Journal of Experimental Psychology: General
2013, Vol. 142, No. 4, 1001–1005

0096-3445/13/$12.00 DOI: 10.1037/a0030655

BRIEF REPORT

When Cheating Would Make You a Cheater: Implicating the Self Prevents Unethical Behavior

Christopher J. Bryan
University of California, San Diego

Gabrielle S. Adams
London Business School

Benoît Monin
Stanford University

In 3 experiments using 2 different paradigms, people were less likely to cheat for personal gain when a subtle change in phrasing framed such behavior as diagnostic of an undesirable identity. Participants were given the opportunity to claim money they were not entitled to at the experimenters' expense; instructions referred to cheating with either language that was designed to highlight the implications of cheating for the actor's identity (e.g., "Please don't be a cheater") or language that focused on the action (e.g., "Please don't cheat"). Participants in the "cheating" condition claimed significantly more money than did participants in the "cheater" condition, who showed no evidence of having cheated at all. This difference occurred both in a face-to-face interaction (Experiment 1) and in a private online setting (Experiments 2 and 3). These results demonstrate the power of a subtle linguistic difference to prevent even private unethical behavior by invoking people's desire to maintain a self-image as good and honest.

Keywords: moral identity, moral self-image, language and thought, labeling, dishonesty

Supplemental materials: http://dx.doi.org/10.1037/a0030655.supp

Think of a number from 1 to 10. Imagine that, before you reveal it, we tell you we are studying the prevalence of cheating and will give you $5 if your number is even. If you thought of an odd number (as most people do), would you tell us? Would you be more honest if, instead of the prevalence of "cheating," we told you we were studying the prevalence of "cheaters"? In this article, we propose that such subtle linguistic cues can influence ethical decisions by invoking identity concerns.

Specifically, we focus on the implications for ethical decision making of framing behavior as reflecting one's identity. A long tradition of research in moral psychology demonstrates that individuals motivated to engage in unethical behavior deploy strategies to weaken the behavior–identity link (e.g., Bandura, 1999; Mills, 1958). To reconcile their unethical behavior with their desire to see themselves as good and ethical (Blasi, 1980; Dunning, 2005; Monin & Jordan, 2009; Steele, 1988), people downplay the seriousness of their ethical lapses and tell themselves that occasional instances of cheating do not make one a dishonest person (Mazar, Amir, & Ariely, 2008). In doing so, an individual can engage in dishonest behavior while avoiding the correspondent inference (Jones & Nisbett, 1972; Ross, 1977) that he or she is the *kind of person* who behaves dishonestly, allowing that individual to have his or her cake (reap the benefits of unethical behavior) and eat it too (preserve a positive self-image).

Thus, one way to decrease the incidence of unethical behaviors might be to strengthen the link between such behaviors and their associated undesirable identities. In this article, we test whether highlighting the identity implications of cheating by using a subtle manipulation of phrasing can keep people honest. We refer to dishonest behavior with either the self-relevant noun *cheater* (e.g., "please don't be a cheater," "how common cheaters are") or some variation of the verb *to cheat* or the verbal noun *cheating* (e.g., "please don't cheat," "how common cheating is"). The "cheater" wording frames dishonest behavior as the enactment of an identity—a reflection of the kind of person one is—and should make it more difficult for people to ignore the implications of this behavior for their desired view of themselves as honest.

This wording manipulation is inspired by research showing that subtle differences in language can affect people's perceptions of themselves and others (Gelman & Heyman, 1999; Walton & Banaji, 2004). Nouns of the sort used in our manipulation directly characterize the actor, signaling that an attribute is representative

This article was published Online First November 5, 2012.

Christopher J. Bryan, Department of Psychology, University of California, San Diego; Gabrielle S. Adams, Organisational Behaviour, London Business School; Benoît Monin, Graduate School of Business and Department of Psychology, Stanford University.

We thank Elizabeth Mullen, Gregory Walton, and the Monin–Mullen Morality Lab for input and Hattie Bluestone and Ashli Carter for assistance.

Correspondence concerning this article should be addressed to Christopher J. Bryan, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109. E-mail: cbryan@ucsd.edu

of the person's essential identity (Gelman, Hollander, Star, & Heyman, 2000). Consistent with this, recent research found that exposure to a survey referring to voting in an upcoming election with a self-relevant noun (e.g., "How important is it to you to be a voter [vs. "to vote"] in tomorrow's election?") caused more people to vote the next day (Bryan, Walton, Rogers, & Dweck, 2011). Apparently the "voter" wording signaled to participants that, by voting, they could claim a desirable identity, which motivated them to vote. Thus, self-relevant nouns, like *cheater* and *voter,* ascribe symbolic significance to behavior, suggesting it has implications for the kind of person one would be by performing it.[1]

So far, self-relevant noun wording has only ever been shown to cause approach effects (e.g., motivating voting). This leaves open the possibility that the effect is caused not by the motivation to assume an identity but rather by a more purely cognitive self-perception process—that the "voter" wording, for example, caused participants simply to see themselves as voters and they behaved accordingly (Bem, 1972). Such an account would suggest that self-relevant noun wording should always increase people's tendency to act in line with the noun label. But our theory suggests the opposite prediction in the case of undesirable identities: self-relevant nouns should cause people to avoid the behavior.

Overview of Research and Theoretical Contributions

In three experiments, participants engaged in a task with real financial stakes in which they had the opportunity to claim money they were not entitled to and their individual cheating could not possibly be discovered. We manipulated the specific wording used to refer to cheating and predicted that participants would claim more money in the "cheating" than in the "cheater" condition.

Showing this would make important theoretical contributions in several areas of psychology. First, by showing that self-relevant noun wording not only increases the appeal of positive behavior (Bryan et al., 2011) but also decreases the appeal of negative behavior, it would provide support for our emerging theory that such nouns influence behavior by emphasizing its implications for identity. Second, it would provide direct empirical support for recent theoretical models asserting the importance of the self in regulating ethical behavior (Mazar et al., 2008; Monin & Jordan, 2009; Zhong, Lount, & Murnighan, 2010). Third, it would build on the rich tradition of research on causal attribution and correspondent inference (Jones & Nisbett, 1972; Ross, 1977) by showing that manipulating the availability of internal (or person) attributions for people's own actions—before they even happen—can affect their behavior. Finally, it would contribute to the growing literature examining ways in which small and seemingly incidental features of language have profound effects on the way we think and behave (Dils & Boroditsky, 2010; Fausey & Boroditsky, 2010; Thibodeau & Boroditsky, 2011).

Experiment 1

Method

Participants. Participants were approached on the campus of Stanford University by a student experimenter, who was unaware of the hypotheses, and asked if they would be willing to participate in a 3-min study for a chance to win $5. Fifty-one people agreed; however, participation was limited a priori to people who were native English speakers, which left a final sample of 50 (23 women; M_{age} = 22.98 years).

Procedure. Participants were randomly assigned to either the "cheater" or the "cheating" condition. The manipulation was embedded in the study instructions. The content of the instructions was identical in the two conditions; the only difference was the specific wording used to refer to cheating:

> We're interested in how common [cheating is/cheaters are] on college campuses. We're going to play a game in which we will be able to determine the approximate [rate of cheating/number of cheaters] in the group as a whole but it will be impossible for us to know whether you're [cheating/a cheater].

Next, participants were asked to think of a number from 1 to 10 without revealing the number to the experimenter. Once they had thought of the number, they were told they would receive $5 if their number was even but nothing if it was odd (Williams, Pizarro, & Ariely, 2009). Participants were then asked to reveal their number and paid (or not) as promised. We intentionally paid for even numbers because previous research has found that participants instructed to generate a random number typically show a strong bias toward odd numbers (Kubovy & Psotka, 1976), so we expected that this payment scheme would make more participants lose and face the temptation to cheat.

Results

As predicted, only a small proportion of participants in the "cheater" condition reported having thought of an even number (5 of 24, or 20.8%), whereas this proportion more than doubled (13 of 26, or 50.0%) in the "cheating" condition, $\chi^2(1) = 4.61$, $p = .032$, $w = .30$.

To confirm the previously documented bias toward odd numbers (Kubovy & Psotka, 1976), we approached 26 additional participants on Stanford's campus and simply asked them to think of a number from 1 to 10 with no promise of reward. Few (5 of 26, or 19.2%) thought of an even number, a rate nearly identical to that in the "cheater" condition, $p > .89$. Thus, it appears that many participants in the "cheating" condition misreported their number and collected money they did not deserve, but there is no evidence that anyone in the "cheater" condition did so.

Experiment 2

In Experiment 2, we sought to rule out the possibility that the effect observed in Experiment 1 relies on the presence of another person, which may have triggered self-presentation concerns. Experiment 2 was conducted in the more private and impersonal setting of an online study in which participants never met or expected to meet the experimenters.

Experiment 2 also used a new task (coin flipping) in which the expected outcome in the absence of cheating was more straight-

[1] That the voting effect was observed many hours after the wording manipulation suggests that the manipulation indeed changed the meaning of the behavior instead of, for example, increasing objective self-awareness, a more evanescent and situation-bound state (Duval & Wicklund, 1972).

forward. This allowed us to interpret more directly the difference between wording conditions relative to what would be expected by chance.

Finally, whereas Experiment 1 simply evoked the cheating-vs.-cheater framing in the absence of any direct admonition to the participant (e.g., "It will be impossible for us to know whether you are [cheating/a cheater]"), Experiment 2 tested the bolder prediction that, even when participants in both conditions were directly asked to be honest, "cheater" wording ("Please don't be a cheater") would curb dishonesty more effectively than "cheating" wording ("Please don't cheat").

Method

Participants. Participants were members of a university-administered online participant pool who volunteered for a study advertised as being about "psychokinesis." Eighty-eight people volunteered, but four did not meet the a priori criterion that they be native English speakers. Five additional people were excluded for having completed the experiment faster than pilot testing suggested was reasonable for a person participating in good faith (see the online supplemental material for details). Thus, the final sample included 79 participants (62 women; M_{age} = 39.87 years).

Procedure. Online instructions explained that a recent controversial article claimed to report the first scientific evidence for paranormal phenomena (a vague reference to an article by Bem, 2011, which had received considerable media attention). Participants were told they should find a coin and flip it 10 times, while trying to influence the outcome of each toss with their minds, making the coin land heads as often as possible. They were told that to ensure that they were "properly motivated," they would receive $1 for every toss landing heads. To forestall any perception of experimental demand to cheat, the instructions signaled that the present experimenters were skeptical that psychokinesis is real. Participants were randomly assigned to either the "cheater" or the "cheating" condition. The manipulation was embedded in the instructions that followed:

> NOTE: Please don't [cheat/be a cheater] and report that one or more of your coin flips landed heads when it really landed tails! Even a small [amount of cheating/number of cheaters] would undermine the study, making it appear that psychokinesis is real.

The instructions acknowledged that the laws of probability dictate that people would, on average, make $5, although some would "make as much as $10 just by chance" and others would "make as little as $0." The manipulation was also embedded in the instructions on the next page, where participants logged the outcomes of their 10 coin flips. At the top of the page, in large, red, capital letters, was the message, "PLEASE DON'T [CHEAT/BE A CHEATER]." We used the average number of heads participants claimed to have obtained to estimate cheating rates.

Results

As predicted, participants in the "cheating" condition claimed to have obtained significantly more heads (M = 5.49, SD = 1.25) than did those in the "cheater" condition (M = 4.88, SD = 1.38), $t(77) = 2.06$, $p = .043$, $d = 0.46$. Moreover, the mean number of heads reported in the "cheating" condition was significantly higher than the 5.00 that would be expected by chance, $t(38) = 2.43$, $p = .020$, $d = 0.39$, suggesting that cheating occurred. The average number of heads reported in the "cheater" condition was not different from chance, $t(39) = 0.570$, $p > .50$ (see Figure 1A).

Although we observed dishonesty in the "cheating" condition, the "cheater" wording apparently eliminated it completely. Furthermore, by replicating the essential finding from Experiment 1 in a relatively anonymous setting, Experiment 2 demonstrates that self-relevant noun wording reduces cheating even when it merely raises the private specter of taking on an undesired identity.

Experiment 3

Experiment 3 replicated the design of Experiment 2, adding a baseline condition with no reference to cheating. This allowed us to test whether the "cheating" condition had any effect relative to no message at all and to ascertain that our effect results from decreased cheating in the "cheater" condition and not from increased cheating in the "cheating" condition. To further rule out impression management, we also ensured that participants would feel anonymous and disconnected from the experimenters by using an ad hoc sample with no relationship with the university.

Method

Participants. Participants were users of Facebook in the United States who clicked on an advertisement for a "Stanford web study." Of 154 volunteers, 131 met the a priori criterion that they be native English speakers. Of those, 99 (54 women; M_{age} = 22.94 years) also met our completion time criterion for good-faith participation (see the online supplemental material for details).

Procedure. The procedure was identical to that in Experiment 2 except that a baseline condition was added in which cheating was not mentioned.

Results

The omnibus effect of condition was significant, $F(2, 96) = 4.38$, $p = .015$. Participants in the "cheating" condition claimed to

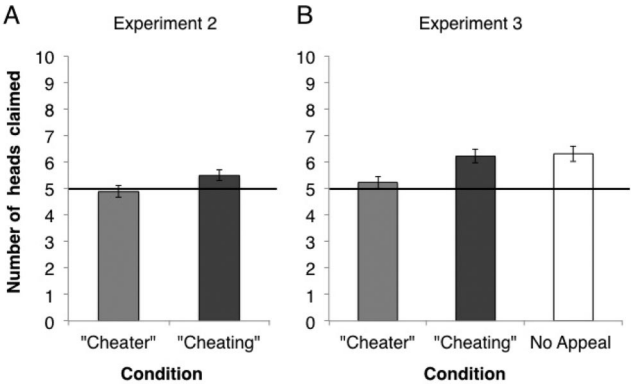

Figure 1. In Experiment 2 (Panel A), the mean number of heads claimed in the "cheating" condition was significantly greater than chance, suggesting cheating, and was significantly greater than the number claimed in the "cheater" condition. In Experiment 3 (Panel B), the mean numbers of heads claimed in both the "cheating" and baseline conditions were significantly greater than chance and greater than the number claimed in the "cheater" condition. Error bars indicate standard errors.

have obtained significantly more heads ($M = 6.22$, $SD = 1.55$) than did participants in the "cheater" condition ($M = 5.23$, $SD = 1.18$), $t(96) = 2.52$, $p = .013$, $d = 0.71$. Participants in the baseline condition also claimed to have obtained significantly more heads ($M = 6.31$, $SD = 1.72$) than did participants in the "cheater" condition, $t(96) = 2.95$, $p = .004$, $d = 0.66$. The numbers of heads claimed in the "cheating" and baseline conditions were similar, $t(96) = 0.25$, $p > .80$.

Further, the numbers of heads claimed in both the "cheating" and the baseline conditions were significantly higher than chance, $t(36) = 4.79$, $p < .0005$, $d = 0.79$, and $t(35) = 4.55$, $p < .0005$, $d = 0.78$, respectively. Finally, there was no evidence of cheating in the "cheater" condition; the number of heads claimed in that condition was not different from chance, $t(25) = 1.00$, $p > 0.30$ (see Figure 1B; see the online supplemental material for additional analyses).

Discussion

In three studies, we showed that simply using the self-relevant noun *cheater* rather than the verb (or verbal noun) *cheating* to refer to unethical behavior curbed cheating. In Experiment 1, participants in the "cheater" condition were half as likely to say they had thought of a winning number as were those in the "cheating" condition. In Experiments 2 and 3, participants in the "cheater" condition reported that their coin flipping resulted in chance rates of heads, whereas those in the "cheating" and baseline conditions reported above-chance earnings. These effects obtained in face-to-face interactions (Experiment 1) and private online settings (Experiments 2 and 3). In all three studies, the impossibility of identifying individual-level cheating ensures that participants were motivated by private concerns and not by worries about being caught or exposed as cheaters.

One intriguing finding is that direct appeals for honesty that used *cheating*-based wording were completely ineffective. In Experiment 3, participants cheated to the same degree in the "cheating" condition as they did in the baseline condition, where there was no appeal for honesty. But a simple shift to self-relevant noun wording appears to have eliminated cheating completely. This may be because, in this online context, the most salient rationale for honesty in the "cheating" condition was that someone the participant had never met and had no reason to care about was asking him or her not to cheat. But the *cheater*-based appeal changed the significance of cheating, suggesting it would say something about the participant's identity. It is fascinating to consider that institutions may unwittingly greatly moderate the effectiveness of such admonitions with arbitrary choices between seemingly equivalent phrasings (e.g., "Please don't litter" vs. "Please don't be a litterbug"; "Don't drink and drive" vs. "Don't be a drunk driver"). Awareness of the effect documented here holds the promise of increasing the effectiveness of appeals for prosocial behavior at little cost.

That the transgressions committed by participants in the "cheating" and baseline conditions were relatively minor does not diminish the importance of these findings. Indeed, Ariely (2012) argued that such minor transgressions, committed frequently and by large numbers of people, compose the lion's share of society's dishonesty. The fact that the *cheater* wording reduced cheating to the extent that none could be detected suggests the enormous potential of such subtle language manipulations to curb socially harmful behavior on a large scale.

Although the potential of self-relevant noun wording to reduce the incidence of unethical behavior in society is exciting, it is important to consider a possible risk our theory suggests might be associated with such techniques. Because such wording signals that cheating has implications for identity, it is unclear what the effect might be on someone who is exposed to this treatment and then goes on to cheat anyway. Such a person might come to see being a cheater as part of his or her identity (Miller, Brickman, & Bolen, 1975) and be more likely to cheat in the future.

In conclusion, these findings add to an emerging perspective suggesting that the self plays a central role in governing ethical behavior. Further, this effect demonstrates how even subtle linguistic cues can prevent dishonesty by harnessing people's desire to maintain a view of themselves as ethical and honest. This suggests the potential for simple interventions to help curb dishonest behavior in society.

References

Ariely, D. (2012). *The honest truth about dishonesty: How we lie to everyone—especially ourselves*. New York, NY: HarperCollins.

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

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). New York, NY: Academic Press.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 407–425. doi:10.1037/a0021524

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin, 88,* 1–45. doi:10.1037/0033-2909.88.1.1

Bryan, C. J., Walton, G. M., Rogers, T., & Dweck, C. S. (2011). Motivating voter turnout by invoking the self. *PNAS: Proceedings of the National Academy of Sciences, USA, 108,* 12653–12656. doi:10.1073/pnas.1103343108

Dils, A. T., & Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. *PNAS: Proceedings of the National Academy of Sciences, USA, 107,* 16396–16400. doi:10.1073/pnas.1009438107

Dunning, D. (2005). *Self-insight: Roadblocks and detours on the path to knowing thyself*. New York, NY: Psychology Press. doi:10.4324/9780203337998

Duval, S., & Wicklund, R. A. (1972). *A theory of objective self-awareness*. New York, NY: Academic Press.

Fausey, C. M., & Boroditsky, L. (2010). Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic Bulletin & Review, 17,* 644–650. doi:10.3758/PBR.17.5.644

Gelman, S. A., & Heyman, G. D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science, 10,* 489–493. doi:10.1111/1467-9280.00194

Gelman, S. A., Hollander, M., Star, J., & Heyman, G. D. (2000). The role of language in the construction of kinds. In D. Medin (Ed.), *Psychology of learning and motivation* (Vol. 39, pp. 201–263). New York, NY: Academic Press.

Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution:*

Perceiving the causes of behavior (pp. 79–94). Morristown, NJ: General Learning Press.

Kubovy, M., & Psotka, J. (1976). The predominance of seven and the apparent spontaneity of numerical choices. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 291–294. doi:10.1037/0096-1523.2.2.291

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45,* 633–644. doi:10.1509/jmkr.45.6.633

Miller, R. L., Brickman, P., & Bolen, D. (1975). Attribution versus persuasion as a means for modifying behavior. *Journal of Personality and Social Psychology, 31,* 430–441. doi:10.1037/h0076539

Mills, J. (1958). Changes in moral attitudes following temptation. *Journal of Personality, 26,* 517–531. doi:10.1111/j.1467-6494.1958.tb02349.x

Monin, B., & Jordan, A. H. (2009). The dynamic moral self: A social psychological perspective. In D. Narvaez & D. Lapsley (Eds.), *Personality, identity, and character: Explorations in moral psychology* (pp. 341–354). New York, NY: Cambridge University Press.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173–220). New York, NY: Academic Press.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261–302). New York, NY: Academic Press.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PLoS ONE, 6,* Article e16782. doi:10.1371/journal.pone.0016782

Walton, G. M., & Banaji, M. R. (2004). Being what you say: The effect of essentialist linguistic labels on preferences. *Social Cognition, 22,* 193–213. doi:10.1521/soco.22.2.193.35463

Williams, E. F., Pizarro, D., & Ariely, D. (2009, February). *Visceral states influence moral decision making*. Poster presented at the meeting of the Society for Personality and Social Psychology, Tampa, FL.

Zhong, C. B., Ku, G., Lount, R. B., & Murnighan, J. K. (2010). Compensatory ethics. *Journal of Business Ethics, 92,* 323–339. doi:10.1007/s10551-009-0161-6

Received April 23, 2012
Revision received September 20, 2012
Accepted September 22, 2012 ■