</tr>
<tr>
<td></td>
<td>(1.56)</td>
<td>(1.55)</td>
<td>(1.49)</td>
<td>(2.06)</td>
<td>(1.54)</td>
</tr>
<tr>
<td>Obs</td>
<td>3.66</td>
<td>4.50</td>
<td>3.95</td>
<td>5.25</td>
<td>5.65</td>
</tr>
<tr>
<td></td>
<td>(1.44)</td>
<td>(1.70)</td>
<td>(1.54)</td>
<td>(1.42)</td>
<td>(1.36)</td>
</tr>
<tr>
<td><strong>How justified?</strong></td>
<td>Self 3.62</td>
<td>Self 2.46</td>
<td>Self 2.89</td>
<td>Self 2.76</td>
<td>Self 2.68</td>
</tr>
<tr>
<td></td>
<td>(1.74)</td>
<td>(1.81)</td>
<td>(1.45)</td>
<td>(1.74)</td>
<td>(1.68)</td>
</tr>
<tr>
<td>Obs</td>
<td>3.06</td>
<td>2.26</td>
<td>2.71</td>
<td>2.30</td>
<td>2.10</td>
</tr>
<tr>
<td></td>
<td>(1.59)</td>
<td>(1.42)</td>
<td>(1.43)</td>
<td>(1.31)</td>
<td>(1.47)</td>
</tr>
<tr>
<td><strong>How likely to do it again?</strong></td>
<td>Self 3.15***</td>
<td>Self 2.12***</td>
<td>Self 2.61***</td>
<td>Self 2.45***</td>
<td>Self 1.86***</td>
</tr>
<tr>
<td></td>
<td>(1.40)</td>
<td>(1.30)</td>
<td>(1.55)</td>
<td>(1.78)</td>
<td>(1.39)</td>
</tr>
<tr>
<td>Obs</td>
<td>4.53</td>
<td>4.65</td>
<td>5.14</td>
<td>5.15</td>
<td>5.23</td>
</tr>
<tr>
<td></td>
<td>(1.56)</td>
<td>(1.58)</td>
<td>(1.47)</td>
<td>(1.63)</td>
<td>(1.53)</td>
</tr>
<tr>
<td><strong>Situation- Behavior Pattern</strong></td>
<td>Self 2.27***</td>
<td>Self 2.96***</td>
<td>Self 2.74***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.78)</td>
<td>(1.59)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.63)</td>
<td>(1.53)</td>
</tr>
<tr>
<td>Obs</td>
<td>3.18</td>
<td>4.17</td>
<td>4.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.53)</td>
<td>(1.68)</td>
<td>(1.72)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Trait</strong></td>
<td>Self 2.89</td>
<td>Self 2.98***</td>
<td>Self 3.36***</td>
<td>Self 2.46***</td>
<td>Self 2.73***</td>
</tr>
<tr>
<td></td>
<td>(1.55)</td>
<td>(1.69)</td>
<td>(1.64)</td>
<td>(1.51)</td>
<td>(1.62)</td>
</tr>
<tr>
<td>Obs</td>
<td>3.03</td>
<td>4.30</td>
<td>4.18</td>
<td>3.67</td>
<td>5.13</td>
</tr>
<tr>
<td></td>
<td>(1.46)</td>
<td>(1.59)</td>
<td>(1.60)</td>
<td>(1.74)</td>
<td>(1.63)</td>
</tr>
<tr>
<td><strong>Bad person?</strong></td>
<td>Self 2.14</td>
<td>Self 2.24**</td>
<td>Self 1.83</td>
<td>Self 1.67***</td>
<td>Self 2.03**</td>
</tr>
<tr>
<td></td>
<td>(1.14)</td>
<td>(1.46)</td>
<td>(1.26)</td>
<td>(1.05)</td>
<td>(1.28)</td>
</tr>
<tr>
<td>Obs</td>
<td>2.30</td>
<td>3.11</td>
<td>2.17</td>
<td>3.67</td>
<td>3.32</td>
</tr>
<tr>
<td></td>
<td>(1.31)</td>
<td>(1.65)</td>
<td>(1.23)</td>
<td>(1.74)</td>
<td>(1.74)</td>
</tr>
</tbody>
</table>
Note. Obs = Yoked observers' ratings of actors. Ratings could range from 1 to 7. For "How justified?" question, higher values indicate more positive ratings. For all other rows, higher numbers indicate more negative evaluations. Standard deviations are in parentheses. Situation-behavior patterns: cheated on test = bad student; drove under the influence = reckless driver; cheated on a significant other = bad boyfriend/girlfriend. No person-situation contingencies were included for the following behaviors: intentionally hurt someone's feelings; stole something. Trait: Intentionally hurt someone's feelings = mean; stole something = dishonest; cheated on a test = dishonest; drove under the influence = dangerous; cheated on a significant other = untrustworthy. *p < .05. **p < .01. ***p < .001.
for behavioral predictions and situation-behavior patterns. Moreover, self-protection effects extended beyond this "sweet spot" for self-protection.

For construals of the concrete behavioral acts, actor participants construed their behaviors to be less negative than did yoked observers for two of the behaviors. Participants construed their instances of driving under the influence less negatively ($M = 4.44$) than yoked observers did ($M = 5.25$), $t(54) = -2.40, p = .020$. Similarly, when reflecting on a time that they cheated on a significant other, participants' judgments of their own behaviors were less negative ($M = 4.94$) than the judgments of target persons ($M = 5.54$), $t(54) = -2.61, p = .014$.

At the more abstract level of trait judgments, evidence of self-protection was found. Whereas significant self-other discrepancies at the trait level were only found for a single behavior in the actors-only analysis, self-serving trait judgments were found for four of the five behaviors when the judgments of yoked observers were used as the reality standard. For these four traits, actors rated themselves less negatively than yoked observers at the .001 significance level (See Table 9).

The construals at the highest levels of abstraction asked participants to indicate the extent to which the negative behaviors suggested that the perpetrator was a bad person. Whereas significant self-other discrepancies for these global attributions were only found for a single behavior in the actors-only analysis, self-serving global attributions were found for three of the five behaviors when the judgments of yoked observers were used as the reality standard. When judging the extent to which their
behaviors indicated that they were bad people, actors' global attributions were less negative than those of yoked observers (See Table 9).

In summary, for construals at intermediate levels of abstraction, the results of the yoked observer analyses coincided with those of the actors-only analyses. Furthermore, in the yoked observer analyses, manifestations of self-protection motivations were found outside the "sweet spot" of self-protection. Specifically, self-protection effects were found for construals of two concrete behaviors, abstract trait judgments for four behaviors, and global attributions for three behaviors.

Based on the evidence presented in this section, it appears that the pervasiveness of self-protection effects depends, to some extent, on which accuracy criteria are used. When yoked observers' construals of actors were employed as the accuracy standard, self-protection appeared across more levels of abstraction than was the case when actors' construals of target persons were used as the accuracy standard. This raises an interesting question. Why were self-protection effects less pervasive when actors' construals of target persons were utilized as the reality criteria? The next section explores the possibility that the key to unlocking this apparent puzzle is related to how target persons were judged by actors and neutral observers.

Target Person's Behaviors

In the procedure for actors, after construing their own behaviors, actors formed inferences about five target persons. In the procedure for yoked observers, after construing actors' behaviors, yoked observers formed inferences about these five target persons as well. As discussed in the previous section, this design made it possible to
compare actors' self-construals to the construals that yoked observers formed of them. In addition, this design made it possible to compare actors' construals of target persons to yoked observers’ construals of the same people. Actors and yoked observers had access to abundant contextual information regarding the behaviors of target persons.

For each question, a correlated $t$-test was conducted to compare actors' construals of target persons to yoked observers' construals of target persons. Yoked observers' construals of target persons represented the reality standard. Prior to running the analysis, I speculated that actors' might cast themselves in a less negative light by forming harsh judgments of target persons. As shown in Table 10, for four of the five behaviors, actors actually judged target persons more leniently than yoked observers did for construals at low, intermediate, and high levels of abstraction. In the following paragraphs, I discuss two possible reasons that actors judged target persons more leniently than yoked observers did.

The first reason involves the manner in which the experimental design heightened the salience of actors' own misdeeds. Before actors evaluated target persons, they evaluated themselves on the basis on the basis of their own negative behaviors. Naturally, answering these questions made their own potentially immoral behaviors salient to them. In this context, actors would have felt grossly hypocritical if they had provided extremely harsh evaluations of other people who engaged in the same misdeeds as themselves.

A related reason involves research on the relative preference effect, which refers to the tendency to evaluate other people more favorably, or less unfavorably, when their ethical decisions correspond to one's own (Alicke, 1993). The relative preference effect
Table 10

*Means Ratings of Target Persons by Actors and Yoked Observers in Study 3*

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Hurt Someone's Feelings</th>
<th>Stole Something</th>
<th>Cheated on Test</th>
<th>Drove Under Influence</th>
<th>Cheated on Significant Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>How bad is behavior?</td>
<td>Act 3.51 (1.37)</td>
<td>Act 3.98* (1.79)</td>
<td>Act 3.41* (1.55)</td>
<td>Act 4.63* (1.57)</td>
<td>Act 4.55* (1.72)</td>
</tr>
<tr>
<td></td>
<td>Obs 3.61 (1.48)</td>
<td>Obs 4.73 (1.69)</td>
<td>Obs 4.00 (1.53)</td>
<td>Obs 5.19 (1.60)</td>
<td>Obs 5.45 (1.33)</td>
</tr>
<tr>
<td>How justified?</td>
<td>Act 3.11 (1.60)</td>
<td>Act 2.27 (1.58)</td>
<td>Act 2.85 (1.52)</td>
<td>Act 2.41 (1.73)</td>
<td>Act 2.67</td>
</tr>
<tr>
<td></td>
<td>Obs 3.25 (1.55)</td>
<td>Obs 2.19 (1.36)</td>
<td>Obs 2.56 (1.35)</td>
<td>Obs 2.22 (1.36)</td>
<td>Obs 2.27</td>
</tr>
<tr>
<td>How likely do it again?</td>
<td>Act 4.18 (1.68)</td>
<td>Act 3.89* (1.78)</td>
<td>Act 3.13* (1.57)</td>
<td>Act 4.53* (1.73)</td>
<td>Act 4.09*</td>
</tr>
<tr>
<td></td>
<td>Obs 4.30 (1.59)</td>
<td>Obs 4.69 (1.99)</td>
<td>Obs 3.89 (1.67)</td>
<td>Obs 5.15 (1.63)</td>
<td>Obs 5.00</td>
</tr>
<tr>
<td>Situation-Behavior Pattern</td>
<td>Act</td>
<td>Act 2.84* (1.36)</td>
<td>Act 2.96* (1.72)</td>
<td>Act 3.70*** (1.55)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Obs</td>
<td>Obs 3.32 (1.54)</td>
<td>Obs 3.59 (1.61)</td>
<td>Obs 5.15 (1.79)</td>
<td></td>
</tr>
<tr>
<td>Trait</td>
<td>Act 3.02 (1.52)</td>
<td>Act 3.30** (1.58)</td>
<td>Act 3.13*** (1.57)</td>
<td>Act 2.87** (1.55)</td>
<td>Act 3.91**</td>
</tr>
<tr>
<td></td>
<td>Obs 2.91 (1.58)</td>
<td>Obs 4.15 (1.70)</td>
<td>Obs 3.89 (1.67)</td>
<td>Obs 3.63 (1.76)</td>
<td>Obs 5.00</td>
</tr>
<tr>
<td>Bad person?</td>
<td>Act 2.30 (1.18)</td>
<td>Act 2.48 (1.47)</td>
<td>Act 1.85 (.97)</td>
<td>Act 2.06*** (1.17)</td>
<td>Act 2.61</td>
</tr>
<tr>
<td></td>
<td>Obs 2.43 (1.51)</td>
<td>Obs 3.00 (1.58)</td>
<td>Obs 2.17 (1.41)</td>
<td>Obs 3.63 (1.76)</td>
<td>Obs 3.33</td>
</tr>
</tbody>
</table>
Note. Act = Actors' ratings of target persons. Obs = Yoked observers' ratings of target persons. Ratings could range from 1 to 7. Standard deviations are in parentheses. For "How justified?" row, higher values indicate more positive ratings. For all other rows, higher numbers indicate more negative evaluations. Situation-behavior patterns: cheated on test = bad student; drove under the influence = reckless driver; cheated on a significant other = bad boyfriend/girlfriend. No person-situation contingencies were included for the following behaviors: intentionally hurt someone's feelings; stole something. Trait: Intentionally hurt someone's feelings = mean; stole something = dishonest; cheated on a test = dishonest; drove under the influence = dangerous; cheated on a significant other = untrustworthy. *p < .05. **p < .01. ***p < .001.
occurs not only for choices that produce positive consequences, but also for choices that produce negative consequences (Alicke, Yurak, & Vredenburg, 1996). Consistent with research on the relative preference effect, it appears that, to some extent, actor participants relied on their own values when evaluating the behaviors of target persons.

Summary of All Study 3 Analyses

Study 3 was designed to provide additional evidence to support my argument that self-protection motives affect construals of negative behaviors. When serving in the observer role, participants had tremendous access to the thoughts, motives, and intentions of the target persons. By providing ample contextual information about target persons' negative behaviors, the procedure utilized in Study 3 dramatically reduced the information disparity between actors and observers. Nevertheless, numerous manifestations of self-protection motivations were found across multiple reality criteria.

In the actors-only analyses, actors' self-construals were compared to their construals of target persons. For the most part, self-protection effects were not found for construals of the concrete behaviors. Modest evidence for self-protection was found for some abstract trait judgments and global attributions. For construals at intermediate levels of abstraction, however, actors consistently evaluated themselves less unfavorably than target persons who engaged in the same behaviors as themselves. Based on the results of the actors-only analysis, it appears that behavioral predictions and situation-behavior patterns lie within the "sweet spot" of self-protection.

In other sets of comparisons, the judgments of yoked observers provided additional accuracy criteria. The findings from these analyses complemented the findings from the actors-only analyses. In the first set of comparisons in the yoked observers’
analyses, yoked observers' construals of actors were compared to their construals of target persons. One purpose of these comparisons was to rule out any alternative explanation for the actor-observer differences that might suggest that there was some sort of difference between actors' behavioral descriptions and those of observers. Across every question for every behavior, the construals from these neutral judges were not significantly different from one another. The lack of significant differences across all questions and behaviors provides empirical verification that the favorability of actors' behavioral descriptions did not differ from those of target persons.

The results of this first set of comparisons indicated that actors' evaluated themselves more leniently than yoked observers did. Apparently, actors' attempts to justify their own behaviors fell on deaf ears. Although actors may have been successful at convincing themselves that their negative behaviors did not represent "the real me," they were unsuccessful at convincing neutral observers that their negative behaviors were unrelated to their dispositional characteristics.

The second set of comparisons in the yoked observers' analyses shed additional light on how the experimental paradigm in Study 3 affected actors' judgments of target persons. Before construing the behaviors of target persons, actors provided construals of their own potentially immoral behaviors. For construals at every level of abstraction, actors evaluated target persons more leniently than yoked observers did. There are two likely reasons for these discrepancies. First, it would have been blatantly hypocritical to provide extremely harsh evaluations of people who engaged in behaviors that participants had just acknowledged committing. Although participants did engage in self-protection,
they did so to the extent to which they could regard themselves as reasonable rather than delusional (Sedikides & Gregg, 2003, 2008). Second, people tend to ascribe less blame to people who make moral decisions that are similar to their own (Alicke, et al., 1996).
GENERAL DISCUSSION

People have a strong desire to view themselves as moral and decent human beings. The two most prominent processes by which people pursue this goal are self-enhancement, the tendency to exaggerate one's virtues, and self-protection, the tendency to de-emphasize one's shortcomings. Whereas the purpose of self-enhancement is to approach the goal of a favorable self-image, the purpose of self-protection is to maintain a favorable self-image by avoiding decrements to self-esteem (Alicke & Sedikides, 2009, 2011; Sedikides & Gregg, 2008). Although the end goal of both motivations is a favorable self-image, research suggests that these two motives should be examined separately because each of them operates via distinct mechanisms (Alicke & Sedikides, 2011; Elliot & Mapes, 2005; Sedikides & Gregg, 2008). In the social psychological literature, far more attention has been paid to self-enhancement than to self-protection (Sedikides, 2011). To address this imbalance, the present research focused on self-protection.

In order to focus on a domain that has important implications for one's status as a good person, the present research examined how self-protection motivations affect people's construals regarding their own potentially immoral behaviors. Most of the research on this topic has focused on the excuses, rationalizations, and self-serving attributions that people employ to argue that their behaviors are not as bad as they appear. Less is known about how people construe their behaviors that they know are bad. To fill this gap in the literature, the current research sought a more nuanced view of people's
beliefs about their own bad behaviors. Specifically, the present studies examined how people's negative behaviors become integrated into their self-concepts.

The overarching goal in these three studies was to provide experimental evidence that the magnitude of self-protection varies as a function of levels of construal or abstraction. To this end, I designed three actor-observer studies in which participants provided concrete construals of identical behaviors committed by themselves and others. In addition, participants provided abstract construals regarding the degree to which these behaviors were linked to their self-concepts. In Study 1, participants judged themselves more leniently than target persons for construals at all levels of abstraction. At the concrete level, participants construed their own behaviors to be less negative than those of others. At the abstract level, participants predicted that they would be less likely than others to perform the same actions again in the future. Likewise, in comparison to their construals of other people, participants perceived identical negative behaviors to be more representative of other people than themselves.

Although I argue that self-protection motivations are at least partially responsible for the results of Study 1, a compelling nonmotivational account based on an information disparity between the self and others merits discussion. According to this account, actors have more information about their own thoughts, motives, and intentions than those of others (Jones & Nisbett, 1972). This was certainly the case in Study 1. When construing the behaviors of target persons, actors based their judgments entirely on the actual behaviors because they did not have access to target persons' reasons for their actions. In
contrast, when construing their own behaviors, actors recalled plenty of information about their reasons for committing the behaviors.

In Study 2, an important procedural step was taken to reduce this information disparity and further equate the procedures for actors and observers. Specifically, participants in Study 2 were granted access to the reasons that target persons engaged in the behaviors. For concrete construals, in contrast to Study 1, participants did not judge their own behaviors to be less negative than those of others. Rather, participants perceived their own behaviors to be just as negative as those of target persons.

Acknowledging the negativity of their past behaviors, however, did not prevent participants from engaging in self-protection for construals at greater levels of abstraction. Replicating the results of Study 1, actors predicted that they would be less likely than target persons to repeat the negative behaviors in the future. On a similar note, participants felt that identical negative behaviors were more representative of other people than themselves. Given that participants had access to the reasons that target persons committed the negative behaviors, it appears that mundane availability biases cannot fully account for the two self-other discrepancies that were consistently found across the first two studies. Rather, the evidence suggests that these self-other discrepancies were at least partially accounted for by self-protection motives.

Before You Judge Me, Let Me Explain

Some self-other studies are criticized for adopting procedures that "stack the deck" against comparison targets such that it is virtually inevitable that people will rate themselves more favorably, or less unfavorably, than comparison targets. It could be
argued that the procedures in Study 1 were guilty of "stacking the deck" in this manner because actors did not know why target persons committed the negative behaviors. Hence, the information to which participants had access was quite lopsided: they had plenty of information about themselves and minimal information about comparison targets (Preuss & Alicke, 2009).

In the real world, people whose wrongdoings become public knowledge often defend themselves in the court of public opinion. In Study 2, procedures were adopted to mimic this commonplace real-world phenomenon. Specifically, target persons were given the opportunity to provide justifications for their behaviors. When construing these behaviors and forming inferences about target persons, actors had access to these justifications. As a result of this important procedural change, the disparity in contextual information between the behaviors of actors and those of target persons was reduced. The increase in the available amount of contextual information surrounding the behaviors of target persons represents a step in the right direction. Nevertheless, it may be a stretch to argue that actors knew just as much information about the circumstances surrounding others' behaviors as they did about the circumstance surrounding their own behaviors.

To address this concern, important procedural steps were taken in Study 3 to ensure that actors had access to as much contextual information as possible about target persons' behaviors. As was the case in Study 2, target persons provided justifications for their negative behaviors. Also, participants had additional opportunities to provide other potentially mitigating circumstances surrounding their misdeeds. For example, all target persons were given the opportunity to disclose how recently the misdeed occurred. Just as
politicians sometimes distance themselves from past misdeeds by describing them as college shenanigans and tomfoolery that were committed by the "old me," target persons also had an opportunity to engage in this form of "temporal bracketing" (Baumeister et al., 1990). Similarly, by giving target persons the chance to provide detailed information about the precise details of their actions, actors were able to reach informed judgments regarding the severity target persons' negative behaviors. Furthermore, similar to a court of law, target persons had the opportunity to express remorse for their negative actions. Because actors had access to a wealth of details about the context in which target persons' behaviors occurred, it is difficult to argue that the procedures in Study 3 unfairly "stacked the deck" against target persons.

Despite the increased amount of contextual information available to actors about the behaviors of target persons, the self-other discrepancies found in Study 3 mirrored those found in Study 2. More importantly, the results of Study 3 provided a more nuanced view regarding the relationship between people's beliefs about their own negative behaviors and their self-concepts. As described in the following paragraphs, the results of Study 3 suggested that the magnitude of self-protection effects varies as a function of whether construals are at low, intermediate, or high levels of abstraction.

At the level of the concrete behaviors committed by the self and others, self-protection effects were virtually nonexistent, as participants were willing to acknowledge that their own past behaviors were equally as negative as those of others. For these low-level construals, although participants were not accountable to others per se (Lerner & Tetlock, 1999; Sedikides, Herbst, Hardin, & Dardis, 2002), it seems likely that they
wished to make claims that were, at the very least, believable to themselves (Sedikides & Gregg, 2003, 2008). Moreover, upon reading rich and detailed contextual information regarding the behaviors of target persons, participants fully recognized that their own behaviors were not more justified than those of others. Hence, it would have been difficult for participants to convince themselves that their behaviors were less negative than those of others.

As construals proceeded to intermediate and high levels of abstraction, the emphasis shifted from the concrete behaviors to interpretations of the meaning of these actions for higher-level self-concepts. Stated another way, these construals were designed to measure the extent to which participants felt the potentially immoral behaviors revealed diagnostic information about their own dispositions. For construals at intermediate levels of abstraction self-protection effects were quite robust. For example, when judging identical behaviors committed by the self and others, participants predicted that they would be much less likely than others to perform these negative behaviors again in the future. In a sense, participants asserted that while their past negative behaviors were not part of their own dispositions, these same behaviors were part of the dispositions of target persons. Similarly, in other construals at intermediate levels of abstraction, self-protection effects were found for situation-behavior patterns (e.g., bad student, bad boyfriend, reckless driver).

At the greatest levels of abstraction, self-protection effects were typically not found for trait judgments and global attributions. When indicating the degree to which negative behaviors implied that the perpetrators possessed negative traits or were bad
people, participants did not judge others more harshly than themselves. In most cases, not only were participants reluctant to claim that their own negative actions implied that they possessed negative traits or were bad people, but they also were reluctant to form negative trait judgments or globally negative attributions about others on the basis of a single behavior.

Why were self-protection effects more robust for mid-level construals than high-level construals? It may be the case that reality constraints were stronger for high-level construals than mid-level construals. In the case of construals at the highest levels of abstraction, participants were directly asked to make trait judgments and global attributions about themselves and others on the basis of a single behavior. If participants consistently judged others more harshly than themselves under these circumstances, they ran the risk of appearing hypocritical, if not delusional, to themselves.

In contrast, construals at intermediate levels of abstraction queried participants about the dispositions of themselves and others in a less direct manner. For example, in the case of behavioral predictions, rather than being directly asked to link single actions to decontextualized trait judgments, participants were asked to indicate the likelihood that they (and others) would perform each action again in the future. To a certain extent, making a prediction about the likelihood that someone would repeat a behavior involved a dispositional attribution. For example, if participants claimed that they were less likely than others to cheat on a test again in the future, then they were indirectly ascribing a lower level of dishonesty to themselves than others. Nevertheless, making comparatively optimistic behavioral predictions does not directly challenge reality constraints to the
degree that decontextualized trait judgments do because distal time perspectives activate idealized representations of the self (Kivetz & Tyler, 2006). Furthermore, optimistic behavioral predictions regarding behaviors in the moral arena are facilitated by the greater salience of people's own intentions relative to those of others (Epley & Dunning, 2000, 2006; Kruger & Gilovich, 2004).

Other examples of construals at intermediate levels of abstraction involved situation-behavior patterns. In contrast to decontextualized trait judgments, situation-behavior patterns were localized on a specific context. To illustrate the subtle, yet important, distinction between decontextualized trait judgments and contextualized situation-behavior patterns, I return to the example of cheating on a test. Based on a single instance of cheating on a test, labeling others as more dishonest than themselves, regardless of context, would have unreasonably stretched the bounds of believability. With regard to academic contexts, however, participants who cheated on a test could label target persons who did the same thing as worse students than themselves while still believing that they had a reasonable basis for their self-serving comparative judgments.

To summarize, generally speaking, self-protection effects were absent for low-level and high-level construals. However, self-protection effects were incredibly robust for mid-level construals. Specifically, participants predicted that they would be less likely than others to commit the negative behaviors again in the future. Similarly, for situation-behavior patterns, participants evaluated themselves more leniently than others. In summary, because reality constraints were least severe for mid-level construals, it appears that the "sweet spot" for self-protection lies at intermediate levels of abstraction.
Is Self-Protection Adaptive?

As this paper nears its conclusion, it is important to reflect on the potential benefits of self-protection. Before doing so, I briefly discuss the theoretical origins of research on self-protection. The study of how people deal with internal or external events that threaten preferred favorable self-images has its origins in psychoanalytic theory. In an influential body of work, Sigmund Freud (1915/1961a, 1923/1961b, 1926/1961c) and Anna Freud (1936/1946) proposed a set of defense mechanisms. According to their argument, the purpose of defense mechanisms was to keep socially unacceptable aggressive and sexual urges from reaching conscious awareness and being associated with the self (i.e., the ego).

Modern personality and social psychology has steered away from the idea that personality is heavily based on attempts to hide unacceptable aggressive and sexual impulses. Although the idea that people need defense mechanisms remains quite strong (Baumeister, Dale, & Sommer, 1998), the hypothesized purpose of defense mechanisms has undergone a shift in emphasis. One of the pioneers of this shift, Otto Fenichel (1945) postulated that defense mechanisms are functional because they help protect people's self-esteem. In modern times, the notion of psychological defense has come to encompass virtually any self-aspect that an individual values (Alicke & Sedikides, 2009). As a result, contemporary researchers in personality and social psychology explore potential benefits to the usage of defense mechanisms in normal populations.

Are defense mechanisms adaptive among normal populations? To answer this question, consider the plight of people who rarely utilize defense mechanisms. Every
year, thousands of people visit mental health professionals to learn how to deal with the negative affect and depressed mood that they experience when reflecting on their perceived shortcomings. For many depressed individuals, the negative thoughts associated with a single event can open the gates of a flood of negative cognitions that consume the entire self-system. Defense mechanisms have the potential to quell these self-threats (Baumeister et al., 1998).

At the other end of the spectrum, the overuse of defense mechanisms has been associated with psychopathology and various maladaptive consequences. Although both the complete absence of and the overreliance on defense mechanisms have detrimental side effects, mild forms of defense mechanisms appear to promote well-being. By exploring the attempts of individuals in normal populations to maintain favorable self-images in the face of threats, the current research focused on self-protection in this "optimal margin of illusion" (Baumeister, 1989).

As the previous paragraph illustrates, self-protection enables individuals to maintain a favorable image without delusionally claiming that 100% of their past behaviors were moral and virtuous. This brings us to an important theoretical question: When people engage in self-protection, what exactly are they protecting? According the theoretical framework outlined by Alicke and Sedikides (2009), people defend self-interests that fall at different levels of a hierarchical network. Interests that pertain to the specific behavioral event lie at the bottom of the hierarchy. As one moves toward the top of the hierarchy, interests become more general and abstract. Concerns related to global
self-esteem lie at the top of the hierarchy. When people insist that their negative behaviors do not make them bad people, they are defending a high-level interest.

Judgments regarding personality traits lie just beneath the highest level of the hierarchy. The abstract nature of traits such as honesty and trustworthiness grants people a great deal of latitude in determining the criteria upon which to base their self-evaluations on these traits (Dunning, Meyerowitz, & Holzberg, 1989). In the example of people who have cheated on a significant other (Study 1 and Study 3), rather than evaluating their trustworthiness on the basis of their past infidelity, people may base their self-ratings on this trait on behaviors in other areas of their life where they have behaved more admirably (Critcher, et al., 2011). Moving down the hierarchy to low-level interests that are more closely tied to the actual act of unfaithfulness, people may attempt to employ excuses (e.g., "I was really drunk") (Alicke & Sedikides, 2009; Snyder, 1989; Snyder, et al., 1983) or rationalizations (e.g., "my boyfriend treats me like crap") to justify behaviors that are inconsistent with their typically favorable self-conceptions (Aronson, 1968).

According the theoretical framework outlined by Alicke and Sedikides (2009), the connections among the hierarchical network of interests can be either facilitative or inhibitory. When depressed people reflect on past behaviors that fell short of their moral standards, facilitative connections lead their negative feelings regarding the behaviors to spread to higher-level interests. Self-protection is normative because it prevents this spread from occurring. To a certain extent, the flood of negative cognitions that accompanies depression represents a breakdown of this normative bias (Alloy, Wagner,
Black, Gerstein, & Abramson, 2011). Cognitive therapy is often employed in clinical settings to inhibit this spiral of negative cognitions (Beck, 1976).

In normal populations such as the one used in the current research, people utilize subjective construals to keep negative self-relevant thoughts localized at the level of the concrete behavioral act. That is, when people reflect on their past misdeeds, inhibitory connections are enlisted to serve self-protective functions. These inhibitory connections prevent the negative feelings associated with the behavior from spreading to higher-level interests regarding traits and global self-esteem (Alicke & Sedikides, 2009).

To summarize, the multifaceted and hierarchical nature of the self allows people to acknowledge the negativity of their misdeeds while simultaneously denying the implications of these behaviors for higher-level self-concepts. Stated another way, admitting that they have done some bad things does not prevent people from feeling good about themselves at the global level. Instead, self-protection allows people to convince themselves and others that their negative behaviors do not reflect "the real me."

Limitations

Although I favor a motivational account for the self-other discrepancies that were found in the current set of studies, it is important to revisit the possibility that these discrepancies can be accounted for by mechanisms that make no reference to self-protection or other goal-oriented processes. The most viable non-goal-directed explanation for the results of the current studies involves the notion that participants had more information about their own behaviors than those of others. In Study 3, monumental efforts were taken to address this alternative explanation. Despite the attempts to
minimize the information disparity between the self and others, participants continued to judge themselves less unfavorably than target persons for construals that fell at intermediate levels of abstraction.

Some critics might argue that the procedures in Study 3 did not fully equate the procedures for the self and target persons. In future research, the procedures for the self and other could be fully equated by utilizing a two-phase paradigm similar to the "better than myself" paradigm employed by Alicke et al. (2001). In the first phase, participants would be asked to indicate whether they had committed one or more of a series of socially undesirable behaviors. Additionally, participants would be asked to indicate the reasons that they engaged in each behavior. In the second phase, which would occur several weeks later, participants would evaluate another person who also committed these same behaviors. Participants would also have access to the reasons that the other person committed each of these behaviors. Unbeknownst to participants, the behaviors and the accompanying reasons to which they are granted access would actually be their own. In order to disguise this fact, the reasons would need to be paraphrased in some cases. If the self-other discrepancies found in the previous studies replicated in this novel paradigm, then this "less bad than myself" effect would provide evidence for self-protection motivations in a design where the procedures for the self and other were fully equated.

Another potential limitation involves a statistical concern regarding potential Type I errors. Because participants judged multiple target persons, it was not possible to conduct a Multivariate Repeated-Measures Analysis of Variance to control the experimentwise error rate. Thus, it is possible that several of the self-other discrepancies
reported in this paper are Type I errors. Despite this legitimate concern, it is important to highlight the remarkably consistent pattern of self-serving construals that were found for behavioral predictions and situation-behavior patterns in Study 3. The robust nature of the results for these mid-level construals bolsters my claim that there is a "sweet spot" for self-protection.

In future research, to control the experimentwise error rate, the procedures that were employed in Study 3 could be modified by conducting a two-part study. The first phase would be identical to the procedure that was utilized in Study 3. In the second phase, rather than evaluating multiple target persons, participants would form construals of an individual target person who committed exactly the same behaviors as themselves.

By adopting this procedure, a repeated measures Multivariate Analysis of Variance (MANOVA) could be conducted. There are at least two important benefits to modifying the procedures in this manner. First, a repeated measures MANOVA would help to determine the appropriateness of examining univariate effects and to reduce the likelihood that any significant self-other discrepancies would be due to chance. Second, because actors would evaluate a single target person, the results of the actor-observer comparisons for each question could be aggregated across multiple behaviors. Hence, the results could be displayed and summarized in a clear, concise, and organized manner.

Conducting a two-part study in this manner offers another crucial advantage. In contrast to the procedures for actors in the second phase of Study 3, actors would not form evaluations of other people immediately after their own negative behaviors were made salient. By minimizing the salience of actors’ past misdeeds, actors’ inferences of
target persons would not be colored by their desire to avoid forming blatantly hypocritical judgments of others. In a two-part study, I speculate that self-protection effects would be found not only for mid-level construals, but also for many high-level construals. Specifically, I would expect find results similar to those found in Study 3 in the first set of comparisons involving yoked observers’ judgments (See Table 9).

When interpreting the results of the current set of studies, I frequently drew distinctions among low-level, mid-level, and high-level construals. These distinctions regarding the degree of abstraction of these construals were based on the theoretical frameworks outlined by Alicke and Sedikides (2009) and Liberman et al. (2007). Although my argument regarding the degree of abstraction of each construal is theoretically-based, it is possible that participants did not perceive each construal for each behavior to be more abstract than the one that preceded it.

To address this limitation in future research, the abstractness of each construal should be pretested in an independent sample of raters (e.g., Schellekens, Verlegh, & Smits, 2010). If the researcher asks participants to "rate the level of abstraction" of various judgments, then some participants may not fully comprehend the questions. Therefore, during this pretesting, it will be important to phrase the questions in ways that are easily interpretable by undergraduate participants with a wide range of academic abilities. To clarify, in the example of cheating on a test, my claim that traits are more abstract than situation-behavior patterns could be tested by asking participants to answering the following question: "How much does the following tell you about what a person is like in general?" First, participants would be asked to answer this question with
regard to the phrase "bad student." Second participants would be asked to answer this question with regard to the word "dishonest."

The specific behavior examples used in the current studies were based on the results of a pilot study in which participants listed the three worst things that they had ever done. The behaviors that were mentioned most frequently by participants in the pilot study were included in the current set of studies. In each of the self-other comparisons in Study 1 and Study 2, participants' responses to each question were aggregated across the behaviors that they acknowledged committing. Because some behaviors were perceived by participants to be more negative than others, aggregating across behaviors poses some potential interpretational problems.

To address this limitation in future research, extensive pretesting of the behaviors should be conducted (for an example, see Chadwick, Bromgard, Bromgard, & Trafimov, 2006). Although it will be a costly effort, extensive pretesting will help assure the equivalence of the perceived negativity of behaviors. As a result of extensive pretesting, the researcher will have a stronger basis for conclusions that are based on judgments that are aggregated across multiple behaviors.

Finally, regardless of which direction researchers pursue in future examinations of the role that self-protection motivations play in construals of potentially immoral behaviors, it will be important to examine how self-protection effects in this context are moderated by individual differences. Specifically, it would be interesting to better understand the subset of individuals who do not exhibit self-protection tendencies. Evidence from Study 2 suggested that self-protection effects for one of the dependent
variables were less pronounced for low self-esteem individuals. Future research should explore the relationship between self-esteem and self-protection in greater detail.

Future Directions

The current research focused on people's construals of their own negative behaviors. Future research should expand the scope beyond the realm of negative behaviors. After all, behaviors in the moral arena represent one of many types of negative self-relevant information that people might wish to minimize. One potentially fruitful avenue for future research would be to examine people's construals of what they perceive to be their biggest character flaws.

In a recent pilot study, I asked participants to indicate their biggest character flaw. For purposes of clarification, the phrase "worst thing about you as a person" was listed in parentheses. Participants were also asked to "give specific examples of three different behaviors that reveal this flaw of yours."

The aforementioned prompts effectively generated honest responses. Based on extant research on self-protection, I was somewhat concerned that people might minimize their shortcomings by only mentioning minor flaws. As it turned out, this concern was unfounded. Not only were people quite willing to admit their character flaws, but they appeared to list things about themselves that they truly did not like. Some of the most frequently mentioned flaws were related to procrastination, anxiety, low self-esteem, personal insecurities, and other negative self-relevant thoughts. The informative responses from participants in this pilot study point toward some promising directions for future research. Some of these future directions are discussed below.
First, future research on self-protection should examine whether people think that their character flaws are less bad than the flaws of others. In addition to listing their own five biggest character flaws, undergraduate participants would list the five biggest character flaws of the typical student at their institution. Next, participants would rate the negativity of both their own character flaws and those of the average student at their institution. Evidence for self-protection would be demonstrated if participants perceive their own flaws to be less negative than those of the typical college student. To solidify claims based on the motivation to minimize one's shortcomings, participants' judgments could be compared to the external reality check provided by the judgments of neutral observers.

Second, future research should explore the degree to which people think that others share their flaws. In other words, do people think that everybody has the same flaws that they do, or do they think that their flaws are unique to themselves? Two possibilities are evident. On the one hand, research on the false consensus effect (e.g. Krueger, 2000) suggests that people would overestimate the degree to which others share their character flaws. On the other hand, people might underestimate the degree to which others have the same character flaws as themselves. Recent research on pluralistic ignorance has demonstrated that people believe they experience negative self-relevant emotions more frequently than others. People think they are more alone in their emotional struggles than they actually are because social norms discourage people from publicly discussing many of their insecurities with others (Jordan, Monin, Dweck, Lovett, John, & Gross, 2011).
Third, future research should examine people's metaperceptions regarding their greatest flaws. Metaperceptions refer to people's beliefs about what others think of them. Are people aware of what other people think are their worst character flaws? It is quite likely that people have incomplete information about how others view them, especially with regard to their own negative characteristics. When forming their social networks, people tend to gravitate towards people who provide predominately positive feedback (Frey & Tropp, 2006; Herbert & Vorauer, 2003; Kenny & Depaulo, 2003; Shrauger & Schoeneman, 1979). Even when people communicate negative feedback to others, social norms encourage them to sugarcoat their criticisms (Blumberg, 1972; DePaulo & Bell, 1996). As a result, people might underestimate the severity of their flaws or perhaps even remain blissfully unaware of their worst personality characteristics.

In this section, future research directions on three ways in which people might minimize their character flaws have been discussed. First, participants might ignore, fail to notice, or not even be exposed to information from peers about their character flaws. Second, even when people are aware of their character flaws, they might convince themselves that their flaws are not as bad as those of others. Third, people might claim that almost everyone has the same flaws that they do. By no means do these three possibilities represent an exhaustive list of strategies that people might employ to minimize their character flaws. Without a doubt, a full menu of other self-protection options awaits further empirical exploration.

In yet another self-protection option, a fourth potential line of research could explore whether people think their own characters flaws are fixed or malleable. To
provide a concrete example, consider people who perceive a quick temper to be their biggest character flaw. Some individuals might minimize this flaw by envisioning attainable possible selves (Markus & Nurius, 1986) that do not possess this negative characteristic.

To empirically examine this possibility, participants would provide a self-rating on this negative characteristic. Next, participants would predict their standing on this characteristic ten years from now. To provide a standard of comparison, participants' predictions would be compared to those of well-acquainted observers. Based on the self-other comparisons found in the current research for predictions regarding negative behaviors, I predict that people would make optimistic predictions regarding their potential self-improvement on their biggest character flaw.

To extend this idea, one fifth potential line of research could explore the degree to which people think their flaws characterize their overall personalities. To empirically examine this issue in future research, a single-item pictorial measure could be created (an example of such a pictorial measure in another domain is Aron, Aron, and Smollan's Inclusion of Others in the Self Scale, 1992). In this pictorial measure, respondents would select a picture that best describes the degree to which their biggest flaw characterizes their overall personality from a set of Venn-like diagrams. Each of these diagrams would represent a different degree of overlap of two circles. One circle would represent a person's overall personality. The other circle would represent that person's greatest character flaw. The figures would be designed so that the degree of overlap progresses
linearly, creating a seven-step, interval-level scale. Greater overlap between these two circles would indicate that the flaw is a major component of the person's personality.

To provide a reality criterion with which to measure self-protection, the degree of overlap indicated by the self would be compared to the degree of overlap ascribed to the person by an acquaintance who knows the person well. Self-protection would be empirically demonstrated if participants claim that the flaw represents a less significant portion of their overall personality than neutral observers do. In other words, in contrast to the judgments ascribed to them by observers, people might think that their character flaws represent minor blemishes on an otherwise rosy backdrop of positive qualities.

Concluding Thoughts

To conclude, although research suggests that self-enhancement biases are pervasive, many researchers and laypersons have qualms about this portrayal of human nature because it makes people seem way too self-confident. The evidence garnered from the extant literature on self-evaluative biases, the current research, and the future lines of research that I have proposed may eventually converge to describe people's self-evaluative motivations in a manner that can be embraced by psychologists from a wide variety of perspectives. In the future, researchers may conclude that self-enhancement and self-protection are complementary forces that help people deal with some fundamental insecurities. If this occurs, social psychology will have came full-circle to suggest a notion that is not far removed from the idea of defense mechanisms that Freud introduced over seventy years ago.
REFERENCES


doi:10.1207/s15327957pspr1003_5


