

January 01, 2008

Beyond emotion and reason: the social function of morality

Piercarlo Valdesolo
Northeastern University

Recommended Citation

Valdesolo, Piercarlo, "Beyond emotion and reason: the social function of morality" (2008). *Psychology Dissertations*. Paper 4.
<http://hdl.handle.net/2047/d10016302>

This work is available open access, hosted by Northeastern University.

Beyond Emotion and Reason:
The Social Function of Morality

A Dissertation Presented By

Piercarlo Valdesolo

To

The Department of Psychology

In partial Fulfillment of the requirements for the degree of

Doctor of Philosophy

Northeastern University

Boston, MA

June, 2008

Beyond Emotion and Reason:
The Social Function of Morality

by

Piercarlo Valdesolo

ABSTRACT OF DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate School of Arts and Sciences of
Northeastern University, June, 2008.

ABSTRACT

Three studies explore the hypothesis that morality is defined by evolved social-cognitive abilities and social emotions tailored to the development of trusting and cooperative relationships with others. Study 1 will show a fundamental bias in moral judgment that functions to elevate one's conception of one's own moral reputation relative to others, and show that this bias extends to group-level social identities. Study 2 will address the cause of this bias, showing that the phenomenon does not result from self-serving automatic intuitions, but rather from the effect of self-serving motivated reasoning which operates in direct competition with more basic and automatic selfless intuitions. Study 3 will extend the social model beyond moral judgment to moral action, investigating the implications of group-membership on the experience of prosocial emotions as well as the frequency and degree of altruistic behavior. Taken together these studies suggest a need for a shift away from traditional arguments debating the relative importance of emotion and reason in moral judgment, and towards a consideration of the function of our moral capacities. Subjects were recruited from the Northeastern University introductory psychology participant pool.

TABLE OF CONTENTS

Abstract	3
Table of Contents	4
Introduction	5
Study 1	16
Study 2: The Duality of Virtue: Deconstructing the Moral Hypocrite	23
Study 3: For Whom do we feel? Similarity as a determinant of empathy and altruism	34
General Discussion	41
References	48

INTRODUCTION

“In-group morality” has been posited as a fundamental moral intuition (Haidt and Graham, 2007) – binding individuals together into cooperative and co-dependent relationships within social groups, while also fomenting divisive ideological, religious, and political beliefs and fueling intergroup conflict. Such intuitions are thought to play a causal role in moral judgment and action. However, their origin can only be surmised and, experimentally, little is known about the nature of the moral intuitions that subserve group-level social identities and how they might contribute to in-group cohesion or intergroup discord. The present studies provide an initial examination of these phenomena, revealing the underlying processes and contextual sensitivities of moral judgments regarding transgressions enacted by and/or committed against members of social ingroups and outgroups.

By utilizing a methodology designed to induce *in vivo* moral transgressions in the lab, I will show a fundamental bias in moral judgment, which I term moral hypocrisy. Moral hypocrisy occurs when individuals evaluate their own transgressions much more leniently relative to others’ identical transgressions. I will show that this hypocrisy readily extends to the group level, affording even tangentially-similar others the same moral latitude with which we judge ourselves (Study 1). The ease with which hypocrisy spreads to group-level social identities suggests that in addition to overcoming normative differences in morality across cultures, individuals must be wary of the seemingly fundamental ingroup-oriented biases activated when discussing moral issues. I will also present evidence suggesting that this phenomenon does not result from self-serving automatic intuitions, but rather from the

effect of self-serving motivated reasoning which operates in direct competition with more basic and automatic selfless intuitions (Study 2).

Using a similar paradigm, I will also examine moral perceptions with respect to the nature of the victim (as opposed to the transgressor). Here, I show that transgressions committed against in-group members are perceived to be much more morally objectionable than the same actions committed against nonaffiliated others. Additionally, I will demonstrate that this asymmetric sensitivity is directly driven by enhanced feelings of empathy for in-group victims, which consequently motivates not only harsher judgments against transgressors but also costly prosocial behavior aimed at helping victims of our ingroups (Study 3).

Reason vs. Emotion

Philosophers have long been concerned about the role of emotion and reason in moral judgment. Traditionally most theorists championed a rationalist model wherein emotion was considered to be at best unrelated to morality (Leibniz, Descartes), and at worst the seed of all sin and the root of base desires. The passions were often characterized as carnal, animalistic impulses that needed to be reigned in and stifled by the powers of reason in order to lead a virtuous life.

This worship of reason began to erode in the 18th century as philosophers began to argue for a necessary role of emotion in moral life, believing that the attainment of moral knowledge flowed from our immediate feelings as opposed to our reasoned thoughts (Hume, 1739-1740/1969). These theorists argued that humans have a built-in sense of right and wrong, with feelings of positivity associated with moral acts and feelings of negativity

associated with immoral acts. “Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them” (Hume, p. 462).

The rift between rationalist and intuitionist models widened in the late 19th century. Freud saw moral judgments as driven by nonconscious motives and feelings (Freud, 1900/1976). To a large extent, his theory of psychosexual stages and emphasis on the important role of the unconscious sought to explain how children deal with and manage their strongly felt immoral desires through the development of the super-ego. Towards the middle of the century, behaviorists began to argue that morality was little more than the acts that societies happened to teach, either through imitation, reward, or punishment, belittling mental processes. Indeed, the famous Bobo doll study (Bandura, Ross & Ross, 1961) illustrated that children could learn aggressive behavior detached from the influence of emotional or reasoning processes. In this study, children exposed to aggressive exemplars, for example an adult violently hitting an inflatable doll with a mallet, were more likely to engage in subsequent aggressive behavior than children exposed to neutral exemplars.

The cognitive revolution of the 1960’s drove the most divisive stake between rationalist and emotion-based models. Lawrence Kohlberg spearheaded the movement with a research program dedicated to undermining “irrational emotive theories” (1971, p.188) and replacing them with a cognitive-developmental approach based on the work of Jean Piaget (1932, 1965). Using Piaget’s method of interviewing children to discern how they solve moral dilemmas, Kohlberg presented participants with hypothetical dilemmas and studied how they resolved these conflicts. The most well-known scenario, known as the Heinz dilemma, reads as follows:

A woman was near death from a special kind of cancer. There was one drug that the doctors thought might save her. It was a form of radium that a druggist in the same town recently discovered. The drug was expensive to make, but the druggist was charging ten times what the drug cost him to produce. He paid \$200 for the radium and charged \$2,000 for a small dose of the drug. The sick woman's husband, Heinz, went to everyone he knew to borrow the money, but he could only get together about \$1,000 which is half of what it cost. He told the druggist that his wife was dying and asked him to sell it cheaper or let him pay later. But the druggist said: "No, I discovered the drug and I am going to make money from it". Heinz got desperate and broke into the man's store to steal the drug for his wife. Should Heinz have broken into the laboratory to steal the drug for his wife? Why or why not?

Kohlberg developed a 6-stage model conceptualizing the progression of the manner in which participants reason about moral issues such as the Heinz dilemma. People in the first two stages act in accordance with a principle of avoiding punishment and satisfying self-interest (pre-conventional stage). Respondents in this stage might argue that it is fine for Heinz to steal the drug since it was in his wife's best interest. The next two stages are characterized by motives to conform to the rules of society and to obey authority (conventional). Respondents in this stage might argue that it was wrong to steal the drug since theft would violate an important societal norm. People in the final two stages view laws as social contracts and act in accordance with universally applied ethical principles (post-conventional). Respondents in this stage might argue that Heinz acted appropriately by adhering to a principle of causing the least harm to others. Kohlberg inspired several decades of research devoted to understanding how we think about moral rules, and to what degree our action covaries with these rules (Darley, 1993; Nucci & Turiel, 1978; Turiel, 1983; Turiel, Killen, & Helwig, 1987). As a result, arguments accumulated for the causal role of conscious reasoning in moral judgment, to the exclusion of theories emphasizing a role for emotion-based processes.

This focus on moral reasoning was short-lived, however, as the affective revolution of the 1980's sparked a resurgence of interest in the moral emotions. Reacting to Kohlberg's

theories, researchers began to turn to findings in evolutionary psychology (Trivers 1971; Hamilton, 1964) and primatology to argue that morality was, in fact, grounded in a set of automatic intuitions that make people *feel* for one another, and to bring matters of cooperation, cheating and trust to the forefront of moral concerns. This new trend argued that emotions such as sympathy, empathy, shame, gratitude, and vengeance directly affect our moral lives, and contribute more directly to moral judgment than reasoning processes. Haidt, Koller, and Dias (1993) supported this “intuitionist” explanation with cross-cultural studies which revealed that moral judgments were more often driven by emotional reactions to situations than reasoned arguments. In this study, participants were asked to read situations involving offensive but harmless acts (e.g. eating one’s pet dog, cleaning one’s toilet with the national flag). While respondents stated that no one was directly harmed by these acts, they nonetheless considered them to be universally wrong. Similar results were found in a study of American political conservatives and liberals. Opinions regarding moral issues such as homosexuality, incest, and masturbation were predicted by affective reactions to the scenarios as opposed to reasoned arguments (Haidt & Hersh, 2001), suggesting that ideological divides may be driven more by differences in intuitive reactions rather than reasoned principles. According to Haidt, the frequency with which people condemn an action in the absence of supportive reasons undermines the causal role of moral reasoning in decision making, suggesting that moral judgments are driven by gut feelings of right and wrong, while reasoned arguments act merely as *ex post facto* rationalizations. He calls this phenomenon “moral dumbfounding” – the inability to generate supportive reasons for a judgment combined with an adamant refusal to change that judgment. Research on the automaticity of social cognition lent further support for an intuitionist model of moral

judgments, finding that many automatic processes can cause cognitive and behavioral changes in the absence of conscious reflection or awareness (Bargh and Chartrand, 1999).

Dual Process Models

Despite the accumulation of evidence for the role of automatic processes in moral judgment, our moral lives cannot be fully explained by an intuitionist account. Consider the following dilemma as an illustrative example. A runaway trolley is hurtling down the tracks towards five workmen who will be killed if it proceeds on its present course. You stand at a fork in the tracks next to a large switch. You can save these five workmen by flipping the switch and diverting the trolley onto a different set of tracks, one that has only one workman on it, but if you do this that person will be killed. Is it morally permissible to flip the switch, turning the trolley and thus saving five workmen at the cost of one? An overwhelming majority of people say “yes” (Greene 2001, 2004; Valdesolo & DeSteno, 2006).

Now consider a variant of this dilemma. Once again, a runaway trolley is hurtling down the tracks towards five workmen. This time, you are standing on a footbridge overlooking the tracks next to a large stranger. The only way to save the five workmen is to push this stranger off the bridge and into the path of the oncoming trolley. The stranger’s body will knock the trolley off the tracks, killing the stranger but saving the five workmen. Is it morally permissible to push the stranger, saving the five workmen at the cost of one? An overwhelming majority of people say “no” (Greene 2001, 2004; Valdesolo & DeSteno, 2006). It would be logically consistent for those who would flip the switch to also push the stranger, as the tradeoff in both dilemmas is equivalent. Yet respondents give logically

inconsistent responses. What makes it morally permissible to sacrifice one life to save five in the switch dilemma but not in the footbridge dilemma?

Neuroimaging has revealed that dilemmas such as the footbridge dilemma produce increased activation in emotion-related brain centers as well as in centers normally used for deliberative reasoning; considering moral violations, such as inflicting direct harm, elicits prepotent negative reactions that appear designed to inhibit amoral acts. Accordingly, in the absence of a prepotent reaction to the consideration of a moral violation, such as in the switch dilemma, neuroimaging reveals decreased activation in emotion-related brain centers in combination with an increase in centers associated with deliberative reasoning (Greene et al., 2001). In short, the footbridge dilemma engages people's emotions in a way that the switch dilemma does not. The intuitive aversion to physically pushing someone to their death drives respondents to the non-utilitarian response for the footbridge dilemma, and the absence of such an aversion to flipping a switch leads to relatively more utilitarian decision making in the switch dilemma. The logically inconsistent responses to these two dilemmas reflect the differential contributions of emotion-based and reason-based processes to each decision.

A second neuroimaging study found that the infrequent selection of the logically appropriate option in the footbridge dilemma is associated with heightened activation of deliberative centers aimed at cognitive control, suggesting that the automatic negative reaction must be disregarded if a utilitarian judgment is to be made (Greene et al., 2004). Furthermore, increased deliberation, as measured by reaction time, was predictive of utilitarian judgment, suggesting that the ultimate arbiter of ethical choice for such dilemmas resides in individuals' abilities and motivations to engage in controlled analysis (Greene et al.,

2004; Valdesolo & DeSteno 2006). However, a later study predicted that choice could be altered not only by increased engagement in reasoning processes, but also by increased input from emotion-based processes. This was found to be the case, as environment-induced feelings of positivity at the time of judgment reduced the perceived negativity, or aversion “signal,” of the potential moral violation (e.g. inflicting direct harm) and, thereby, directly altered moral judgment (Valdesolo & DeSteno, 2006). In sum, both reason based and emotion based processes have been shown to play an important causal role in moral judgment.

As a result, researchers have begun to converge on a dual-process model of moral judgment (Cushman, Young, & Hauser, 2006; Greene Nystrom, Engell, Darley, & Cohen, 2004; Haidt, 2001; Pizarro & Bloom, 2003; Valdesolo & DeSteno, 2006). According to this view, an intuitive process, which automatically alters hedonic states in response to specific types of socially relevant stimuli, is theorized to work in tandem with more domain-general, consciously-guided processes that underlie abilities for abstract reasoning, simulation, and cognitive control. Processes at both levels are sensitive, to differing degrees, to morally-relevant events or principles (e.g., cause no direct harm, utility, self-protection), with the eventual decision output representing some confluence of the processes.

Beyond Emotion and Reason

Though a dual-process model of moral judgment represents significant progress from wholly rational or intuitive models, most research has now turned to determining which of these processes contributes relatively more to particular domains of moral life. While parsing the specific roles that emotion and reason play in any given decision may lead to a greater understanding of the nuts and bolts of our moral circuitry, a broader picture of

morality is too often ignored. In this dissertation, I will hypothesize a possible function of the processes involved in moral judgment. Specifically, that at every level, these processes have evolved in order to promote adaptive social functioning.

Crudely put, the processes guiding moral judgment may have been forged by evolutionary pressures sensitive to group-level concerns. The ability to coexist peacefully and cooperate with others within groups composed of kin and non-kin alike has been characterized as an adaptation of immense import for survival and genetic replication (Hamilton, 1964; Trivers, 1971). The protection and wider availability of resources afforded by group living conferred advantages to group members that those unable to adjust to group life were denied. Consequently, it stands to reason that humans, as well as close primate ancestors, should have evolved psychological mechanisms suited to effectively functioning within a group context. I believe that an important component of such mechanisms falls within the moral domain.

In short, moral judgments should be sensitive to group-level social identities. The cognitive processes which contribute to these kinds of moral decisions, whether explicitly reasoned or based on feeling states, should help us successfully navigate social dilemmas and ultimately help us share in the survival benefits associated with group membership. Specifically, the processes underlying morality answer questions such as, who can I trust? With whom should I cooperate or help? Who deserves punishment or blame? By answering these questions, individuals attempt to identify those in their environment with whom they stand to benefit from engaging in costly exchanges. Group membership stands as a cue for an increased probability in the likelihood of forming relationships based on this reciprocal altruism (Trivers, 1971). Once such a relationship has been established, its perpetuation, via

the personal benefits one accrues in the long-term by acting selflessly in the short-term, reinforces the strength of the group at large. Social cohesion increases and social order is cemented.

I hypothesize that moral emotions and moral reasoning processes are importantly related to these kinds of group relations. This paper will describe three separate experiments exploring this possibility.

The Studies

Study 1 will investigate the flexibility of our moral principles. If morality is inherently social, then judgments should be sensitive to group issues such as reputational concerns.

People should be motivated to elevate their moral status relative to those around them, as doing so would help garner the social rewards associated with being seen as a moral individual. Reciprocal altruism will flourish only with those who emit signals suggesting a high likelihood of reciprocating, so it is in the interest of the individual to promote one's moral reputation. Therefore, as opposed to consistently applying moral rules to all, individuals should be more lenient when judging their own transgressions relative to unfamiliar others' identical transgressions. This hypocritical bias should extend to the group level affording similar others the same moral leniency with which we judge ourselves. Such decision making would function to solidify and promote our group members' moral standing, presumably protecting them from censure and conferring upon them the same social rewards that a moral reputation would bring oneself.

Study 2 will address the processes involved in these biases, determining how our feelings and thoughts regarding our own, our group members', and unfamiliar others'

transgressions interact to shape our judgments. That is, a primary goal of Study 2 will be to uncover the levels at which the hypocrisy bias may emerge. Finding a discrepancy in how we reason about these transgressions, and consequently a stability in the way we feel about our own versus others' transgressions, would reveal the deep-seated nature of our sensitivity to moral concerns, consistent with a social model of morality. Indeed, having an aversion to others' transgressions would be functional in an environment in which adaptive behavior entails forming secure and trusting relationships, and several theorists have speculated that emotions operate to promote and solidify precisely these types of relationships (Trivers, 1971; Frank, 1988; Bartlett & DeSteno, 2006).

Study 3 tests the applicability of the social model to moral action. It may be the case that sensitivity to social concerns applies to judgments of others' transgressions but not necessarily to actions – we may be willing to condemn immoral behavior but to what extent will we engage in costly action to help a victim? If the establishment of cooperative social relationships is indeed advantageous, and if group membership serves as a cue to identify those individuals most likely to reciprocate, then costly prosocial behavior will be preferentially directed towards ingroup victims relative to unfamiliar victims. We should be motivated to help those with whom we perceive some sort of similarity. Study 3 will also test the mechanism by which this helping occurs, hypothesizing that group differences in altruism will be driven by group differences in empathy. Simply put, we should feel more empathy for similar victims than unfamiliar victims, and this discrepancy will drive helping behavior. Empathy functions to bind groups together by being preferentially triggered in the face of a suffering ingroup member.

Testing the Hypotheses

To elicit the phenomena of interest I modified a paradigm developed by Batson et al. (1997). In the current paradigm, a target individual faces a dilemma representing a conflict between self-interest and the interest of another. The individual is required to distribute a resource (i.e., time and energy) to themselves and another, and can do so either fairly (i.e., through a random allocation procedure) or unfairly (i.e., through personal selection). In conditions where the question of interest relates to fairness judgments of one's own actions, the participant plays the role of the target and is later asked to evaluate the morality, or fairness, of his actions. In conditions where the question of interest relates to fairness judgments of others' actions or fairness judgments of transgressions against particular victims, the participant plays the role of an observer wherein they view another individual, a confederate, acting in the unfair manner (i.e. selecting the better option for herself), and subsequently evaluate the morality of this act. All studies employ the same methodology with additions which will be noted as necessary.

Study 1. Moral Hypocrisy: Social Groups and the Flexibility of Virtue

As noted above, Study 1 was interested in the degree to which participants would judge their own and in-group members' fairness transgressions differently than the same transgressions enacted by others. We hypothesized that participants would exhibit a hypocritical bias in their evaluations of these judgments (i.e. judge their own transgressions to be less objectionable than others' identical transgressions), and that this hypocrisy would extend to members of participants' in-groups. If such hypocrisy commonly exists, there is good reason to theorize that it might extend to individuals beyond the self. Specifically, group affiliation might stand as a basic limit on the radius of one's "moral circle," qualifying ingroup members for the same leniency which individuals apply to their own transgressions.

To the extent that the group stands as an important source of self-definition, one may have an interest in protecting the sanctity of that entity. Indeed, “ingroup morality” has been posited as a fundamental moral intuition (Haidt & Graham, 2007). Use of minimal groups to demonstrate such variability would constitute the most strict and compelling test of our hypothesis, and was therefore employed.

I defined hypocrisy as the discrepancy between the fairness judgments for a transgression when committed by the self or by the other. To determine if any emerging hypocrisy would extend beyond the self, I measured the discrepancy between the fairness judgments for a transgression when committed by an ingroup member or an outgroup member. If hypocrisy emerged here as well, it would suggest flexibility in the radial boundaries of hypocrisy as a function of target affiliation with the self, and provide strong support for the sensitivity of moral concerns to social relations.

Methods

Seventy-six subjects were randomly assigned to one of four conditions. In all conditions, subjects judged the fairness of the same action, and this judgment served as the primary dependent variable. Condition 1 and Condition 2 measured judgments of one’s own transgression and judgments of another’s transgression, respectively. Conditions 3 and 4 were designed in order to measure judgments of moral transgressions when the enactor was either an ingroup or outgroup member, respectively.

Condition 1: Judging One’s Own Transgression.

On entering the lab, a participant was seated at an individual workstation, given a brief introduction to the experiment, and told to begin the computerized tasks. The

instructions explained that the experimenters were examining performance on two different types of tasks, and that any participant would only complete one of the tasks. The first task (i.e., the green task) consisted of a brief survey combined with a short photo hunt that would take 10 minutes to complete. The second task (i.e., the red task) consisted of a series of math and logic problems combined with a longer and somewhat tedious mental rotation task that would take 45 minutes to complete. Following the task descriptions, participants were informed by the experimenter that the research team was also evaluating a new participant assignment protocol meant to reduce experimenter bias. Therefore, certain participants would be randomly selected to make condition assignments for themselves and others.

Participants then read the following instructions:

In order for the experimenters to remain blind to condition assignments, you must assign either yourself or the next participant to the green condition and the other of you to the red condition. Some people feel that giving both individuals an equal chance is the fairest way to assign the tasks.

If you would like to use a randomizer to assign conditions, please move to the computer behind you and follow the instructions. The decision is entirely up to you. You can assign yourself and the other participant however you choose. The other participant does not and will not know that you are assigning conditions.

The randomizer was a computer program designed to assign the participant to the “red” condition following a few demonstration trials conducted by the experimenter in which it alternated between conditions to guard against participant suspicion. The experimenter then left the room.

After assigning conditions to themselves and another, all participants completed a questionnaire designed to assess their affective state. Participants rated how well 4 different emotion descriptors (*sad, down, negative, gloomy*) represented their current feeling state using 7-point scales. Mean scores on this measure were used to assess negative affect (Chronbach's $\alpha = .94$).¹ Following this measure, participants responded to questions regarding the assignment procedure which were presented as a way to collect opinions on the new protocol. Embedded in a small set of distractors was the target question: "How fairly did you act?" Participants responded to this question on a 7-point scale ranging from "extremely unfairly" to "extremely fairly." The session was then terminated and participants debriefed.

Condition 2: Judging Another's Transgressions.

In this condition, participants' primary task involved evaluating the actions of another individual who completed a procedure identical to the one completed by the participant in Condition 1. Here, participants were informed that their role was to act as an impartial observer to provide feedback to experimenters regarding use of the new assignment protocol by other participants. These other participants were in fact confederates. To accomplish this goal, participants were informed that they would be seated in the room with an individual taking part in an experiment and therefore able to observe his actions and responses to the experimental protocol through the use of a yoked computer. That is, participants would be able to see on their screen what the other participant was reading and selecting in real time. Participants received the following instructions on their screen:

Your computer is connected to the adjacent computer. Another participant will be completing an experiment on that computer and you will be asked to follow along and observe on your screen everything that he reads and does. Note that the other participant will be unaware that this is happening. After approximately 5 minutes of observing, you will be asked to rate the new assignment protocol in terms of clarity and design as well as answer some questions concerning the performance of the participant.

Participants were asked if they understood their task, and if so to click the mouse to connect the two computers. From this point on, they were presumably observing the other participant's screen and were asked not to touch their computer until it disconnected and automatically moved them along to the evaluations.

After the computers had "connected," the participant waited in her seat while the experimenter brought in the second participant (i.e., the confederate). The confederate was told that all instructions would be on the computer and to begin the experiment by clicking the mouse. The confederate then simultaneously clicked his mouse as well as a second mouse surreptitiously connected to the back of the participant's computer. The mouse clicks set off a timed presentation which created the illusion that the participant was observing, on her own monitor, the confederate go through the instructions and assign himself the "green" condition and a future participant the "red" condition without using the randomizer. After observing the confederate's choice, the participant's computer "disconnected" and brought her to an evaluation section where, embedded in a set of distractors, she answered the following target question: "How fairly did the participant act?" using the same scale as in Condition 1.

Condition 3: Judging an outgroup member's transgression

Two confederates played the roles of other participants in each session. Participants were brought into the lab along with the two confederates and all were seated at individual work stations. The experimenter instructed one of the confederates to work on a separate task while the participant and remaining confederate completed a questionnaire requiring frequency estimates for different types of events (e.g. *How many black bears are there in Massachusetts?*). They were then “categorized” by the computer into one of two groups: overestimator or underestimator (cf. DeSteno, Dasgupta, Bartlett, & Cajdric, 2004). Participants were always categorized as the opposite group of the confederate. Following the group manipulation, the experimenter entered the laboratory and escorted both confederates to a different laboratory, presumably to work on the next part of the experiment. The experimenter then returned to the laboratory, and told the participant that their next task would be to observe and evaluate a new experimenter-blind assignment procedure that one of the other participants was about to complete (the out-group confederate). From that point forward, the experiment unfolded as in the “judging other” condition.

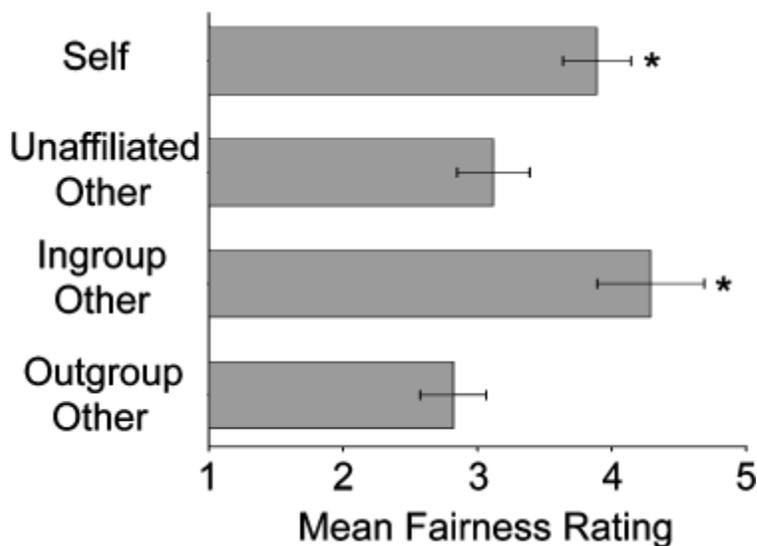
Condition 4: Judging an ingroup member's transgression

Condition 4 mirrored Condition 3 with the exception that the participant was always categorized in the same group as the confederate during the minimal groups manipulation.

Results and Discussion

Two subjects in Condition 1 were removed from analysis for acting altruistically or using the randomizer; all others assigned themselves to the green condition. In accord with predictions, a planned contrast (with contrast weights of 1, -1, 1, and -1 for Conditions 1

through 4, respectively) confirmed the existence of hypocrisy at both the individual and the group levels, $F(1, 72) = 11.75, p = .001$. As shown in the figure below, hypocrisy appears to be a fundamental bias in moral reasoning: Individuals perceived their own transgressions to be less objectionable than the same transgression enacted by another person. Moreover, this hypocritical view extended to judgments of others. Subjects readily excused other individuals' unfair acts if these others belonged to subjects' social groups.



While at face value this finding may not be entirely surprising for groups with long histories of conflict (e.g., Israeli vs. Palestinian factions), this study revealed the fundamental nature of this bias by demonstrating its emergence even among novel, or minimal, groups. Individuals more readily excused others' fairness transgressions even when they bore no other similarity to themselves beyond receiving identical feedback on a trivial numerical estimation task. Indeed, subjects viewed transgressions committed by emergent in-group members to be as acceptable as their own. Though no direct measure of group identification

was taken, minimal group manipulation have consistently been shown to create group-level social identities (Tajfel, H. & Turner, J.C., 1986).

Evidence of hypocrisy at both the individual and group levels adds to the growing view of the context-dependent nature of moral reasoning (cf. Valdesolo & DeSteno, 2006). At a basic level, preservation of a positive self-image appears to trump the use of more objective moral principles, keeping our moral status elevated relative to unknown others and qualifying us for the social rewards commensurate with a moral reputation. This stain of hypocrisy actively spreads to group-level social identities, presumably protecting group members from censure while also potentially inflaming intergroup discord.

Study 2: The Duality of Virtue: Deconstructing the Moral Hypocrite

As described in Study 1, *moral hypocrisy* refers to a fundamental bias in moral judgment in which individuals evaluate a moral transgression enacted by themselves to be less objectionable than an identical transgression enacted by others. Of high import for intergroup relations, this asymmetric leniency extended to others as a function of their relation to the self: a transgression enacted by a member of an ingroup is perceived to be of equal acceptability to the same transgression enacted by the self, but to be more acceptable than the identical behavior enacted by an outgroup member or non-affiliated other (Valdesolo & DeSteno, 2007).

Given both its apparent elemental status and practical import, moral hypocrisy stands as a phenomenon quite worthy of further investigation. At present, the existence of

moral hypocrisy is clear, but the mechanisms that underlie it remain clouded. Accordingly, the present experiment focuses on examining the process(es) by which moral hypocrisy emerges. Doing so would potentially cast further light on the inherently social nature of moral decision making. Hypocrisy may be driven by how individuals *feel* or how they *reason* about their own transgressions versus others. A discrepancy in reasoning processes, and consequently a likeness in the way we feel about our own versus others' transgressions, would reveal the deep-seated nature of our sensitivity to how we treat others, consistent with a social model of morality. Having an intuitive aversion to cheating others would be functional in an environment in which adaptive behavior entails forming secure and trusting relationships.

Uncovering the Hypocritical Mind

To elicit hypocrisy, we used the same paradigm as described in the previous study. Individuals faced a dilemma representing a conflict between self-interest and the interest of another (Valdesolo & DeSteno, 2007; cf. Batson, Thompson, Seufferling, Whitney, & Strongman, 1999). In this paradigm, some participants were required to divide a resource (i.e., expended time and energy) between themselves and another, and could do so either fairly (i.e., through a random allocation procedure) or unfairly (i.e., through personal selection of the preferred option). They were later asked to evaluate the morality, or fairness, of their actions. Other participants viewed a separate individual, who was a confederate, acting in the unfair manner toward another (i.e., selecting the better option for herself) and subsequently evaluated the morality of this act. Again, we defined hypocrisy as the discrepancy between the fairness judgments for this same transgression when committed by the self or by the other.

By modeling hypocrisy as discrepant moral judgments, we might expect that its underlying mechanisms would operate in a fashion similar to that of any other moral evaluation. As previously discussed, recent research in the psychology of morality has begun to converge on a dual-process model of moral judgment (Cushman, Young, & Hauser, 2006; Greene Nystrom, Engell, Darley, & Cohen, 2004; Haidt, 2001; Pizarro & Bloom, 2003; Valdesolo & DeSteno, 2006). We believe that moral hypocrisy can be understood within this framework.

Conceptualizing hypocrisy as a dual process model, however, leads to competing predictions regarding precisely how these two classes of processes interact to produce the phenomenon. More specifically, hypocrisy could be driven by a discrepancy in automatic intuitions in response to one's own versus another's transgressions. That is, individuals might display an automatic positivity bias for their own transgressions relative to others', with higher order processes simply functioning to create post hoc justifications for "gut-level" decisions (cf. Haidt, 2001). Alternatively, hypocrisy might be driven by differential activation of higher order cognitive processes geared toward justification and rationalization of one's own transgressions. That is, although individuals might have negative automatic reactions to both their own and others' transgressions, they may engage in more consciously motivated reasoning when judging their own transgressions in order to maintain a positive self-view.

Distinguishing between these two competing explanations has important practical implications for developing strategies geared toward curbing this disturbingly familiar phenomenon. Indeed, deciding whether intuitions should be fostered or overcome hinges upon whether or not people have automatic aversions to their own as well as others'

violations of fairness norms. Additionally, disentangling these two explanations would speak directly to the fit of a social model of moral judgment. The unearthing of an intuitive aversion to one's own transgressions would be suggestive of a process that developed in order to foster cooperative and trusting social relationships.

Two Alternative Models

As noted, there is reason to believe hypocrisy could emerge in two ways based on a dual process model of moral judgment. Mounting evidence suggests that humans may have evolved an intuitive aversion to violations of equity, with similar aversions evidenced by comparative primate species (Brosnan & de Waal, 2003; Hauser, 2006). It has also been hypothesized that humans and our close ancestors have evolved specific social emotions designed to foster cooperation and trust with others (Bartlett & DeSteno, 2006; Frank, 1988, Haidt & Graham, 2007), suggesting an important role for emotional responses meant to stymie self-serving behavior and, thereby, avoid negative social consequences. Accordingly, violations of fairness stand as a strong candidate to engender a spontaneous and immediate negative reaction regardless of the enactor, suggesting that hypocrisy might emerge from more deliberative processes.

Several lines of research suggest that higher order processes could indeed be driven by strong biases aimed at rationalizing and justifying a transgression (Bandura, 1990,1996; Haidt, 2001; Trivers, 1985). In this case, the intuitive system would favor a more "moral" judgment in accord with a basic fairness norm, but conscious control systems might work to generate a more "immoral" judgment that nevertheless may serve to protect one's self-

image. However, when judging another's transgression, higher order processes should not temper the intuitive response as the motive for self-image preservation is less relevant.

Alternatively, recent findings demonstrate that disruption of brain regions responsible for certain types of higher order cognitive control decreases aversion to inequity when playing certain economic games (Knoch, Leone, Meyer, Treyer, & Fehr, 2006), suggesting that automatic reactions might be geared toward engendering self-serving, as opposed to fair, behavior. Indeed, this finding aligns with much research suggesting that humans possess an automatic positivity bias with respect to evaluations involving the self. For instance, tests of implicit self esteem consistently reveal a seemingly ubiquitous generalized positive evaluation of self (Greenwald and Farnham, 2000; Yamaguchi et al., in press). In a similar vein, much work has suggested that exaggerated perceptions of mastery and unrealistic optimism are characteristic of normal human thought (Taylor & Brown, 1988). When taken in combination with recent research demonstrating that both motivational state (Balcetis & Dunning, 2006) and chronic views regarding one's abilities (Ehrlinger & Dunning, 2003) can influence more basic cognitive processes such as perception, these findings suggest that chronic views of oneself as a moral individual, as well as motives to appear as such, might lead to positively biased spontaneous evaluations of one's own transgressions relative to those of others.

If so - if the intuitive system does not generate an immediate aversion, or generates a lesser one, to one's own transgression - then hypocrisy might simply arise as a result of discrepant, spontaneous evaluative responses. According to this view, the intuitive system would favor a more "moral" judgment in accord with a basic fairness norm when contemplating other's transgressions, but favor a more "immoral" judgment in accord with

an automatic positivity bias when contemplating one's own. One might not be as sensitive to transgressions that bring one immediate benefits. If true, these intuitions would work in concert with higher order processes which would serve to provide justification and rationalization for the behavior.

Testing the Dual-Process Model of Hypocrisy

The present experiment seeks to disentangle these competing explanations. If hypocrisy derives from competition between a negative affective response to any violation of fairness coupled with conscious efforts aimed at justifying the behavior when enacted by oneself, then hypocrisy should disappear when efforts aimed at conscious control are constrained (cf. Greene et al, 2004; Valdesolo & DeSteno, 2006). However, if hypocrisy arises because of discrepant automatic intuitions generated in response to our own versus another's transgressions, then constraining conscious control should have no effect on judgments of one's own transgressions.

To examine this question, we used a factorial design, crossing judgments of self- and other transgressions with a manipulation of cognitive constraint: a 2 (Enactor: Self vs. Other) X 2 (Load: Control, Cognitive Load). In the control conditions, differences between the fairness judgments of participants who judged their own behavior after acting unfairly (i.e., assigning themselves a preferable task and another a less preferable task via personal choice) relative to judgments of a confederate's decision to act unfairly in the same manner, were expected to demonstrate the usual hypocrisy effect identified in the previous study (Valdesolo and DeSteno, 2007). In line with this previous finding, we predicted that participants would rate their own transgressions as more "fair" than the same behavior when

enacted by another. Moreover, we predicted that reduced ability for controlled processing would alter the relative causal force of processes contributing to judgment, directly addressing the nature of the dual mechanisms underlying hypocrisy. If manipulation of cognitive constraint has no influence on judgments of participants' own transgressions, it would suggest that the model is one wherein hypocrisy arises from biased automatic intuitions. However, if increased cognitive constraint results in more "moral" judgments of participants' own transgressions (i.e., actions are judged to be more unfair) and thereby attenuates hypocrisy, these findings would suggest that hypocrisy arises from discrepant volitional efforts aimed at justifying transgressions when enacted by the self relative to others. We expect that these manipulations will have no effect on participants' judgments of the confederate's transgressions, as motivated reasoning processes should not be engaged when judging others' violations. Conditions involving the judgments of others will function not only as a baseline for computation of the hypocrisy measure, but also to show that any effects of the manipulations do not result from influencing moral judgments in a global manner (e.g., increased cognitive constraint decreases the perceived fairness of any actions, whether enacted by the self or another).

Method

Ninety one individuals (58 females, 33 males) participated and were randomly assigned to one of four experimental conditions. Conditions 1 and 2 mirrored the "judging one's own transgression" and "judging another's transgression" described previously. As the load conditions are variants of these conditions, only the primary differences in design will be noted. In all conditions, participants judged the fairness of an identical action, which served as the primary dependent variable. As noted, we employed a 2x2 experimental design

crossing self and other judgments with a cognitive load variable. Presentation of materials and data collection were accomplished using pc's running MediaLab software.

Condition 3 was a replication of Condition 1 with the exception that participants made fairness judgments under cognitive load. The load manipulation came directly after participants assigned tasks to themselves and the other, thereby affecting only moral judgment and not behavior. Cognitive load was manipulated using a digit-string memory task (cf. Gilbert & Hixon, 1991). Participants were told that the experimenters were interested in how people make judgments when they are distracted. To simulate distraction, they would be asked to remember a string of digits at the same time that they were responding to a series of questions. Participants were told that a string of seven digits would appear on the screen before each question. They would then have to answer the question within 10 seconds, immediately after which they would have to recall the digit string that had preceded the question. Participants were also told that it was extremely important to provide the most accurate answers possible for questions comprising the assignment evaluation measure. The primary dependent variable consisted of the fairness question presented and scaled as in Condition 1 and embedded in the series of distractor questions completed under load.

Condition 4 mirrored Condition 2 with the exception that participants made judgments under cognitive load, using the same load manipulation as in Condition 3.

Results and Discussion

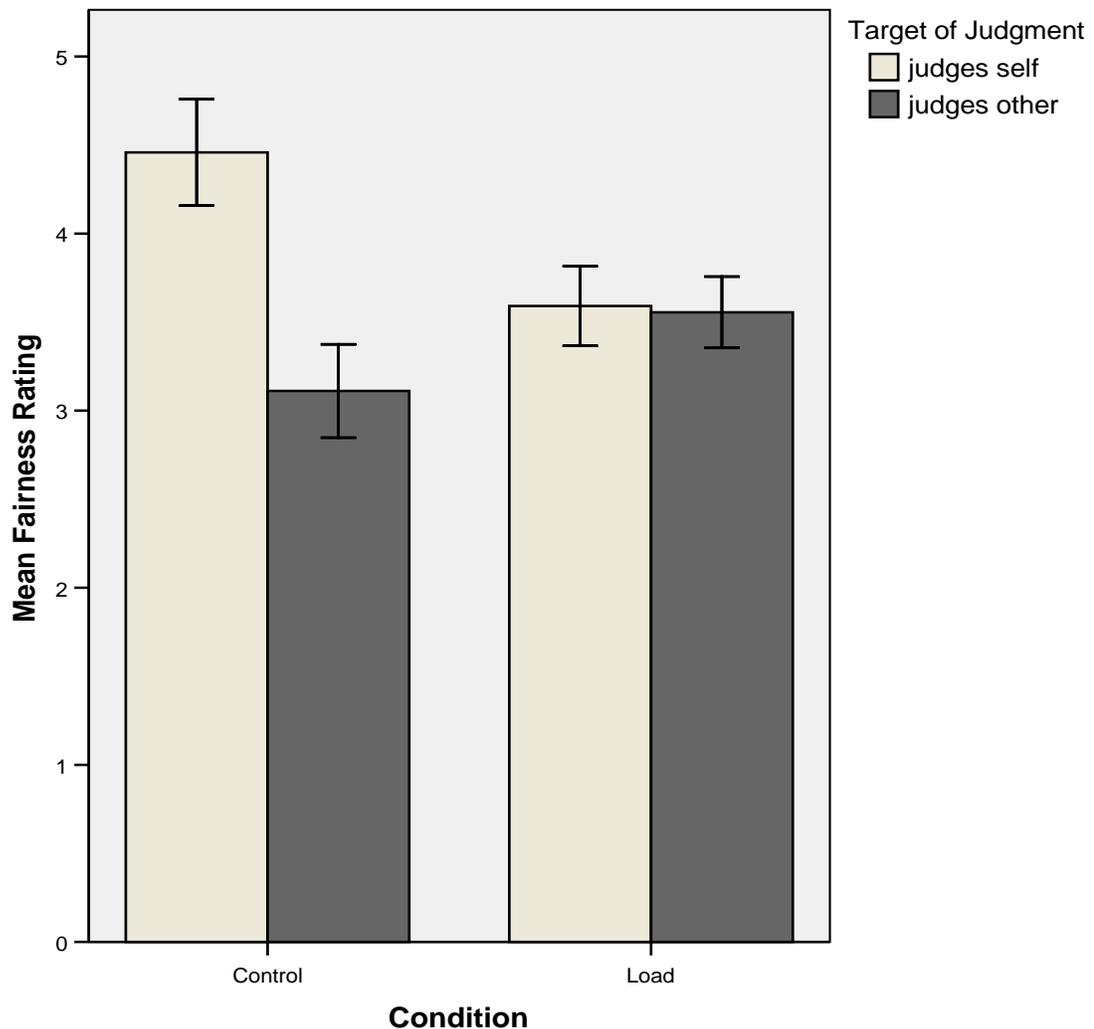
Participants in conditions involving judgments of their own transgressions were removed from analysis if they did not commit a transgression. That is, only those participants who assigned themselves the “green” (i.e., preferable) condition and who did

not use the randomizer were included in the analysis. As in the previous study, those who immediately acted either altruistically or in accord with the fairness norm were a substantial minority; this group consisted of 7 (8%) participants spread almost equally across the two relevant conditions (i.e., Conditions 1 and 3).

Moving to the full factorial design, an ANOVA revealed the predicted moderation of moral judgment as an interactive function of the load factor, $F(1,87) = 6.20, p = .02, \eta^2 = .067$ (see figure below – error bars represent one SE). Constraints on effortful correction (i.e., cognitive load) resulted in the disappearance of the hypocrisy effect; participants experiencing load judged their own transgressions to be as unfair as the same behavior when enacted by another, $t(38) = .115, p = .909, d = .031$. Indeed, a planned contrast revealed that only judgments of one's own actions in the control condition (i.e., Condition 1) significantly exceeded judgments in any of the other three conditions, which showed no reliable differences among themselves, $F(1, 87) = 5.119, p = .003$.

A comparison of Conditions 1 and 2 revealed that moral hypocrisy readily emerged in the absence of load. The same fairness transgression was judged to be substantially more moral when enacted by the self than when enacted by another, $t(49) = 3.39, p = .001, d = .949$, thereby replicating previous findings which demonstrate that hypocrisy represents a fairly

basic bias in moral judgment.



The present study provides strong evidence that moral hypocrisy is governed by a dual process model of moral judgment wherein a prepotent negative reaction to the thought of a fairness transgression operates in tandem with higher order processes to mediate decision making. Hypocrisy readily emerged, but the effect disappeared under conditions of cognitive constraint. Inhibiting control prevented a tamping down or override of the intuitive aversive response to the transgression. Of import, these findings rule out the

possibility that hypocrisy derives from differences in automatic affective reactions towards one's own and others' transgressions. Rather, when contemplating one's own transgression, motives of rationalization and justification temper the initial negative response and lead to more lenient judgments. Motivated reasoning processes are not engaged when judging others' violations, rendering the prepotent negative response more causally powerful and leading to harsher judgments.

These findings are also noteworthy for demonstrating that controlled processing need not always function to "correct" more basic, intuitive responses (cf. Greene et al., 2004), but rather can be subject to less admirable motives - one possible motive being the protection of self-image. Indeed, they show that the interplay between intuitive and volitional moral reasoning is sensitive not only to abstract moral principles but also to more selfish motivations, as evidenced by the overwhelming majority of participants who acted unfairly when assigning tasks.

Despite this disconcerting result, the unearthing of a prepotent negative response to one's own transgressions, and conversely the absence of an automatic positivity bias, reveals an adventitious relationship between moral judgment and hypocrisy. The detection of a low-level sensitivity to fairness transgressions, even at the cost of one's own potential short-term gain, adds to the growing body of evidence dispelling theories which describe morality as a tenuous and fragile "vener" laid over a core of selfish impulses (de Waal, 2006). Instead, it seems likely that humans have evolved strong intuitions which, though selected to promote long term self-interest via reciprocal altruism, can represent moment-to-moment instances of pure selfless concern.

Study 3: For whom do we feel? Similarity as a determinant of empathy and altruism

The previous two studies have been exclusively concerned with moral judgment. However, a social model of morality should not only predict our decisions, but more importantly, our moral actions. Study 3 explores why, to what extent, and with whom we will engage in costly prosocial behavior. If morality is inherently social then our actions should reflect a sensitivity to certain characteristics of the victim (i.e. whether they reflect upon the self). Consequently, group membership might stand as a social cue, signaling an increased likelihood of reciprocity in the future, which should in turn increase the likelihood of engaging in helping behavior. In short, we should more frequently help others with whom we perceive some degree of similarity. Moreover, I believe that empathy is sub served by low-level perceptions of similarity between oneself and a victim, thus serving the specific function of motivating target-specific helping behavior. The degree to which empathy is elicited, and consequently the degree to which it motivates prosocial behavior, depends upon the degree of similarity perceived with a victim.

Traditionally there has been a divide between theorists who argue that empathy fosters prosocial behavior via motivating the desire to relieve one's own distress at the sight of others' pain (Cialdini, R.B., Schaller, M., Houlihan, D., Arps, K., Fultz, J., & Beaman, A.L., 1987; Neuberg, S.L., Cialdini, R.B., Brown, S.L., Luce, C., Sagarin, B.J., & Lewis, B.P., 1997), and those who believe that empathy fosters prosocial behavior via motivating a selfless desire to relieve others' distress (Batson, D.C., Batson, J.G., Griffitt, C.A., Barrientos, S., Brandt, J.R., Sprengelmeyer, P., & Bayly, M.J., 1989; Batson, D.C., Sager, K., Garst, E., Kang, M., Rubchinsky, K., & Dawson, K, 1997). The former theorists have shown that perceptions of similarity with a victim increase empathic responding, reasoning that the

more similar one is to a victim, the more they will be able or motivated to feel his pain, and the more likely they will be to help. Indeed, egoistic accounts of helping behavior argue the absence of any truly selfless motives claiming that helping behavior necessarily reflects upon the self. In response, supporters of the latter hypothesis (the “empathy-altruism” hypothesis) have shown that selfless empathic responding and prosocial behavior cannot be explained by perceptions of similarity, suggesting the possibility of truly altruistic motivations.

Unfortunately, it seems that any measure of perceived similarity has become inextricably linked with “egoistic” accounts of empathy. I believe this is premature and results primarily from methodological and theoretical difficulties in measuring and conceptualizing perceived similarity. It may be the case that low-level perceptions of similarity function much like appraisals, interacting with motivation and affect to elicit an emotional response with the goal of relieving another’s distress when similarity (whether defined by group membership or some other criterion) has been perceived. Contrary to egoistic accounts, this theory allows for the existence of truly selfless empathic responses, and contrary to empathy-altruism accounts, this theory allows for an important role of perceptions of similarity.

The present study tested manipulated similarity between the participant and a victim and measured the effect of this manipulation on fairness judgments, reported empathy, willingness to help, and time spent engaging in costly helping behavior. The addition of a measure of costly prosocial behavior disentangles “egoistic” accounts of empathy and helping from selfless accounts of empathy and helping. If perceptions of similarity do elicit selfless empathic responses with the goal of easing others’ distress, then we would expect that, given a choice between relieving one’s own distress by avoiding a long and difficult task and relieving another’s distress by helping with a long and difficult task, participants who

perceive some degree of similarity with a victim will be more likely to engage in costly helping behavior. It is predicted that participants will judge transgressions against similar victims to be less fair than transgressions against unknown victims; that they will report greater empathy for similar victims than unknown victims; that they will be more willing to engage in costly prosocial behavior on the behalf of similar victims; and that they will spend more time helping similar victims than unknown victims.

Method

49 individuals participated and were randomly assigned to one of two experimental conditions. The experimental conditions were variants of Conditions 3 and 4 from Study 1, the primary difference being that while Study 1 was designed to manipulate group membership with the enactor of the moral transgression, the present conditions were designed to manipulate group membership with the victim of the moral transgression.

Two confederates played the roles of other participants in each session. Participants were brought into the lab along with the two confederates and all were seated at individual work stations. The experimenter instructed one of the confederates to work on a separate task while the participant and remaining confederate completed a questionnaire requiring frequency estimates for different types of events (e.g. *How many black bears are there in Massachusetts?*). They were then “categorized” by the computer into one of two groups: overestimator or underestimator (cf. DeSteno, Dasgupta, Bartlett, & Caidric, 2004). Participants in the in-group condition were categorized as the same group as the confederate, and participants in the out-group condition were categorized as the opposite group of the confederate. Following the group manipulation, the experimenter entered the laboratory and escorted both confederates to a different laboratory, presumably to work on

the next part of the experiment. The experimenter then returned to the laboratory, and told the participant that their next task would be to observe and evaluate a new experimenter-blind assignment procedure that the confederate who was not a part of the group manipulation was about to complete. At that point, the experiment unfolded as in the “judging other” condition.

After the confederate playing the role of the transgressor had been “observed” unfairly assigning the two tasks, he left the rooms ostensibly to find the experimenter, and did not return. Shortly after, the confederate playing the role of the victim was brought back into the room and seated at the computer adjacent to the participant. The experimenter explained to the confederate that he had been randomly assigned to the “red” task which consisted of a series of math and logic problem solving tasks. This was done to increase the salience of the victim’s plight to the participant.

Participants then responded to questions regarding the assignment procedure which were presented as a way to collect opinions on the new protocol. Embedded in a small set of distractors was the target question for moral judgment: “How fairly did the other participant act?” Participants responded to this question on a 7-point scale ranging from “extremely unfairly” to “extremely fairly.” In addition, the following questions were included in order to measure felt empathy for the victim: “How much sympathy do you have for him/her”, “How much pity do you have for him/her?”, “How sorry do you feel for him/her?” Participants responded on a 7-point scale.

In order to measure willingness to help and time spent helping, participants were given the opportunity to help the victim with the long and tedious “red” task. The last screen participants saw said the following:

You have now completed your portion of the experiment. You can see the experimenter to receive your credit. As you know, one of the other participants has to complete a time consuming task. It is not important to the experimenter who completes this task- it is merely a quantity of material that needs to be completed. If you'd like to help please tell the experimenter as much when you get your credit. Thank you for participating.

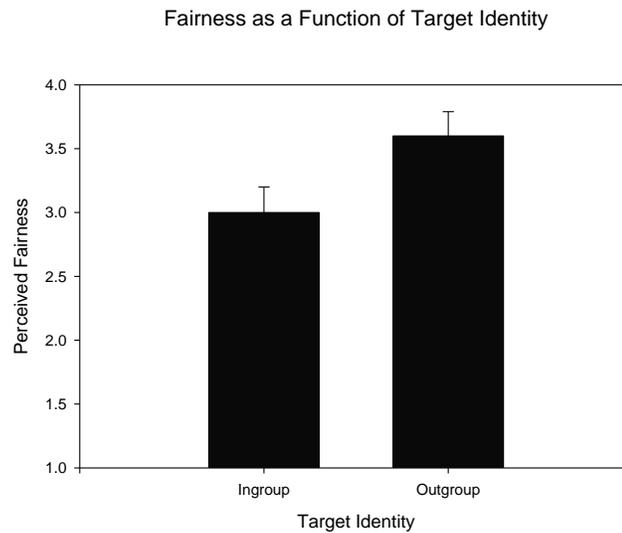
At this point the participant left the room and found the experimenter. In order to stifle any effects of authority on decision to help, the experimenter made no mention of the opportunity to help and simply asked if the participant had any questions about the experiment in general. The participant then either left with his/her credit, or offered to help the victim. In the latter case, the experimenter escorted the participant down the hall to a desk, sat him/her down and said the following:

This part of the experiment is a series of math problems, similar to SAT problems. You can do as much as you want. After you have finished, just leave all the materials on the desk and I will come back to get them when the other participant is done with his current task. Whatever you do not complete, the other participant will finish up at that point.

The experimenter then thanked the participant for his time, handed him the credit slip, walked back to the laboratory, and had no further contact with the participant. In order to ensure that any helping was due to a desire to help the victim, the participants were made aware that they had completed all that was necessary of them, that they had received their credit, and that they would not be in contact with anyone in the lab again. A hidden video camera surreptitiously recorded the participant as he completed the task while a research assistant timed how long he helped from the laboratory control room.

Results and Discussion

In accord with predictions, participants judged transgressions against in-group members as significantly more unfair than transgressions against out-group members, $t(47) = 2.17, p = .04$.

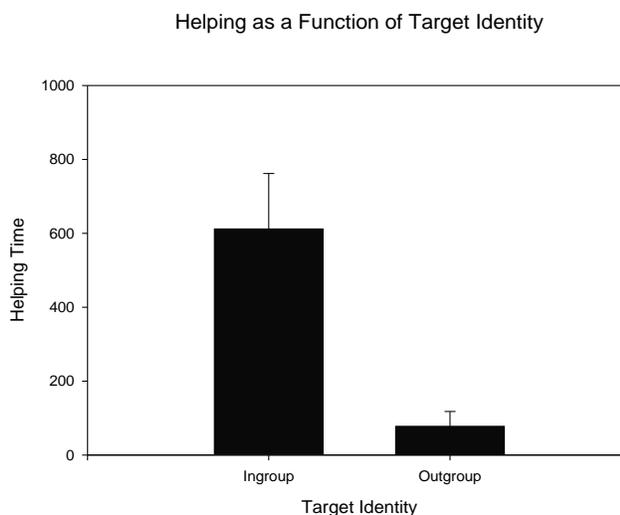


Participants were significantly more willing to offer help to in-group as opposed to out-group victims, Chi-Square (1, N=49) = 9.44, $p = .002$.

	Helped	No Help
Ingroup	14	10
Outgroup	4	21

$$\chi^2(1, N=49) = 9.44, p = .002$$

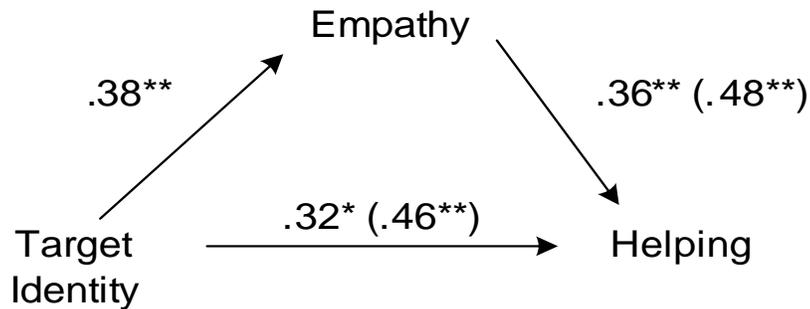
Participants helped in-group victims for a significantly longer period of time relative to out-group members, $t(47) = 3.51, p = .001$ (y-axis represents seconds spent helping).



Finally, participants reported more empathy for in-group victims ($M = 4.1806$) as opposed to out-group victims ($M = 3.52$), $t(47) = 2.84, p = .007$.

In order to demonstrate the causal power of empathy, time spent helping victims was regressed upon both level of empathy and condition. The figure below presents the standardized regression coefficients and zero-order correlations (in parentheses) for this path model. As expected, empathy remained a reliable predictor, but the effect of condition dropped significantly. Thus, it appears that part of the ability of the group manipulation to engender increased helping was mediated by the empathy it induced (Sobel = 1.96, $p = .05$). It should also be noted that there was a significant correlation between empathy and helping in both the in- and out-group conditions ($r = .508, p = .011$; $r = .416, p = .038$). However, as noted above, the empathy felt for the victim and, thus the level of helping, was greater when the victim was a member of the participant's group. Here again, one can readily see the

sensitivity of moral action to social alliances.



The present study provides strong evidence for the effect of group membership on, and consequently the role of similarity in, the experience of prosocial emotions and the frequency with which people engage in and devote effort to prosocial behavior. These findings are consistent with the idea that group membership signals an increased likelihood of reciprocity on the part of a victim. Being preferentially more empathic for similar others and displaying an increased willingness to engage in costly helping behavior on their behalf, aligns with theories of reciprocal altruism. The present study suggests one way in which reciprocal altruism develops and propagates within groups. Greater empathy is felt for those with whom we perceive similarity, and since group membership serves as a proxy for similarity, we tend to feel more for and help group members more often, thus increasing the likelihood of developing mutually beneficial relationships with group members.

General Discussion

Accepting the notion that group life was an important feature of our evolutionary

history opens the possibility that humans may have “a coevolved set of cultural practices and moral intuitions that are not about how to treat other individuals but about how to be a part of a group” (Haidt, 2007, p. 317). It may be the case that as a consequence of a history marked by life in groups (small and large) humans have evolved social-cognitive abilities and social emotions specifically tailored to the development of trusting and cooperative relationships with members of one’s ingroup. The three studies presented illustrate this fundamental role of moral emotions and moral reasoning processes in group-related issues. I found a fundamental bias in moral judgment (hypocrisy) that seems to operate to protect one’s reputation and the reputation of one’s group members (Valdesolo & DeSteno, 2007). I have unearthed a negativity to transgressions committed not only by others, but by oneself – suggesting that, even when individuals stand to gain rewards in the short-term they have intuitions which stymie this behavior (Valdesolo & DeSteno, 2008). Finally I have shown that people are more empathic towards group members and are more likely to engage in helping behavior on behalf of group members.

Several limitations of the studies are worth noting. While Study 1 found a significant difference in ratings of fairness transgressions across groups, we can only speculate as to the nature of such a bias. Whether lenience in judging group members’ transgressions results from a discrepancy in intuitive or more controlled processes is a question to address in future research. It could be the case, in accordance with the findings from Study 2, that participants exert the same degree of cognitive effort to rationalize and justify group members’ transgressions that they devote to justifying their own. Alternatively, it is possible that the leniency with which participants judge group members’ transgressions reflects an adherence to societal norms dictating the protection of one’s own. Group-level hypocrisy

might also result from a discrepancy in intuitive reactions to transgressions, with an automatic positivity bias extending to group members. Past research has found that learning regarding the moral status of others happens automatically and alters neural activity associated with social cognition (Singer, T., Kiebel, S.J., Winston, J.S., Dolan, R.J., & Frith, C.D., 2004). Specifically, it was found that reactions to individuals who had previously defected in a trust game elicited relatively less activity in areas of the brain associated with empathic responding. Given these findings, it may be the case that categorizing individuals as members of one's ingroup or outgroup could differentially influence the output of brain areas necessary for similar prosocial responses, creating the group-level hypocrisy effect. Whatever the process, determining the level at which hypocrisy operates for close others would be an important step in curbing this phenomenon.

Study 2 sought to reveal the process of hypocrisy. While I believe the proposed dual-process model best accounts for the pattern of results, there is an alternative which warrants mention. Upon first glance, it seems that simple actor-observer (A-O) dynamics can account for the pattern of results. A-O theory predicts that individuals will make more dispositional attributions when observing others' actions, while making more situational attributions for their own actions. With regards to Study 2, one could argue that in the neutral condition participants make more dispositional attributions for others' transgressions – driving judgments of “fairness” down – and, conversely, make more situational attributions for their own transgressions. The load manipulation prevents participants from engaging in the cognitive work necessary to attribute our unfair acts to situational as opposed to dispositional factors, thereby eliminating hypocrisy.

First, it is necessary to note that the present judgments are different in kind than

those normally found to produce the A-O effect, and therefore might not be explainable by traditional A-O theory. Questions regarding the fairness of an act do not equate with questions regarding the fairness or “morality” of an individual. Despite this, two separate conditions inducing negative affect at the time of judgment were run in order to rule out this potential confound.

Previous research has shown that inducing negative mood should reduce dispositional attributions (Forgas, 1998). Specifically, inducing a negative mood should increase cognitive processing allowing correction of an initial judgment for situational contingencies, or increased motivated reasoning, thereby strengthening the hypocrisy effect. I found that inducing negative mood does precisely the opposite; it eliminates hypocrisy by bringing judgments of self in line with judgments of others, $t(44) = .133$, $p = .895$. In accordance with this view, negative mood should not reduce a corrective bias.

While this research suggests that it is indeed the amplification of the prepotent response and not A-O effects that is pushing the evaluative judgment around, Forgas’ work deals with the effect of negative affect on *observers*. One could argue that it does not directly inform the present study since hypocrisy depends on the controlled processes of *actors*.

No study, to my knowledge, has tested the effect of induced negativity on actor’s attributions, but a recent meta-analysis of 113 A-O studies found that contemplating negative events (e.g. a moral transgression) heightened the A-O effect relative to when contemplating neutral events (Malle, 2006). Given that contemplating negative events is a standard affect-induction procedure, it is reasonable to predict that inducing negativity in a different way (in this case, watching negative video clips) should have the same effect on

attributions. However, it is also true that actors in our paradigm experienced both the negativity from contemplating their unfair actions, as well as the negativity experienced from the affect induction. It is possible that the combination of these two forms of affect inductions would have a unique effect on judgments, different than what they would each have individually. Rather than exacerbating the A-O effect, this combination of emotion inductions could be predicted to eliminate the effect, and thereby account for the pattern of results found. Future research might study the effect of different combinations of emotion on the attribution process to address this possibility, as the present study cannot rule this alternative out.

Future work will also aim to define how our “moral” intuitions can be fostered to exert a more consistent effect on behavior. The unearthing of a prepotent negative response to one’s own transgressions belies the frequency with which participants acted unfairly. Finding techniques to bridge the disconnect between feeling and action would be an important step in promoting prosocial behavior. Also this finding highlights the inconsistencies in arguments which equate evolved mechanisms with “selfish” motivations (Dawkins, 1989). “Selfish” genes need not produce selfish individuals. Indeed, it seems that the propagation of one’s genes would be best achieved by acting prosocially and cooperating with others when in environments characterized by stable groups and frequent communication. While success at the genetic level is the ultimate result, its proximate manifestation for individuals can be altruistic motivations. Importantly, these motivations seem to be directed primarily towards close others. While repeated hypocrisy on the part of an individual would surely lead to dislike and rejection, group-level hypocrisy strengthens social bonds by offering group members the reputational benefits that accompany more

lenient moral judgments.

With regards to Study 3, I argued for a crucial role of perceptions of similarity in empathic responding and prosocial behavior, though no measure of perceiving similarity with a victim was included. Rather, it was assumed that group membership stood as a proxy for similarity. Future studies will attempt to illustrate that the effect of perceived similarity on felt empathy and helping behavior is not restricted to the group level. One could argue that the effect of group membership on reported empathy and helping behavior was not due to low-level perceptions of similarity but rather a result of adherence to group- norms (i.e. one ought to protect one's own). To rule out this interpretation, empathy and helping behavior must be elicited by perceiving similarity in a manner detached from group membership. To this end, I plan to manipulate similarity via rhythmic synchrony. Put simply, I have designed a methodology wherein participants either keep the same beat with a confederate or keep a different beat. A participant and a confederate sit across from each other at a table and listen to an audio clip of rhythmic tones through individual headphones. They are asked to keep the beat by tapping a sensor on the table in front of them (i.e. tap the sensor once for each tone they hear). In the synchrony condition, the participant and the confederates listen to the same series of tones, whereas in the asynchrony condition they listen to different series of tones. I predict that rhythmic synchrony will create a heightened sense of similarity with another, relative to rhythmic asynchrony. Furthermore, I predict that relatively more empathic responding and helping will be elicited for those victims with whom participants have been synchronized, and that this will be mediated by an increase in perceptions of similarity. Lastly, participants will rate fairness transgressions committed against synchronized others more harshly than fairness transgressions committed against

asynchronized others.

Demonstrating this, in combination with the previous study, would provide a compelling case for the mediating role of perceived similarity in eliciting empathic responses and helping behavior. More generally, this line of research will provide new insight into the nature of our intuitions regarding fairness transgressions, committed against group members and strangers alike.

In sum, conceptualizing morality as a uniquely social phenomenon has important societal implications. Individuals need to be wary of the seemingly automatic biases that are activated when considering moral decisions and action. Tempering hypocrisy might lead to more open and productive discussions and negotiations amongst individuals with differing cultural values (e.g. Americans and Radical Islamists). Furthermore, acknowledging that intuitions regarding the sanctity of one's in-group are universal (Haidt, 2007) might be a first step in bridging political divides. For example, recognizing that patriotism and out-group suspicion seem to be grounded in important psychological mechanisms might give political liberals a better understanding of political conservatives and their motives. While valuing the in-group can clearly lead to undesirable consequences and beliefs (e.g. xenophobia), it is important to remember that it also elicits desirable prosocial emotions and behavior.

Indeed, morality seems to bind people together. My work has aimed to emphasize this point, representing what I hope is an important step in shifting research in moral psychology away from the emotion vs. reason framework which has defined it for so long, towards a social functional model.

REFERENCES

- Balcetis, E., & Dunning, D. (2006). See What You Want to See: Motivational Influences on Visual Perception. *Journal of Personality and Social Psychology, 91*(4), 612-625.
- Bandura, A., Ross, D., & Ross, S.A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal Psychology, 63*, 575-582.
- Bandura, A. (1990). Selective activation and disengagement of moral control. *Journal of Social Issues, 46*(1), 27-46.
- Bandura, A. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology, 71*(2), 364-374.
- Bargh, J., & Chartrand, T. (1999). The unbearable automaticity of being. *American Psychologist, 54*, 462-479.
- Bartlett, M.Y., & DeSteno, D.A. (2006). Gratitude and Prosocial Behavior: Helping When It Costs You. *Psychological Science, 17*(4), 319-325.
- Batson, C.D., Batson, J.G., Griffitt, C.A., Barrientos, S., Brandt, J.R., Sprengelmeyer, P., & Bayly, M.J. (1989). Negative-state relief and the empathy-altruism hypothesis. *Journal of Personality and Social Psychology, 56*(6), 922-933.
- Batson, C.D., Sager, K., Garst, E., Kang, M., Rubchinsky, K., & Dawson, K. (1997). Is empathy-induced helping due to self-other merging? *Journal of Personality and Social Psychology, 73*(3), 493-509.
- Batson, D.C., Kobrynowicz, D., & Dinnerstein, J. L. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology, 72*(6), 1335-1348.
- Batson, C.D., Thompson, E.R., Seufferling, G., Whitney, H., & Strongman, J. (1999). Moral Hypocrisy: Appearing moral to oneself without being so. *Journal of*

Personality and Social Psychology, 77, 525-537.

Brosnan, S.F., & de Waal, F.B.M. (2003). Monkeys reject unequal pay. *Nature*, 425, 297-299.

Cialdini, R.B., Schaller, M., Houlihan, D., Arps, K., Fultz, J., & Beaman, A.L. (1987). Empathy-based helping: Is it selflessly or selfishly motivated? *Journal of Personality and Social Psychology*, 52(4), 749-758.

Cushman, F.A., Young, L., & Hauser, M.D. (2006). The role of reasoning and intuition in moral judgments: Testing three principles of harm. *Psychological Science*, 17(12), 1082-1089.

Darley, J. (1993). Research on morality: Possible approaches, actual approaches. *Psychological Science*, 4, 353-357.

Dawkins, R. (1989). *The Selfish Gene*. Oxford, United Kingdom: Oxford University Press.

DeSteno, D.A., Dasgupta, N., Bartlett, M.Y., & Caidric, A. (2004). Prejudice from thin air: The effect of emotion on automatic intergroup attitudes. *Psychological Science*, 15, 319-324.

De Waal, F.B.M (2006). *Primates and Philosophers: How Morality Evolved*. Princeton: Princeton University Press.

Ehrlinger, J., & Dunning, D. (2003). *How chronic self-views influence (and potentially mislead) estimates of performance*, 84(1), 5-17.

Frank, R.H. (1988). *Passions within reason: The strategic role of the emotions*. New York: Norton.

Freud, S. (1976). *The interpretation of dreams* (J. Strachey, Trans.). New York: Norton. (Original work published 1900).

Gilbert, D. T., & Hixon, G.J. (1991). The trouble of thinking: Activation and

application of stereotypic beliefs. *Journal of Personality and Social Psychology*, 60(4), 509-517.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., & Cohen, J.D. (2001). An fMRI study of emotional engagement in moral judgment. *Science*, 293, 2105-2108.

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., & Cohen, J.D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44, 389-400.

Greenwald, A.G., & Farnham, S.D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology*, 79(6), 1022-1038.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814-834.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316, 998-1002.

Haidt, J., & Hersh, M. (2001) Sexual Morality: The cultures and reasons of liberals and conservatives. *Journal of Applied Social Psychology*, 31, 191-221.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, 65, 613-628.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*.

Hamilton, W.D. (1964). The genetical evolution of social behavior. In George C. Williams (Ed.), *Group Selection*. Chicago: Aldine Atherton, Inc.

Hauser, M.D. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York: Harper Collins.

Hume, D. (1969). *A treatise of human nature*. London: Penguin. (Original work published 1739-1740).

- Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*(5800), 829-832.
- Kohlberg, L. (1971). From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development. In T. Mischel (Ed.) *Cognitive development and epistemology* (p. 151 – 235). New York: Academic Press.
- Neuberg, S.L., Ciladini, R.B., Brown, S.L., Luce, C., Sagarin, B.J., & Lewis, B.P. (1997). Does empathy lead to anything more than superficial helping? Comment on Batson et al. (1997). *Journal of Personality and Social Psychology*, *73*(3), 510-516.
- Nucci, L., & Turiel, E. (1978). Social interaction and the development of social concepts in preschool children. *Child Development*, *49*, 400-407.
- Piaget, J. (1965). *The moral judgement of the child* (M. Gabain, Trans.). New York: Free Press. (Original work published 1932)
- Pizarro, D.A., & Bloom, P. (2003). The intelligence of moral intuitions: Comment on Haidt (2001). *Psychological Review*, *110*, 197-198.
- Singer, T., Kiebel, S.J., Winston, J.S., Dolan, R.J., & Frith, C.D. (2004). Brain response to the acquired moral status of faces. *Neuron*, *41*, 643-652.
- Tajfel, H. and Turner, J. C. (1986). The social identity theory of inter-group behavior. In S. Worchel and L. W. Austin (eds.), *Psychology of Intergroup Relations*. Chicago: Nelson-Hall
- Taylor, S.E., & Brown, J.D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193-210.
- Trivers, R.L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35-57.
- Trivers, R.L. (1985). *Social Evolution*. Menlo Park, CA: Benjamin/Cummings.
- Turiel, E. (1983). *The development of social knowledge: Morality and convention*.

Cambridge, England: Cambridge University Press.

Turiel, E., Killen, M., & Helwig, C.C. (1987). Morality: Its structure, function, and vagaries. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in young children* (p.155-243). Chicago: University of Chicago Press.

Valdesolo, P., & DeSteno, D.A. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*, 476-477.

Valdesolo, P., & DeSteno, D.A. (2007). Moral Hypocrisy: Social Groups and the Flexibility of Virtue. *Psychological Science*, *18*, 689-690.

Valdesolo, P., & DeSteno, D.A. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology* (in press).

Yamaguchi, S., Greenwald, A.G., Banaji, M.R., Murakami, F., Chen, D., Shiomura, K., Kobayashi, C., Cai, H., & Krendl, A. (in press). Apparent Universality of Positive Implicit Self-Esteem. *Psychological Science*.