The Journal of Mind and Behavior
Winter and Spring 2015, Volume 36, Numbers 1 and 2
Pages 27–46
ISSN 0271–0137

Is That Me?
Sense of Agency as a Function of Intra-psychic Conflict

Travis A. Riddle

Columbia University

Howard J. Rosen

University of California, San Francisco

Ezequiel Morsella

San Francisco State University and *University of California, San Francisco*

The sense of agency is based on several cognitive processes, including the perception of a lawful correspondence between *action intentions* and *action outcomes*. We hypothesize that this sense is also modulated by intra-psychic conflict, such that urges (e.g., to smoke) conflicting with current goals (e.g., to not smoke) tend to be perceived as foreign to the self, as captured by the "monkey on one's back" metaphor describing aspects of addiction. Accordingly, in two classic *response interference* paradigms, participants perceived the activation of plans as less associated with the self when the plans conflicted with intended action than when the same plans led to no such interference. Intra-psychic conflict influenced the sense of agency in a dynamic and contextualized fashion. In both paradigms, response interference was associated with weakened perceptions of control and stronger perceptions of competition. These findings illuminate aspects of self-control, volition, and the cognitive construction of the self.

Keywords: sense of agency, authorship processing, cognitive conflict

Traditional psychophysics has focused on the correspondence between subjective experience and the stimuli of an external, objective world. Less attention has been given to the correspondence among modes of cognitive processing in the brain and their subjective effects. Intimately related to the notion of "Dasein" is

I am grateful to Raymond Russ, Editor, for his meticulous editing on an earlier draft of this manuscript.
Correspondence concerning this article should be addressed to Ezequiel Morsella, Ph.D., 1600 Holloway Avenue, EP 301, Department of Psychology, San Francisco State University, San Francisco, California 94132–4168. Email: morsella@sfsu.edu

the sense of existing as an entity/agent, both in the world and in the mind. This sense of agency is based on several cognitive processes, including the perception of a lawful correspondence between action intentions and action outcomes (Wegner, 2003). For example, if one intends to extend one's arm and then the arm happens to extend, one is likely to believe that the movement was willed by the self (Berti and Pia, 2006; Frith, Blakemore, and Wolpert, 2000; Pacherie, 2008; Prinz, 2003). Similarly, in the mental world, if one intends to imagine the shape of a triangle and then happens to experience triangle-like imagery, one is likely to believe that the imagery was willed by the self, even when, in actuality, the percept may have been caused by an experimental manipulation, as in the classic Perky effect (Perky, 1910). (In the Perky effect, experimental subjects are fooled into believing that they are imagining an image that is actually presented physically on a screen.) When intentions and outcomes mismatch, as in action slips and spoonerisms (Botvinick and Bylsma, 2005; Heckhausen and Beckmann, 1990), people are less likely to perceive actions as originating from the self (Wegner, 2002), leading to the cognition, "I didn't intend to do that" or "That thought/action came out of the blue," meaning "not from me."

It seems that similar self-versus-other attributions are found in motivational, intra-psychic conflicts as well (cf., Livnat and Pippenger, 2006), as captured by the "monkey on one's back" metaphor that is often used to describe the tendencies associated with aspects of addiction. Most exemplary, in Freud's (1938) classic framework of the *id*, *ego*, and *superego*, primitive animalistic urges (e.g., libidinal urges from the id) stem from something that is perceived to be distinct from the self (i.e., distinct from the ego). Apart from these speculations and theoretical approaches, there is little empirical evidence to illuminate the relationship between the sense of agency and intra-psychic conflict.

To this end, we hypothesize that intra-psychic conflict influences the sense of agency, such that urges conflicting with goals tend to be perceived as foreign to the self. More specifically, building in part on findings in the addiction literature (Baker, Piper, McCarthy, Majeskie, and Fiore, 2004; Loewenstein, 1996), we hypothesize that the urge to engage in an activity (e.g., smoking) that is incompatible with intentions (e.g., to be healthy) should lead to conscious conflict and such self-versus-not-self attributions (Pacherie, 2008). These intra-psychic, conscious conflicts (Morsella, 2005) stem from incompatible skeletomotor intentions, such as when one suppresses dropping a hot dish of food, suppresses eating behaviors (as in dieting), suppresses emotions, or holds one's breath while underwater (Morsella, Krieger, and Bargh, 2009). To the benefit of the experimenter, research has revealed that such conscious conflicts can also stem from less "hot" (Metcalfe and Mischel, 1999) conflicts, such as those elicited by laboratory response-interference paradigms (Morsella, Gray, Krieger, and Bargh, 2009).

In accord with recent views (Gazzaley and D'Esposito, 2007; van Veen and Carter, 2006), we believe that unraveling the nature of the subjective effects associated

with response conflict in interference tasks such as the Stroop paradigm is essential for understanding the dynamics of the "hot" response conflicts (Metcalfe and Mischel, 1999) of everyday life, such as those involving self-control and motivational conflict (Baker et al., 2004; see review in Morsella, Berger, and Krieger, 2011).

In the classic Stroop interference paradigm (Stroop, 1935), for example, participants are instructed to name the colors in which words are written. When the word and color are incongruous (e.g., RED presented in blue), response conflict leads to increased error rates, response times (RTs), and self-reported "urges to make a mistake" (henceforth, "urges to err"; Morsella, Gray et al., 2009). Moderate interference is also found with non-color, control words (e.g., HOUSE presented in blue). When the color and word are congruous (e.g., RED presented in red), or the color is presented on a neutral stimulus (e.g., a series of *x*'s as in "XXXX"), there is little or no interference (see review in MacLeod and MacDonald, 2000) or urges to err (Morsella, Gray et al., 2009).

Only recently have researchers began to look at the subjective and metacognitive aspects (e.g., urges to err) of interference tasks (Morsella, Wilson et al., 2009). Data suggest that these trial-by-trial subjective effects are not due to participants observing their own behavioral RTs. For example, these subjective effects are still robust in a Stroop-like interference paradigm (Morsella, Wilson et al., 2009) in which participants are instructed to withhold responding to the target stimulus for over a second, which eradicates RT effects (Eriksen and Schultz, 1979). Moreover, the effects are present when participants sustain incompatible intentions (e.g., to point left *and* right) in a motionless state in which no response is required or emitted (Morsella, Gray et al., 2009). In addition, though post-error corrections in interference paradigms involve improved performance (e.g., faster RTs) on trials following a trial involving response interference (e.g., an incongruent trial), reported urges to err actually increase in such a trial, which has been explained as a difference between implicit measures of performance (e.g., RT) and explicit measures (e.g., self-reports about task difficulty; Etkin, Prater, Hoeft, Menon, and Schatzberg, 2010). This research suggests that the explicit (conscious) system can be affected in the opposite manner of the implicit, unconscious, behavioral system, and that urge ratings are not based on observations of one's RTs.

Nevertheless, and despite research showing that urges arise even when RT effects are eradicated by having participants delay responses (Morsella, Wilson et al., 2009), at this stage of understanding, it is difficult, if not impossible, to remove all influence of processing speed, processing fluency, or a general sense of effort (or a combination thereof) on the judgments made by participants (Winkielman, Schwarz, Fazendeiro, and Reber, 2003). Processing speed is introspectable even when behavioral responses are suppressed. More generally, as with other introspective measures, it is challenging to verify what participants are introspecting at the moment that they are making their judgment. Self-reports are far from infallible, even if they occur just seconds after the relevant conscious experience (Block, 2007).

In examining the trial-by-trial subjective aspects of participants' responses in interference paradigms, Morsella, Wilson et al. (2009) concluded that, when there is response conflict (e.g., the Stroop incongruent condition), urges to err tend to be strong; when response interference is low or absent (e.g., the congruent condition), self-reported urges to err tend to be weak. That urges to err are weak for the congruent condition of the Stroop task is interesting because it is known from behavioral and psychophysiological data that participants often do read the stimulus word inadvertently in this condition: "The experimenter (perhaps the participant as well) cannot discriminate which dimension gave rise to the response on a given congruent trial" (see review in MacLeod and MacDonald, 2000, p. 386). Urges to err for the congruent condition are comparable to those of the neutral condition of the Stroop task, in which the color is presented on an illegible letter string (Morsella, Wilson et al., 2009). This intriguing finding has been explained as an instance of *synchrony-blindness*, in which one is unaware that two distinct cognitive operations are activated when the operations lead to the same action plan (Molapour, Berger, and Morsella, 2011). The notion of synchrony-blindness is consistent with the more general view that one is conscious only of the outputs of processes, not of the processes themselves (Fodor, 1983; Lashley, 1951). Synchrony-blindness seems to be featured also in the congruent conditions of countermanding tasks such as the anti-saccade task (Curtis and D'Esposito, 2009).

As this initial research suggests, notable changes in consciousness accompany responses in interference paradigms, rendering the responses in these paradigms qualitatively different from everyday actions (e.g., flicking a switch). Yet, as mentioned above, less has been revealed and documented about the subjective, agency-related aspects of these tasks than about their behavioral and neural aspects. Stemming from research unrelated to the phenomena at hand, one framework (Morsella, 2005) proposes that, of the many forms of integration or binding in the brain, the kinds of subjective and metacognitive effects associated with the sense of agency are most intimately related to one form of binding, namely, efference–efference binding. From this standpoint (Morsella and Bargh, 2011), consciousness and other high-level metacognitive processes play a smaller role in the binding of perceptual features within or between modalities. This afference binding can occur unconsciously, as in perceptual feature binding (e.g., the binding of object shape to color) and intersensory illusions (e.g., the ventriloquism effect). As well, the binding between perceptual and action codes (efference binding; Haggard, Aschersleben, Gehrke, and Prinz, 2002) can occur unconsciously, as when a subliminal stimulus elicits a button press or when one reflexively withdraws one's hand from a painful stimulus or when one reflexively inhales. Thus, consciousness, the sense of agency, and other, high-level metacognitive components (e.g., sense of competition) are most intimately-related to efference-efference binding (defined below).

Efference–efference binding occurs when two streams of efference binding are trying to influence skeletomotor action at the same time. It is important to note that the conscious conflicts associated with this form of binding are intimately related to action selection in the skeletal muscle output system (Morsella, 2005). Conflicts involving non-skeletal muscle effectors (e.g., smooth muscle conflict) do not lead to any subjective effects (Morsella, Gray et al. 2009), and conflicts occurring before the action selection stage of processing (e.g., intersensory conflicts) do not lead to any kind of subjective strife. The pattern of observations is consistent with the view that consciousness integrates high-level outputs for a form of action control, one operating at a higher level than motor control, which is largely unconscious (Morsella and Bargh, 2011).

It has been proposed that, because it is required for integrating two conflicting streams of efference binding, efference–efference binding results in integrated actions such as holding one's breath, carrying a hot dish of food, performing the Stroop task, suppressing socially-inappropriate behavior, or thus modulating another action plan (Morsella and Bargh, 2011). Yet, to date, there is no evidence that a conflicting efference stream is perceived as a "monkey on one's back" and perceived as foreign to the self.

In interference tasks, are self-versus-not-self attributions ephemeral and nebulous, or systematic and reliable? In light of these questions, our goal was to demonstrate for the first time that urges conflicting with one's intended action goals (an instance of efference–efference binding) tend to be perceived as foreign to the self. To this end, in a series of studies, we had participants introspect self-relevant aspects of subjective experience (*perceptions of action authorship* [Wegner, 2003], *control, and competition*) on a trial-by-trial basis while performing classic response interference paradigms.

Study 1

Hypothesis and prediction. We hypothesized that, when an action plan is activated and counters one's action goal, that action plan is perceived as less due to the self than when the concurrently activated action plan does not interfere with one's action goal. In the control and incongruent conditions of the Stroop task, word reading leads to an action plan that counters the participants' goal of naming the color (henceforth, "color-naming"). Hence, we predicted that, in these two conditions of the classic Stroop task, urges to read are less attributed to the self than in the congruent condition, when word-reading does not interfere with, and may actually facilitate, performance on the task (MacLeod and MacDonald, 2000). In our paradigm, participants were asked after each Stroop trial, "How strongly do you feel that the urge to read the word was due to your 'self'?" For brevity, we refer to this as our "reading due to self" dependent measure. We chose the Stroop task because it innocuously captures aspects of the "monkey on one's back" phenomenon, and much is already known about its cognitive, subjective, and neural components.

Extensive piloting revealed that, when presenting this question alone (Pilot Study 1, $n = 17$), and when not including additional clarifications about what was meant in the question by the term "self" (Pilot Study 2, $n = 8$), different participants tended to interpret this question about the self to mean quite different things. For example, piloting revealed that participants often construed "self" as meaning the physical body or organism. As one would expect, this misinterpretation of our question did not lead to informative effects about our experimental manipulation: the Stroop condition did not influence attributions of word reading to the self ($ps > .10$). Similar, ambiguous effects were obtained in Pilot Study 3 ($n = 18$), in which the following question was presented alone: "How strongly do you feel that the urge to read the word was due to your 'self'?" As explained below, for Pilot Studies 1 and 3, after each trial, participants were also asked the following two questions in the following order: "How much personal control did you feel when responding?" (on a 1-to-8 scale in which 1 signified "no control" and 8 signified "absolute control") and "How strong was the thought of a competing response?" (on a 1-to-8 scale in which 1 signified "not strong at all" and 8 signified "very strong").

From piloting we learned that, to remedy these shortcomings, participants need to be presented with a statement (presented below) that explains the difference between the physical self and psychological self. In addition, we learned that, to be understood in the intended manner, the "reading due to self" question could not be presented alone and benefited from being presented along with the question above about color-naming and the self (i.e., "color-naming due to self" question). As evident in previous studies (Morsella, Wilson et al., 2009), questions about subjective experience are answered differently in different contexts and introspecting about one subjective dimension of interest influences judgments based on other dimensions. From this extensive piloting, we became confident that, by clarifying what we mean by "self" and by presenting a comparison question about color-naming, participants would interpret our critical question as we intended.

Method

Participants. San Francisco State University undergraduate students ($n = 32$) participated for class credit. These students were enrolled in psychology courses. The involvement of human participants in our project was approved by the Institutional Review Board at San Francisco State University.

Procedure. Participants were run individually. The session consisted of a block of trials in which participants responded to Stroop stimuli vocally. Each block consisted of 24 Stroop trials having eight congruent (e.g., RED written in red), eight incongruent (e.g., RED in blue), and eight control (e.g., HOUSE in green) stimuli presented in random order. No neutral stimuli (e.g., XXXX in pink) were presented because our "reading due to self" question could not be asked about such stimuli. The eight colors used were correctly identified by all participants. Participants

were instructed, "In this task, you must respond to the words presented on the screen by naming aloud the colors in which the words are written as fast and as accurately as possible. For example, if the word FLOWER is presented in blue, you must utter the color name 'blue.' The microphone will record your response and measure your response time." Vocal responses were detected by microphone (Model 33-3014; Radio Shack; Fort Worth, TX) connected to a PsyScope button box (Response Box; ioLab Systems; UK). Piloting revealed that, for participants to understand that our question was not about the physical self, the experimenter had to explain the nature of the psychological self. Hence, in our experiment, participants were presented with the following statement about the "self."

> There are things that occur in the mind which feel like they come from one's psychological self, and things that feel like they do not come so much from one's psychological self. In psychology, researchers often differentiate between "the bodily self" and "the psychological self." In this study, we are examining the nature of the psychological self.

For this and the following experiments, stimuli were always presented in random order on a white background of a 43 cm Apple iMac computer monitor with a viewing distance of approximately 50.8 cm, and stimulus presentation was controlled by PsyScope software (Cohen, MacWhinney, Flatt, and Provost, 1993). A sample trial proceeded as follows. A blank screen was shown for 700 ms. It was followed by a randomly selected Stroop stimulus (48-point Helvetica), remaining onscreen until a vocal response was detected by microphone. After the response, participants were asked via computer screen, "How strongly do you feel that the urge to read the word was due to your 'self'?," which they rated on an eight-point scale, in which 1 signified "not at all due to self" and 8 signified "absolutely due to self." After inputting their rating and pressing the return key, participants were asked, "How strongly do you feel that the urge to name the color was due to your 'self'?," which they rated using the scale for the first question. This input terminated the trial. The order of presentation of the two questions was counterbalanced across participants.

Results

Primary results. As illustrated in Figure 1, the Stroop condition produced the predicted systematic effects on the measure "reading due to self," F (2, 62) = 10.856, $p < .0001$ ($\eta_p^2 = .26$), in which these attributions were lowest for the incongruent condition ($M = 5.51$, $SEM = .30$), followed by the control ($M = 5.61$, $SEM = .28$) and congruent conditions ($M = 6.33$, $SEM = .30$). Planned comparisons revealed that all differences between conditions were significant ($ps < .01$), except for that between incongruent and control conditions ($p = .53$). Omitted responses and typing errors resulted in the loss of eight (1.0%) of 768 "reading due to self" ratings.

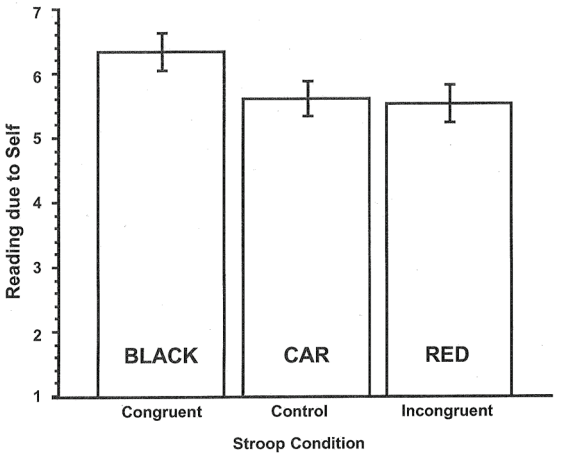

Figure 1: Mean attributions of word-reading urges being due to "the self" as a function of Stroop condition. Error bars indicate *SEM*s.

Reaction time analysis. As in Morsella, Gray et al. (2009), Stroop-task RTs below 200 ms and above 2.5 s were excluded from analysis, resulting in the data loss of 82 (10.6%) out of 768 trials. We replicated the Stroop RT effect: ANOVA analyses revealed that the Stroop condition had systematic effects on RTs, $F(2, 62) = 41.171$, $p < .0001$. Mean RTs were greatest for the incongruent ($M = 1289.96$, $SEM = 51.47$), followed by control ($M = 1166.67$, $SEM = 55.72$) and congruent conditions ($M = 1019.86$, $SEM = 48.28$). Planned comparisons revealed that all differences between conditions were significant ($ps < .01$).

Correlational analysis. Only five of the 32 participants had significant within-person correlations ($rs > .4$ or $< -.4$, $ps < .05$) between RT and "reading due to self" ratings. We used Fisher zr to estimate the population correlation between RT and the ratings (based on 24 trials), and it was nonsignificant ($r = -.05$, $p > .05$). These findings suggest that participants may not have based their "reading due to self" judgment on observing their own RTs. Regarding the "color-naming due to self" ratings, ten of the 32 participants had significant within-person correlations ($rs \leq -.42$, $ps < .05$) between the rating and RT. We used Fisher zr to estimate the population correlation between RT and the ratings (based on 24 trials), and it was nonsignificant ($r = -.25$, $p > .05$).

Supplementary analysis. The "color-naming due to self" question led to an unanticipated and intriguing pattern of results, mirroring that of the "reading due to self ratings," in which these attributions were lowest for the incongruent condition ($M = 5.15$, $SEM = .30$), followed by the control ($M = 5.56$, $SEM = .28$) and congruent conditions ($M = 6.36$, $SEM = .31$), $F(2, 62) = 14.498$, $p < .0001$ ($\eta_p^2 = .33$). *Fisher's PLSD* revealed that only the contrast between the congruent and incongruent conditions was significant, $p < .05$. (Each contrast is significant when

analyzing the data using the same, planned analysis that was used for the "reading due to self" ratings, *p*s < .05.) Omitted responses and typing errors resulted in the loss of 39 (5.1%) of 768 "reading due to self" ratings.

Discussion. As predicted, urges to read were less attributed to the self in the incongruent and control conditions of the Stroop task than in the congruent condition, the only condition in which the automatic action plan of reading does not interfere with performance. At this stage of understanding, it remains unclear why the same pattern of judgments was found for the "color-naming due to self" question. Warranting further investigation and beyond the purview of the present project, which focuses on the strong, automatic actions associated with the task (i.e., the automatic word-reading plan), this finding may reveal additional information about the ways in which participants introspect about, and conceptualize, the process of color-naming, the non-dominant, target action plan (see General Discussion).

Less interestingly, perhaps participants were simply re-inputting the rating that they had inputted for the first question that happened to be presented, or they adopted a strategy in which, when confronted with the incongruent condition, lower ratings were always inputted for the incongruent condition, regardless of the question at hand. To evaluate this uninteresting hypothesis and also learn more about the kinds of agency-related attributions that participants are making as a function of Stroop condition, we re-analyzed the data from the pilot studies (Pilot Studies 1 and 3, *n* = 35) that included two questions about the sense of agency ("How much personal control did you feel when responding?" and "How strong was the thought of a competing response?"), questions that should lead to an opposite patterns of results.

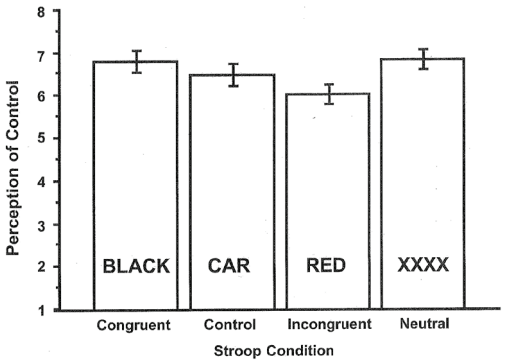

Figure 2: Mean perceptions of control in a vocal version of the Stroop task. Error bars indicate *SEM*s.

Regarding "perception of control," as revealed in Figure 2, Stroop condition led to systematic effects, $F(3, 102) = 18.034$, $p < .0001$ ($\eta_p^2 = .35$), in which perception of control was greatest for the neutral ($M = 6.81$, $SEM = .25$) and congruent ($M = 6.80$,

SEM = .25), followed by control (*M* = 6.48, *SEM* = .25) and incongruent conditions (*M* = 6.00, *SEM* = .24). Planned comparisons revealed that all the differences between conditions were significant (*p*s < .05), except for the differences between congruent and control conditions (*p* = .05), and neutral and congruent conditions (*p* = .94).

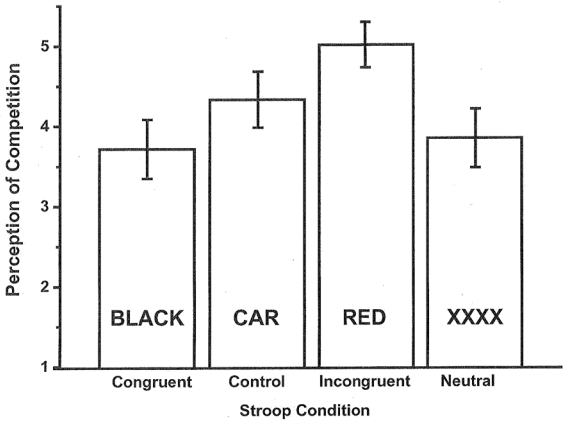

Figure 3: Mean perceptions of competition in a vocal version of the Stroop task. Error bars indicate *SEM*s.

As illustrated in Figure 3, the opposite pattern of results was obtained from the "perception of competition" question, F (3, 102) = 18.648, $p < .0001$ ($\eta_p^2 = .35$), in which perception of competition was greatest for the incongruent condition (*M* = 5.02, *SEM* = .28), followed by the control (*M* = 4.34, *SEM* = .34), neutral (*M* = 3.85, *SEM* = .37) and congruent conditions (*M* = 3.72, *SEM* = .36). Planned comparisons revealed that all the differences between conditions were significant (*p*s < .05), except for the difference between congruent and neutral conditions (*p* = .38). Together, these data replicate in a different context the findings regarding perceptions of control and competition found in Morsella, Wilson et al. (2009). More importantly for present purposes, these data suggest that our primary finding regarding "reading due to self" occurred within a task in which introspections about subjective dimensions associated with agency (e.g., perceptions of control and competition) can be made accurately and reliably. In short, it seems that the judgments obtained in Study 1 co-occur with subjective experiences that one would associate with the sense of agency. In addition, these additional data cast doubt on the alternative hypothesis that our primary finding stemmed, not from participants accurately introspecting about how strongly they felt the urge to read to be attributed to their psychological self, but from participants always reporting lower scores for the incongruent condition, regardless of the question at hand. Data regarding perceptions of control and competition reveal that participants

are capable of responding differentially to questions about subjective aspects of responding that are inversely-related, and suggest that, at a minimum, the unexpected "color naming due to self" finding requires further exploration. Regarding our primary "reading due to self" effect, a more convincing argument about the sense of agency and cognitive interference would be made if it were replicated in a different kind of interference paradigm.

Study 2

In addition to attempting to replicate the primary finding of Study 1 in a different paradigm, in Study 2 we also examined the hypothesis that these self-versus-not-self attributions are malleable and context-dependent: a plan that is intended in one context may be perceived as foreign to the self in another context (e.g., when it is incompatible with current goals). One limitation of using the Stroop task to examine this additional hypothesis is that the interference elicited by the task involves plans that are qualitatively distinct (object-naming versus word-reading) and that possess different "strengths": the color-naming plan is weaker than the automatic, word reading plan (Cohen, Dunbar, and McClelland, 1990; MacLeod and MacDonald, 2000). Hence, to examine our hypothesis, we used a Stroop-like task without these limitations. In the MacLeod and Dunbar (1988) task, participants are trained to name nonsense shapes using color names. For instance, the participant is instructed to name a six-sided polygon as "orange." Following training, participants are instructed to name the colors in which the shapes happen to be presented. On congruent trials, the shape name and color are congruent (e.g., the shape "orange" is presented in orange). On incongruent trials, the shape name and color name are different. For example, the same six-sided polygon will appear in blue and the participant must respond "blue," leading to interference (e.g., increased RTs). In a second phase, participants are instructed to name the shapes and disregard the colors in which the shapes are presented. In the incongruent condition, newly acquired shape-naming plans interfere with color-naming plans (MacLeod and MacDonald, 2000).

Unlike the Stroop paradigm, which examines interference from undesired word-reading plans (Cohen et al., 1990), in this paradigm one can measure within a single session the subjective interference effects of each stimulus-related plan, because the plan that is task-irrelevant in one phase (e.g., shape naming) of the session is task-relevant in the other, and vice versa. Moreover, the paradigm is purer than the Stroop in that intended and interfering plans involve the same kind of action (naming). Together, these advantages allow one to draw better conclusions (cf., MacLeod and MacDonald, 2000).

Hypothesis and prediction. We predicted that, during the shape-naming phase, participants would perceive the activation of color-naming plans as less associated with the self in incongruent than in congruent conditions. Our second prediction

was that, in the color-naming phase, participants would perceive the activation of shape-naming plans as less associated with the self in incongruent than in congruent conditions. Last, we predicted that perceptions of control would be greater for congruent than incongruent conditions and that perceptions of competition would be greater for incongruent than congruent conditions.

Method

Participants. San Francisco State University undergraduates (n = 85) participated for class credit. As with Study 1, these students were enrolled in psychology courses. The involvement of human participants in our project was approved by the Institutional Review Board at San Francisco State University.

Procedure. Procedures followed those of MacLeod and Dunbar (Experiment 1; 1988). Stimuli were presented by computer screen in the same manner as in Study 1. After assessing that participants could identify the colors blue, green, orange, and pink, the session began with a shape-familiarization phase in which participants learned to name shapes by the designations "blue," "green," "orange," and "pink" (Figure 4). Each shape appeared with its corresponding name twice. Thereafter, participants performed a training session in which they had to name each shape aloud. As in Study 1, vocal responses were detected by microphone.

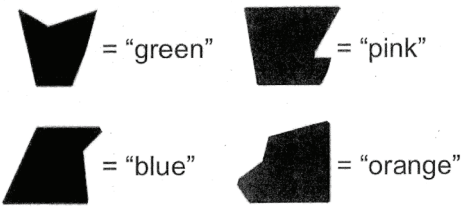

Figure 4: Shape names for the visual stimuli based on MacLeod and Dunbar (1988).

MacLeod and Dunbar (1988) provided participants (n = 22) with 16, 192, 288, or 576 trials of shape-naming training. To examine the sense of agency, it was unnecessary and unfeasible to administer four different degrees of training. Because having participants perform just one of the four training regimens might yield subjective effects that are unique to that regimen, and because administering only a representative regimen (e.g., the average of 268 trials) would not permit us to compare our RT data with those of MacLeod and Dunbar (1988), we decided to administer the two most extreme training regimens (16 trials [participants = 45] or 576 trials [participants = 40]) to get a representative sample of the variable degree of training. Varying the *degree of training* also allowed us to examine whether

introspections could be influenced by degree of training or by its effects on RT. Again, it is difficult to eradicate the influence of processing fluency or speed on subjective judgments (Winkielman et al., 2003): introspective judgments could be based, not on conflict, but on the observation of the speed of the overt response or internal processing. Thus, the two regimens also allowed us to explore further the relationship between RT and judgments.

Following training, participants performed the *color-naming phase*, in which they had to name as quickly as possible the colors in which the shapes were presented. Of the 72 trials, 24 trials were incongruent, 24 were congruent, and 24 were control (a square patch of blue). In the incongruent condition, each color was presented on each shape twice. In the control condition, each patch of color appeared six times. In the congruent condition, each color appeared with the congruent shape six times. Following each trial, participants were asked three questions in the following order: "How much personal control did you feel when responding?," "How strong was the thought of a competing response?," and "How strongly do you feel that the urge to shape-name was due to your 'self'?" Participants rated introspections on a 1–8 scale, in which 1 signified "no control" for the first question, "not strong at all" for the second question, and "not at all due to self," for the third question, and 8 signified "absolute control," "very strong," and "absolutely due to self," respectively. The subsequent, *shape-naming phase* was identical except that participants named aloud the name of the shape, and control stimuli were the shapes presented in black. For both phases, control stimuli were included only because we wanted to follow the procedures of MacLeod and Dunbar (1988) as closely as possible. It is difficult to appreciate whether control stimuli are informative regarding our subjective measures. For this phase, the third question read, "How strongly do you feel that the urge to name the color was due to your 'self'?"

Results

The data from one participant were excluded from analysis because the participant did not follow instructions. We collapsed the introspective data from both degrees of training, because both regimens provided similar results. Because our primary focus was the shape-naming phase (where interference is presumably strongest), we present those results first.

Shape-Naming Phase

Typing errors resulted in the loss of 275 (1.5%) of 18,144 ratings. Participants perceived the urge to color-name as less due to the self during the incongruent than congruent conditions, F (2, 166) = 3.843, $p < .05$ ($\eta_p^2 = 0.04$). [Table 1]. Planned comparisons revealed that all means are significantly different from

Table 1
Mean Introspective Report as a Function of Task and Condition

Shape-Naming Task	Congruent	Incongruent	Control
Color-naming plan due to self	5.54 (.23)	5.31 (.20)	5.31 (.23)
Personal control	6.75 (.20)	6.06 (.17)	6.57 (.17)
Perceptions of competition	2.43 (.19)	3.87 (.18)	2.99 (.20)
Response times	1176.07 (52.61)	1360.94 (55.54)	1264.75 (53.20)
Color-Naming Task	**Congruent**	**Incongruent**	**Control**
Shape-naming plan due to self	5.66 (.25)	5.47 (.24)	5.81 (.26)
Personal control	7.13 (.17)	6.84 (.16)	7.28 (.17)
Perceptions of competition	2.46 (.21)	3.21 (.23)	2.27 (.22)
Response times	1190.35 (46.50)	1341.12 (58.36)	1253.55 (50.63)

Note: *SEM*s in parenthesis.

each other ($ps_{paired} < .05$), except those of incongruent and control ($p_{paired} = .091$). Participants reported stronger perceptions of personal control for congruent than incongruent conditions, F (2, 166) = 22.121, $p < .01$ ($\eta_p^2 = .21$), with all means being significantly different from each other, except for those of congruent and control ($p_{paired} > .05$). Stronger perceptions of competition were reported for the incongruent than congruent conditions, F (2, 166) = 59.688, $p < .01$ ($\eta_p^2 = .42$), with all means being significantly different from each other ($ps_{paired} < .05$).

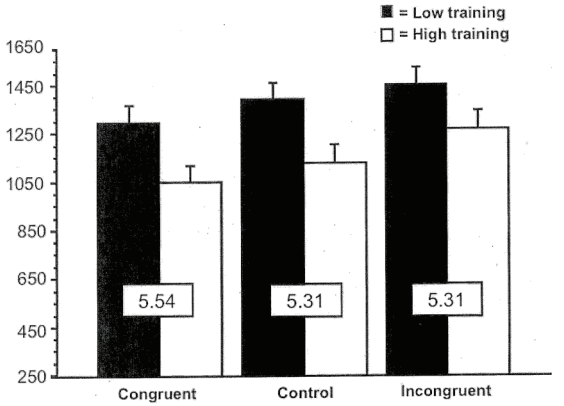

Figure 5: Mean response time (ms) as a function of degree of training and condition. Error bars indicate *SEM*s. Presented in the graph are the mean responses to the question, "How strongly do you feel that the urge to name the color was due to your 'self'?" (1 signified "not at all due to self" and 8 signified "absolutely due to self") for each Stroop-like condition.

It is important to note that, though RT was not the focus of this study, we did replicate the RT effects of MacLeod and Dunbar (1988). Based on previous research (Morsella, Gray et al., 2009; Morsella, Wilson et al., 2009; van Veen et al., 2001; Woodworth and Schlosberg, 1954), we excluded from analysis RTs below 200 ms and above 3.5 s, resulting in the loss of 1,103 (18%) out of 6,048 trials. (Importantly, the same pattern of results is obtained with the unconventional window of 200 ms to 6 s, with only 6% data loss.) As illustrated in Figure 5, there was a main effect of condition (congruent, control, and incongruent), F (2, 164) = 24.777, $p < .0001$ ($\eta_p^2 = .23$) and training, in which highly trained participants (training trials = 576) were faster than those less trained (training trials = 16), F (1, 82) = 5.357, $p < .05$ ($\eta_p^2 = .06$). There was no interaction between the interference condition and degree of training ($p = .30$). In a by-subject analysis, RTs did not correlate with perceptions of action authorship, control, or competition ($ps_{paired} > .05$).

Color-Naming Phase

Typing errors resulted in the loss of 164 (0.9%) of 18,144 ratings. Participants perceived the urge to shape-name as less due to the self during the incongruent than congruent conditions, F (2, 166) = 8.534, $p < .01$ ($\eta_p^2 = .09$), with all the means being significantly different from each other ($ps_{paired} < .05$), except those of congruent and control, which were marginally non-significant ($ps_{paired} = .053$). Participants reported stronger perceptions of personal control for congruent than incongruent conditions, F (2, 166) = 17.716, $p < .01$ ($\eta_p^2 = .18$), with all means differing from each other significantly ($ps < .05$). Stronger perceptions of competition were reported for incongruent than congruent conditions, F (2, 166) = 35.728, $p < .01$ ($\eta_p^2 = .30$), with all means differing significantly from each other ($ps_{paired} < .05$).

Reaction time trimming resulted in the loss of 1,410 (23%) out of 6,048 trials. (The same pattern of results is obtained with data spanning from 200 ms to 6 s, with only 9% data loss.) There was a main effect of condition, F (2, 164) = 15.148, $p < .01$ ($\eta_p^2 = .16$), and a trend in which highly trained participants were faster than less trained participants, F (1, 82) = 2.925, $p = .09$ ($\eta_p^2 = .03$), but no interaction between condition and degree of training ($p = .78$). In a by-subject analysis, RTs did not correlate with any of the ratings ($ps > .05$).

Discussion. Conflict influenced the sense of agency and action-plan authorship in a dynamic and contextualized fashion. As predicted, in the shape-naming phase, participants perceived the urge to color-name as less due to the self during the incongruent than congruent conditions, and, in the color-naming phase, participants perceived the urge to shape-name as less due to the self during the incongruent than congruent conditions. In this study, we demonstrated that the activation of action plans countering current goals is perceived to be less associated with the self than the activation of plans that happen to be compatible with current goals, regardless of the nature of the plan (e.g., color-naming versus shape-naming).

In addition, perceptions of personal control were stronger for congruent than incongruent conditions, and perceptions of competition were greater for incongruent than congruent conditions. Although the focus of this study was not on the complex relationship between RT and judgments, it seems that introspections could not be predicted by knowledge of RT data alone. As in previous studies (Morsella, Wilson et al., 2009), conflict may affect behavior and high-level, conscious metacognitive phenomena in parallel.

General Discussion

As captured by the metaphor "a monkey on one's back," we demonstrated for the first time that when an action plan is activated and counters one's action goal, that action plan is seen as less due to the self than when the concurrently activated action plan does not interfere (or facilitates) one's action goal. In Study 1, urges to read during the Stroop task were less attributed to the self in the incongruent and control conditions of the task than in the congruent condition, the only condition in which the concurrently-activated action plan (word-reading) does not interfere with, and may even facilitate, performance (MacLeod and MacDonald, 2000). This is the first demonstration of such attributions during an interference task. Corroborating our primary finding, perceptions of personal control were stronger for congruent than for incongruent conditions, and perceptions of competition were greater for incongruent than congruent conditions, as found in previous research (Morsella, Wilson et al., 2009). Replicating previous findings (Morsella, Gray et al., 2009; Morsella, Wilson et al., 2009), correlational analyses suggest that participants did not base their judgments only on observing their RTs. It is striking that participants were capable of introspecting such high-level aspects of a cognitive process as fleeting as color-naming, an act lasting less than one second. Although it is known that participants cannot introspect their own RTs at this time scale (Libet, 2004; but see recent evidence to the contrary: Corallo, Sackur, Dehaene, and Sigman, 2008), it cannot be completely ruled out that they were basing their judgments on RTs (see discussion above).

Study 2 replicated the pattern of results found in Study 1 in a different interference paradigm, thereby allaying some of the concerns about the validity of the primary results of Study 1. In Study 2, we demonstrated for the first time that these effects are contextualized and dynamic: in the shape-naming phase, participants perceived the urge to color-name as less due to the self during the incongruent than congruent conditions, and, in the color-naming phase, participants perceived the urge to shape-name as less due to the self during the incongruent than congruent conditions. As in Study 1, perceptions of personal control were stronger for congruent than for incongruent conditions, and perceptions of competition were greater for incongruent than for congruent conditions. Again, although the focus of this study was not on the complex relationship between RT and judgments, it seems that introspections could not be predicted by knowledge of RT data alone.

One unexpected finding that will require further investigation is the pattern of judgments found for the "color-naming due to self" question in Study 1. Beyond the purview of Study 1, which focused on the automatic aspect of the Stroop task (i.e., the automatic word-reading plan), this finding may reveal the ways in which participants introspect about and conceptualize the process of color-naming, the target action plan. Data from our perceptions of control and competition questions imply that, for the color-naming question, participants were not simply re-inputting the rating that they had input for the first question, and were not adopting a strategy in which, when confronted with the incongruent condition, lower ratings were always inputted for the incongruent condition, regardless of the question at hand. Was "color-naming due to the self" more for the congruent condition than for the incongruent and control conditions because the correct action plan was perceived to be stronger? Was this, in turn, because participants were incapable of detecting any interference in this condition, perhaps due to a phenomenon such as synchrony-blindness? Because of such interesting possibilities, this unpredicted pattern of results demands further contemplation and exploration.

Another limitation of the current project is that our sample was restricted to university students. These participants are familiar with laboratory studies and such familiarity may influence performance. Participants' reliable judgments could have been based, not on their experience of conflict, but on their folk beliefs about intra-psychic conflict, the sense of agency, and/or how to comport oneself in a psychological experiment. For example, perhaps participants based their ratings on heuristics such as, "if the Stroop trial is incongruent, then I will report 6 as the rating." This alternative hypothesis has been addressed before (see Morsella, Wilson et al., 2009). Although this cannot be fully ruled out by the present studies, this alternative hypothesis seems unlikely given that participants' ratings tended to vary across trials within each condition. For instance, for incongruent Stroop trials, the first 8 "word-reading due to self" ratings from a participant selected at random from Study 1 were 4, 4, 3, 4, 3, 6, 6, and 7. Of course, it may well be that participants were using a more sophisticated and nuanced heuristic when engendering our primary results. An additional limitation of the current project is that it did not take into account the potential effects of the variables of sex and age on the attributions of agency associated with conflict. Future investigations on cognitive conflict and the sense of agency, involving different kinds of population samples, will certainly be needed to qualify the kinds of conclusions that can be drawn from this present, initial project. We emphasize that this is an initial, and not a conclusive, project on the sense of agency and conflict.

Apart from these considerations, a limitation of this approach is that judgments may simply be based on task difficulty, with the efference–efference binding of incongruent conditions being more difficult than the kinds of bindings (e.g., efference binding) required in the other conditions. Data suggest that efference–efference is qualitatively distinct from the other forms of binding.

For example, in a neuroimaging study, van Veen et al. (2001) demonstrated that, though both response interference (when targets and distracters are associated with a different response, as in the Stroop incongruent condition) and perceptual interference (when distracters and targets look different but are associated with the same response) are associated with differences in performance, only the former (involving efference–efference binding) activates the anterior cingulate cortex, a brain region located on the medial surface of the frontal lobe that is interconnected with many motor areas and is believed to be involved in both conflict detection and willed processing (Botvinick, Braver, Carter, Barch, and Cohen, 2001; Brown and Braver, 2005; Crick, 1995; Mayr, 2004). Consistent with the idea that the conflict among plans is what is primarily driving our sense of agency effects, it has been shown that, independent of suppression or other forms of interference (e.g., perceptual interference), and on the basis of a priori theoretical predictions (Morsella, 2005), merely sustaining incompatible intentions (e.g., to point left *and* right) leads to subjective, metacognitive effects that are greater than those associated with sustaining compatible intentions (e.g., to point left *and* utter a word; Gray, Bargh, and Morsella, 2013; Morsella, Gray et al., 2009). This datum demonstrates that introspections about agency in a cognitive task are due not simply to self-observations of RT.

In conclusion, we hope that these initial findings about the liaison between intra-psychic conflict and the sense of agency (including perceptions authorship, control, and competition) will provide a foundation for a deeper understanding of the cognitive construction of the self, a mental content that is intimately related to Dasein. In addition, we hope that such an experimentally-based approach will one day illuminate the nature of "hotter" conflicts involving self-control and disorders of agency.

References

Baker, T. B., Piper, M. E., McCarthy, D. E., Majeskie, M. R., and Fiore, M. C. (2004). Addiction motivation reformulated: An affective processing model of negative reinforcement. *Psychological Review, 111*, 33–51.

Berti, A., and Pia, L. (2006). Understanding motor awareness through normal and pathological behavior. *Current Directions in Psychological Science, 15,* 245–250.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences, 30,* 481–548.

Botvinick, M., M., and Bylsma, L., M. (2005). Distraction and action slips in an everyday task: Evidence for a dynamic representation of task context. *Psychonomic Bulletin and Review, 12,* 1011–1017.

Botvinick, M. M., Braver, T. S., Carter, C. S., Barch, D. M. and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108,* 624–652.

Brown, J. W., and Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science, 307,* 1118–1121.

Cohen, J. D., Dunbar, K., and McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97,* 332–361.

Cohen, J. D., MacWhinney, B., Flatt, M., and Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavior Research Methods, Instruments, and Computers, 25*, 257–271.

Corallo, G., Sackur, J., Dehaene, S., and Sigman, M. (2008). Limits on introspection: Distorted subjective time during the dual-task bottleneck. *Psychological Science, 19*, 1110–1117.

Crick, F. (1995). *The astonishing hypothesis: The scientific search for the soul*. New York: Touchstone.

Curtis, C. E., and D'Esposito, M. (2009). The inhibition of unwanted actions. In E. Morsella, J. A. Bargh, and P. M. Gollwitzer (Eds.), *The Oxford handbook of human action* (pp. 72–97). New York: Oxford University Press.

Eriksen, C. W., and Schultz, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception and Psychophysics, 25*, 249–263.

Etkin, A., Prater, K., Hoeft, F., Menon, V., and Schatzberg, A. (2010). Failure of anterior cingulate activation and connectivity with the amygdala during implicit regulation of emotional processing in generalized anxiety disorder. *American Journal of Psychiatry, 167*, 545–554.

Fodor, J. A. (1983). *Modularity of mind: An essay on faculty psychology*. Cambridge, Massachusetts: MIT press.

Freud, S. (1938). *The basic writings of Sigmund Freud* [A. A. Brill, Ed. and Trans.]. New York: Modern Library.

Frith, C. D., Blakemore, S. J., and Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London, 355*, 1771–1788.

Gazzaley, A., and D'Esposito, M. (2007). Unifying prefrontal cortex function: Executive control, neural networks and top–down modulation. In B. Miller and J. Cummings (Eds.), *The human frontal lobes: Functions and disorders* (pp. 187–206). New York: Guilford Press.

Gray, J. R., Bargh, J. A., and Morsella, E. (2013). Neural correlates of the essence of conscious conflict: fMRI of sustaining incompatible intentions. *Experimental Brain Research, 229*, 453–465.

Haggard, P., Aschersleben, G., Gehrke, J., and Prinz, W. (2002). Action, binding and awareness. In W. Prinz and B. Hommel (Eds.), *Common mechanisms in perception and action: Attention and performance* (Vol. XIX, pp. 266–285). Oxford: Oxford University Press.

Heckhausen, H., and Beckmann, J. (1990). Intentional action and action slips. *Psychological Review, 97*, 36–48.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior. The Hixon symposium* (pp. 112–146). New York: Wiley.

Libet, B. (2004). *Mind time: The temporal factor in consciousness*. Cambridge, Massachusetts: Harvard University Press.

Livnat, A., and Pippenger, N. (2006). An optimal brain can be composed of conflicting agents. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 3198–3202.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes, 65*, 272–292.

MacLeod, C.M., and Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 126–135.

MacLeod, C. M., and MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: Uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences, 4*, 383–391.

Mayr, U. (2004). Conflict, consciousness, and control. *Trends in Cognitive Sciences, 8*, 145–148.

Metcalfe, J., and Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review, 106*, 3–19.

Molapour, T., Berger, C. C., and Morsella, E. (2011). Did I read or did I name? Diminished awareness of processes yielding identical 'outputs.' *Consciousness and Cognition, 20*, 1776–1780.

Morsella, E. (2005). The function of phenomenal states: Supramodular interaction theory. *Psychological Review, 112*, 1000–1021.

Morsella, E., and Bargh, J. A. (2011). Unconscious action tendencies: Sources of 'un-integrated' action. In J. Decety and J. Cacioppo (Eds.), *Oxford handbook of social neuroscience* (pp. 335–347). New York: Oxford University Press.

Morsella, E., Berger, C. C., and Krieger, S. C. (2011). Cognitive and neural components of the phenomenology of agency. *Neurocase, 17*, 209–230.

Morsella, E., Gray, J. R., Krieger, S. C., and Bargh, J. A. (2009). The essence of conscious conflict: Subjective effects of sustaining incompatible intentions. *Emotion, 9*, 717–728.

Morsella, E., Krieger, S. C., and Bargh, J. A. (2009). The function of consciousness: Why skeletal muscles are "voluntary" muscles. In E. Morsella, J. A. Bargh, and P. M. Gollwitzer (Eds.), *Oxford handbook of human action* (pp. 625–634). Oxford University Press.

Morsella, E., Wilson, L. E., Berger, C. C., Honhongva, M., Gazzaley, A., and Bargh, J. A. (2009). Subjective aspects of cognitive control at different stages of processing. *Attention, Perception, and Psychophysics, 71*, 1807–1824.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition, 107*, 179–217.
Perky, C. W. (1910). An experimental study of imagination. *American Journal of Psychology, 21*, 422–452.
Prinz, W. (2003). How do we know about our own actions? In S. Maasen, W. Prinz, and G. Roth (Eds.), *Voluntary action: Brains, minds, and sociality* (pp. 21–33). London: Oxford University Press.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662 .
van Veen, V., and Carter, C. S. (2006). Conflict and cognitive control in the brain. *Current Directions in Psychological Science, 5*, 237–240.
van Veen, V., Cohen, J. D., Botvinick, M. M., Stenger, V. A., and Carter, C. C. (2001). Anterior cingulate cortex, conflict monitoring, and levels of processing. *Neuroimage, 14*, 1302–1308.
Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, Massachusetts: MIT Press.
Wegner, D. M. (2003). The mind's best trick: How we experience conscious will. *Trends in Cognitive Science, 7*, 65–69.
Winkielman, P., Schwarz, N., Fazendeiro, T., and Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch and K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189–217). Mahwah, New Jersey: Lawrence Erlbaum.
Woodworth, R. S., and Schlosberg, H. (1954). *Experimental psychology* (second edition). New York: Holt, Rinehart & Winston.