



Self-deception and impaired categorization of anomaly

Jordan B. Peterson^{a,*}, Erin Driver-Linn^b, Colin G. DeYoung^a

^a*Department of Psychology, University of Toronto, 100 St. George Street, Toronto, Ontario, Canada M5S 3G3*

^b*Harvard University, William James Hall, Cambridge, MA 02138, USA*

Received 20 February 2001; received in revised form 1 August 2001; accepted 28 August 2001

Abstract

One hundred and forty community volunteers were prescreened for upper and lower quartile scores using the Balanced Inventory of Desirable Responding [BIDR; Paulhus, D.L. (1988). *Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding. Unpublished manual.* Vancouver, British Columbia; University of British Columbia, Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. Wrightsman (Eds.), *Measures of personality and social-psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press, 1991], and classified into stable High ($n = 14$) and Low ($n = 15$) self-deception groups using the Self-Deceptive Enhancement subscale of the BIDR. Participants identified normal and anomalous computer-displayed playing cards [following Bruner, J. & Postman, L. (1949). *Journal of Personality*, 18, 206–223], presented for short (~ 16 ms), then increasingly longer durations. High and low self-deceivers identified the normal cards equally rapidly. Highs, however, took twice as many trials as lows ($M = 11.21$, $S.D. = 9.65$, vs. $M = 5.00$, $S.D. = 3.87$) to identify the anomalous card correctly twice ($t [16.85]$ corrected for unequal variances = -2.25 , $P = 0.019$, one-tailed). Self-deception thus appears associated with impaired categorization of anomaly. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Self-deception; Anomaly; Exploration; Information processing; Categorization

An anomaly is an unexpected event of unspecified significance. It has long been known that anomaly elicits non-specific anxiety, particularly when it is rapidly manifested or proximal in nature (Dollard & Miller, 1950). More recent investigations indicate that anomaly-related anxiety emerges as a consequence of the unexpected or undesired disruption of goal-directed activity (Carver & Scheier, 1982; Gray, 1982, 1987; Oatley & Johnson-Laird, 1987), and that it constitutes

* Corresponding author. Tel.: +1-416-978-7619; fax: +1-416-978-4811.

E-mail address: jpeterson@psych.utoronto.ca (J.B. Peterson).

rapidly generated but relatively low resolution information about potential sources of harm (LeDoux, 1996; Peterson, 1999a; 1999b). This rapidly generated, low resolution, affective information appears akin to Damasio's (1994) "somatic marker," which serves initially to delimit and direct voluntary attention.

Under optimal conditions, when initial wary attention indicates that nothing additionally unpredictable or otherwise threatening is immediately likely to occur, an anomalous event also produces curiosity and exploratory behavior (Blanchard & Blanchard, 1989; Dollard & Miller, 1950; Gray, 1982, 1987). Exploratory behavior may be conceived as a multi-stage process, which transforms the anxiety-inducing anomaly into detailed, explicit utilitarian information. Such information can be used to update currently dysfunctional plans and goals. The stages of exploration appear to include: (1) registration of the undifferentiated anomaly, most likely in the form of "something other than what was expected"; (2) assignment of provisional relevance to the undifferentiated anomaly, in the form of an affective marker; (3) active motoric or abstract examination of the anomaly, undertaken in the attempt to extract more information about its significance, functional and objective; (4) modification of perceptual, emotional, or cognitive category and habit, as a consequence of the integration of newly gained information; and (5) transformation of plans and goals, so that the environment characterized by presence of the anomaly is once again mastered and rendered useful and predictable (Peterson, 1999a).

The information derived from such investigative exploration comes at a price, however. First, exploration is effortful (Ohman, 1979, 1987). This is perhaps because the initial phases of processing novel information activate large cortical areas (Tulving, Markowitsch, Kapur, Habib, & Houle, 1994), at high metabolic cost (Roland, Eriksson, Stone-Elander, & Widen, 1987). Second, exploration is risky. Active exploration of the unknown exposes the explorer to potential danger, both environmental and psychological, and this danger must be taken into account. Rats re-exploring a once-safe area, for example—contaminated by the recent presence of a cat—hunch down and make rapid "corner runs" through the area to protect themselves from detection, while gathering valuable but dangerous information about the area's current safety (Blanchard & Blanchard, 1989). Human beings exploring something novel—whether territory or abstraction—run the additional risk of upsetting their most valuable and time-honored abstract presuppositions (Kelly, 1969; Kuhn, 1970), as the implications of the unexpected can reveal themselves at increasingly basic or paradigmatic, and therefore increasingly troublesome, levels of understanding. In short, there is potent *a priori* or unlearned motivation for avoiding useful exploration in the face of the genuinely dangerous unknown (Peterson, 1999a, 1999b): what you don't yet know *can* hurt you.

Self-deception, a commonly used but problematically defined term (c.f. Mele, 1997), might therefore profitably be construed simply as the opposite of exploration in the face of anomaly (Peterson, 1999a, 1999b). We hypothesize, specifically, that the self-deceptive individual fails to engage in the latter stages (3–5) of the process that turns evidence of error into information useful for the modification of non-productive plans and goals. This failure leads to a mental state in which the self-deceiver continues to hold beliefs that have been indicated as problematic by his or her own affective response (see Greenwald, 1988, 1992, 1997, for an alternative failure-of-information-processing account of self-deception). The greater the perceived anomaly—that is, the broader or more basic the plans, goals, or conceptions it disrupts—the more negative the affective response, and the more potent the motivation for self-deception.

This model avoids the logical quagmire surrounding traditional approaches to self-deception, such as Sackeim and Gur's (1978, 1979), that are predicated on the idea that self-deceivers must hold two conflicting beliefs and, even more problematically, must remain unaware that one of them is held. The requirement that information must be known but not known at the same time is paradoxical. How else, however, could self-deception be differentiated from mere ignorance? The answer, from the perspective outlined above, is that self-deceivers do not hold conflicting beliefs. Instead they ignore evidence that their current beliefs may be in error. Our use of possession of conflicting evidence rather than conflicting belief as a criterion for self-deception is similar to Mele's (1997). Unlike Mele, however, we do not suggest that this evidence is objective; we specify instead that it is affective and thus subjectively defined. This specification avoids Sackeim and Gur's (1978) criticism that possession of conflicting evidence, conceived as objective information, is insufficient to define self-deception because the individual may simply be ignorant of the significance of that evidence.

To test the hypothesis that self-deception is related to failure to process anomaly, we utilized a paradigm employed by Bruner and Postman (1949) to demonstrate that explicit classification of anomaly is more difficult, in general, than explicit classification of something familiar. These researchers asked subjects to identify playing cards, presented for brief but increasing time periods using a tachistoscope (a pre-computer-era device, utilized for the rapid presentation of visual stimuli). Some of these cards were anomalous, with color and suit reversed. Bruner and Postman hypothesized that "for as long as possible and by whatever means available, [participants] will ward off the perception of the unexpected, those things which do not fit [their] prevailing set" (p. 208). They found, as predicted, that recognition of the "trick" cards required longer exposure (300 ms, for 65% of the population) than recognition of the normal cards (10 ms, for 65% of the population).

These results might not be regarded as surprising today. Categorization of familiar stimuli—a process dependent on information gathered during initial exploration—rapidly becomes something predicated on memory and automatized. Our perception makes use of an "internal experiential template" (Miller, Galanter, & Pribram, 1960) that allows us to anticipate "regularities" and to become efficient top-down processors of information (Rumelhart, Smolensky, McClelland, & Hinton, 1986). These "regularities" are not so much objectively stable real-world phenomena as categories with stable functional significance (Tranel, Logan, Frank, & Damasio, 1997; Wittgenstein, 1968), which remain appropriately intact as long as they continue to work effectively in the movement toward one's goals (Miller, 1956; Peterson, 1999a).

Efficient processing of the familiar, however, may weaken the ability to respond to potentially relevant changes in the environment (Johnston, Hawley, & Farnham, 1993; Levin & Simons, 1997). Expectations "bias perception toward what is familiar or expected" and render the organism "relatively insensitive to environmental changes" (Hawley, Johnston, & Farnham, 1994, p. 261). This bias or relative insensitivity appears as a phenomenon potentially characterized by significant individual differences.

Although Bruner and Postman observed "intact, normal organisms" (p. 208), in the form of Harvard and Radcliffe undergraduates, they noticed a great deal of between-subject variance in reaction to anomalous cards. Most participants showed a "dominance" reaction, becoming bound by color (so that a black four of hearts would be seen as a spade) or by form (so that a black four of hearts would be seen as red). As the anomalous card was presented for progressively

longer durations, however, the response types diverged. After noting that something was “wrong” with the card, some individuals soon recognized the nature of the incongruity. Others, however, seemed unable to reorganize their categories so that accurate and explicit perception could become possible. These participants showed either a “compromise” reaction, reporting that the trick card was purple or gray, or a “disruption” reaction, reporting, for example, “I don’t even know what the hell it is now, not even for sure whether it’s a playing card” (p. 214).

Bruner and Postman did not empirically address these individual differences in reaction to anomaly. It appears possible, however, that the tendency to self-deceive might be associated with such differences, at least inasmuch as self-deception entails unwillingness or inability to adjust category or habit in the face of anomaly (i.e. to proceed through stage 4 above). The present study therefore attempted to determine if individuals high in self-deception were characterized by decreased efficiency of recategorization, following exposure to anomalous information. It was additionally hypothesized that individuals low and high in self-deception would be characterized by equal initial affect-driven reaction to the first appearance of anomaly, despite the decreased final efficiency of the highly self-deceptive group.

1. Method

1.1. *Participants: recruitment and selection*

One-hundred and forty women and men, fluent in English, recruited from the general public in Cambridge, Massachusetts, were screened to form a smaller pool of participants who were either high or low in self-deception. One or two researchers approached potential participants sitting alone in one of several public areas around Harvard Square, usually during lunch time, and asked if they would like to fill out a brief questionnaire with a 50% chance to participate in a follow-up psychology study. The cover page assured participants of the anonymity of their responses, and indicated that 50% of respondents would be telephoned later to come in to the laboratory. All screened participants provided their names and phone numbers on the cover page, so that they could be contacted, but were told explicitly that their names would not be associated with their questionnaire responses. Instead, an ID number, written in the corner of each sheet, was used for this purpose. While participants watched, the experimenter removed the cover page from the questionnaire and placed it randomly among others in a large envelope.

The screening questionnaire was the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1988, 1991), which consists of three subscales of 20 items each: Self-Deceptive Enhancement, Impression Management, and Self-Deceptive Denial. Self-Deceptive Enhancement measures the tendency to exaggerate one’s abilities (e.g. “I am fully in control of my own fate”). It assesses a general egoistic or overconfident response tendency (Paulhus & John, 1998). Impression Management measures the tendency to deny socially undesirable behavior (e.g. “I sometimes tell lies if I have to”—reversed). Self-Deceptive Denial measures the tendency to claim that one does not experience socially undesirable thoughts and feelings (e.g. “I could never enjoy being cruel”). Impression Management and Self-Deceptive Denial assess a general moralistic or conforming response tendency (Paulhus & John, 1998). Responses are given on a seven-point Likert scale ranging from *not true* to *very true*, with only extreme high responses (6 and 7, or 1 and 2

for reversed items) being scored (Paulhus, 1991). The BIDR has a high degree of internal consistency (Cronbach's $\alpha = 0.83$) and adequate test-retest reliability (0.65 to 0.69) (Paulhus, 1991).

The BIDR, like other similar, though less differentiated, measures of socially desirable responding (e.g. the Self-Deception Questionnaire, SDQ, Sackeim & Gur, 1978; the Marlowe–Crowne Social Desirability Scale, MCSD, Crowne & Marlowe, 1960), is predicated on the assumption that there are universally occurring, but undesirable, thoughts and behaviors, the denial of which is indicative of bias. High scores on such scales have been consistently associated with behavioral measures reasonably indicative of self-deception. The MCSD, for example, is positively correlated with deficits in memory for negative autobiographical events (Davis, 1990) and with impaired ability to perceive negative emotional stimuli explicitly (Schwartz, 1990), while the BIDR is positively correlated with increased illusion of control in uncontrollable situations (Paulhus & Reid, 1991), greater self-serving bias after failure (Paulhus, 1988), and claimed familiarity with non-existent products (Paulhus, 1988).

Individuals were selected from the pool of screened participants if their scores fell in the top or bottom quartile of the population on the total BIDR score from all three subscales. An independent experimenter entered the data from the screening questionnaires and identified all ID numbers that met this criterion. Another experimenter scheduled participants by telephone, using the contact information from the cover page.

Twenty-nine participants (10 female; 19 male) who met inclusion criteria were willing to report to the laboratory between 3 weeks and 4 months after screening. Participants ranged in age from 20 to 59 ($M = 32$, $S.D. = 10$). Twenty-five were Caucasian American, two were Black American, and two were Hispanic. Self-reported household income ranged from \$3000 to \$120,000 ($M = \$40,000$, $S.D. = \$29,390$). All participants except one had attended college, and five reported graduate degrees.

1.2. Procedure: laboratory testing

Participants provided informed consent upon arriving at the laboratory and then completed (1) a *Demographic Form*, which asked for information about age, sex, ethnic background, household income, and education; (2) the short form of the *Eysenck Personality Questionnaire* (EPQ; Eysenck, Eysenck, & Barrett, 1985), which includes Extraversion, Neuroticism, and Psychoticism Scales (reliability statistics for all scales are reported in Eysenck et al., 1985); and (3) the *BIDR*, for a second time.

Participants then completed a 15-min computerized version of the Bruner–Postman card identification task. All participants were videotaped while completing this task, so that exploratory analyses of their facial and motor responses to the appearance of the trick card might be conducted. Experimenters running participants through the protocol were blind to their level of self-deception.

Each participant was asked to sit approximately 18 inches from a 15" color video monitor. Instructions were presented on the monitor screen. Participants were instructed to describe whatever appeared on the screen during the task, and to tell the experimenter to move on to the next trial when they were satisfied they had provided appropriate identification.

Participants were not told they were identifying playing cards. Instead, they were asked to describe the objects that appeared on the screen, in as much detail as possible. Each participant

was presented with five life-size cards presented centrally, serially, on the computer screen. Four cards were normal and one was anomalous. Cards were presented in two orders, randomly assigned across participants (to control for potential nonspecific order effects). The anomalous card was presented fourth or fifth in the sequence. All participants saw a 9 of hearts (red), a 5 of spades (black), and a 7 of clubs (black), in that order. Half then saw a 4 of hearts (black) and a 3 of diamonds (red), while the other half saw a 3 of diamonds (red), then a 4 of hearts (black). All participants were thus exposed to, and had to correctly identify, a heart and a spade, before seeing the trick card. The club was included to ensure that hearts and spades did not attract undue attention.

Each card was presented at a given duration three times (starting with a presentation period of approximately 16 ms—the screen refresh rate), after which the duration doubled. The 9 of hearts, for example, was presented three times at 16 ms, three times at 32 ms, three times at 64 ms, and so on, until correctly identified twice in a row. However, all cards were shown a minimum of five to eight times (randomly computer-chosen), regardless of number of trials to correct identification. This unpredictable mode of presentation was chosen to eliminate any contextual cues with regard to the oddity of the trick card. Pilot testing had indicated that virtually all participants were able to identify normal cards correctly on the first or second trial. This meant that normal cards had to be presented multiple times, so that multiple presentation of the anomalous card would not in itself appear strange.

The experimenter recorded all correct and incorrect responses by pressing designated keys on the keyboard. At the end of the task, participants were debriefed regarding the general hypotheses of the experimenter, thanked for their participation, and paid \$5.00 for their time.

2. Results

2.1. Final determination of self-deception groups

As described above, participants completed the BIDR at screening and again during the laboratory session. Initial screening narrowed the pool of potential laboratory participants to those who scored in the upper or lower quartile on the BIDR (a score ≥ 21 , or ≤ 11). For the final analysis, it was decided to use a single subscale of the BIDR, Self-Deceptive Enhancement, to produce self-deception groups, due to limited sample size and to the fact that Self-Deceptive Enhancement appears to be the best validated and most commonly used of the three subscales as a measure of self-deception (e.g. Fossum & Barrett, 2000; Paulhus, 1991). Impression Management has been shown to be more susceptible to situational pressures that encourage or discourage honesty, such as instructions to be honest versus to fake good, anonymous response conditions, or bogus-pipeline conditions (Paulhus, 1986, 1991; Paulhus & John, 1998), thus indicating a confounding of self-deceptive bias with conscious exaggeration. Self-Deceptive Denial is very similar to Impression Management in content (their only major difference being that Impression Management items refer to objectively verifiable behaviors—e.g. “I never read sexy books or magazines”—while Self-Deceptive Denial items refer to cognitions and feelings that may only be verified internally—e.g. “I never enjoy watching sexy scenes in movies”), and, perhaps unsurprisingly, the two scales have been found to be highly correlated (Paulhus, 1999;

Paulhus & Reid, 1991; in the present study: $r = 0.71$, $P < 0.001$). Consequently, the BIDR is most commonly used without the Self-Deceptive Denial subscale (Paulhus, 1991, 1999), and Self-Deceptive Enhancement seems the safest scale to use as an indicator of self-deception. Further, given that Self-Deceptive Enhancement is related specifically to an egoistic and overconfident type of bias (Paulhus & John, 1998), it appears especially clearly related to the conceptualization of self-deception as the ignoring of evidence that one is in error. At any rate, Self-Deceptive Enhancement scores were highly correlated with the total BIDR scores in our sample ($r = 0.83$, $P < 0.001$), and similar results emerged using either criterion.

A median-split of the averaged Self-Deceptive Enhancement scores from the screening and the laboratory administration yielded 15 participants low in self-deception and 14 participants high in self-deception. This procedure allowed for categorization of participants on the basis of assessments conducted at two different points in time, enhancing the reliability of the categorization and accounting for regression to the mean. One individual in the upper quartile of scores from the initial screening fell below the averaged median. No individuals moved from the lower screening quartile to the final high self-deception group.

2.2. *Personality correlations*

Pearson correlations were computed between participants' averaged Self-Deceptive Enhancement scores and Eysenck Personality Questionnaire subscale scores. In keeping with the results of previous investigations (Sackeim & Gur, 1979), self-deception was strongly and negatively correlated with the Neuroticism Scale ($r = -0.49$, $P < 0.01$, two-tailed). A negative correlation was also found between self-deception and the Psychoticism Scale, ($r = -0.43$, $P < 0.05$, two-tailed). Self-deception and Extraversion were correlated in the positive direction ($r = 0.47$, $P < 0.05$, two-tailed). This pattern of correlations appears fairly typical (Fossum & Barrett, 2000; Meston, Heiman, Trapnell, & Paulhus, 1998, Sackeim & Gur, 1979), especially if the EPQ is assumed to be comparable to measures of the Five Factor Model of personality (Big Five), with Psychoticism reflecting a combination of the Big Five factors Agreeableness and Conscientiousness, reversed (Costa & McCrae, 1992; Digman, 1997). Such a pattern is also consistent with the tendency to claim heightened social and intellectual abilities and to deny anxiety that characterizes the egoistic response bias assessed by Self-Deceptive Enhancement (Paulhus & John, 1998).

2.3. *Identification of the anomalous card*

Preliminary analyses revealed no significant difference between the number of trials that participants needed to recognize the trick card and its appearance fourth or fifth in sequence, ($t[27] = 0.24$, $P = 0.814$, two-tailed). Furthermore, no significant gender difference emerged, with regards to number of trials needed for recognition ($t[27] = 0.25$, $P = 0.807$, two-tailed). In consequence, subsequent analyses were collapsed across order of trick card presentation and gender. There were no significant differences between individuals low or high in self-deception with regards to trials needed to identify the normal playing cards (low: $M = 1.16$, S.D. = 0.39; high: $M = 1.17$, S.D. = 0.23; $t[27] = -0.08$, $P = 0.935$, two-tailed). However, high self-deceivers needed an average of 9.29 trials (S.D. = 9.81) to identify the trick card correctly once, while low self-deceivers needed an average of only 3.67 trials (S.D. = 3.62; Levene's test for equality of variance,

$F=21.76$, $P<0.001$; t [16.27] corrected for unequal variance = -2.02 , $P=0.030$, one-tailed). Pilot testing had indicated that some individuals identified the trick card correctly, but then reverted to an incorrect categorization, before committing themselves to the correct response. In consequence, trials to second correct identification were also recorded. High self-deceivers needed an average of 11.21 trials (S.D. = 9.65) to identify the trick card correctly twice, while low self-deceivers needed an average of only 5.00 trials (S.D. = 3.87) (Levene's test for equality of variance, $F=15.06$, $P=0.001$; t [16.85] corrected for unequal variance = -2.25 , $P=0.019$, one-tailed). Because the standard deviations for card recognition differed so markedly between the high and low self-deceivers, the data were also considered categorically (following Rosenthal, & Rosnow, 1991, pp. 538–539, with proportions used to compute S_p^2 conservatively replaced with 0.25 to account for low frequency cells), as presented in Table 1.

2.4. Noticeable reaction to the anomalous card

Four raters, blind to participants' level of self-deception, assessed participants' videotaped reaction to the first appearance of the trick card. Raters were asked to assess participants' reactions defined as "changes in verbal or non-verbal behavior upon first presentation of the anomalous card, relative to the first appearance of the other cards." Participants were rated on a scale ranging from 1 (no noticeable reaction to first appearance of trick card, relative to first appearance of other cards) to 7 (strong reaction to first appearance of trick card, relative to the first appearance of other cards). Correlations within individual raters' assessments of the participants ranged from 0.34 to 0.74. The mean reliability of a single judge (an estimate of the reliability of the typical rater in this situation) was $r_1=0.57$. The effective or aggregate reliability of the total set of raters (reliability of mean 4 raters' ratings) was $r_4=0.84$.

All participants were characterized by a noticeable reaction to the first appearance of the trick card, relative to the first appearance of the normal cards ($M=4.07$, S.D. = 1.14). From the videos it was apparent that these reactions took different forms. Participants leaned forward slightly, narrowed their eyes, raised an eyebrow, or paused before verbally describing the card. As hypothesized, there was no significant difference between individuals low ($M=4.00$, S.D. = 1.06) and high ($M=4.15$, S.D. = 1.27) in self-deception (t [27] = -0.38 , $P=0.71$, two-tailed).

Table 1

Proportions of high self-deceivers in groups needing a small, medium or large number of trials to recognize the anomalous card

Level of self-deception	Number of trials to recognition		
	Small (1–5)	Medium (6–15)	Large (16–30)
Low ($n=15$)	11	4	0
High ($n=14$)	5	4	5
Proportion of high self-deceivers	0.31	0.50	1.00

Contrast of proportions $Z=2.65$, $P<0.001$, one-tailed, $r=0.49$ (proportions to compute S_p^2 conservatively replaced with 0.25 to account for low frequency cells).

3. Discussion

When an anomaly invalidates expectations predicated on current belief, thereby provoking an affective response and implicitly casting doubt on ongoing plans and goals, exploration can produce new information and lead to the modification of unsuccessful strategies. Alternatively, dysfunctional plans and goals and the categories and habits on which they depend may remain intact, despite their evident lack of functional utility. The goal in the present study was identification of familiar and anomalous cards. Low and high-self deceivers identified the familiar cards rapidly and equally proficiently, as expected. High self-deceivers, however, took more than twice as many trials to identify the anomalous cards, despite the apparent equivalence of subjective reactions to the initial appearance of these cards, as observed from their videotaped responses. We have no way to verify that participants' subjective reactions included anxiety; nor, given the trivial nature of the anomaly in question, would we expect anything but the slightest twinge of negative affect in response to our manipulation. Nonetheless, that participants emitted noticeable behavioral reactions is compatible with the presence of such a "somatic marker" (Damasio, 1994) and serves, at least, to verify that they perceived something subjectively defined as anomalous.

As outlined in the introduction, exploration of an anomalous event appears to be a process completed in several stages. The results of the present study imply that individuals high in self-deception are less efficient in the completion of at least the final stages of the exploratory process, which involve adjustment to information derived from examination of the unexpected experience. Specifically, they are significantly slower to adjust categorization to accommodate the anomalous card, corresponding to an impediment at stage four. Some of the difference between groups may have been due to a greater desire for accuracy on the part of the low self-deceivers; however, the fact remains that a third of the high self-deceivers took from 16 to 30 trials to identify the anomalous card (Table 1), which meant that the card was presented at durations ranging from half a second to several seconds without being correctly identified. It seems likely that such difficulty in adjusting category and habit will tend to make it difficult—presumably impossible if recategorization is avoided entirely rather than simply slowed—to proceed to the fifth and final stage, transformation of plans and goals. As pragmatic adaptation logically requires such transformation to take place, self-deception appears unlikely to be beneficial, when considered over a sufficient time-frame.

There has been much discussion in the last decade regarding the relative merits of self-deception. Some believe that self-deception is positively related to mental health (Taylor, 1989; Taylor & Brown, 1988, 1994), at least in "optimal doses" (Baumeister, 1989; Sackeim, 1983). Others argue, more classically, that self-deception is pathological, when considered from the larger social perspective or across broad time-frames (Colvin & Block, 1994; Colvin, Block, & Funder, 1995; Goleman, 1985; Shedler, Mayman, & Manis, 1993). Proponents of the former position buttress their hypotheses by referring to the consistently replicated negative relationship that obtains between self-deception and neuroticism (Fossum & Barrett, 2000; Meston et al., 1998; Sackeim, 1983; Sackeim & Gur, 1979; Taylor & Brown, 1988), which was replicated again in the present study. As Colvin and Block (1994) pointed out, however, individuals high in self-deception seem unlikely to provide accurate answers to questions about their negative emotional states. Indeed, in keeping with the definition of self-deception as the ignoring of anxiety-provoking anomaly, there would be good reason to expect self-deceivers to lack a consciously elaborated memory of their experience of negative affect.

Alternatively, it could be that self-deception *is* veridically associated with reduced levels of negative affect, at least in the short term, because any occurrence that induces negative affect remains something steadfastly unexplored. (Note that our model entails that self-deceivers are more likely than others to *ignore* negative affect, not that they are more likely to *experience* negative affect.) This process of rejecting exploration could theoretically protect the self-deceptive individual from painful disruption of his or her current categorization scheme (Grossberg, 1987) and help, at least temporarily, to suppress the spread of negative affect that would otherwise attend exploration of error. It can certainly be unpleasant to face up to the larger implications of one's mistakes. Such protective avoidance, however, cannot logically fail to increase the probability of unfortunate long-term consequences, as the functional mismatch between plans and goals and environment inevitably expands if uncorrected, rendering the environment increasingly hostile (Peterson, 1999a, 1999b). Some evidence for this sort of discrepancy between short- and long-term effect has appeared in investigations that do not rely solely on self report. Paulhus (1998), for example, found that individuals scoring high on Self-Deceptive Enhancement made positive first impressions on unfamiliar partners in social interactions, but that after more extended interaction these same partners evaluated the self-deceptive enhancers negatively.

Relative insensitivity to anomalous information seems bound to preserve the stability of category and habit and, therefore, the moment-to-moment stability of affect. However, behavior that does not reach its goal yet continues to be manifested produces more and more anomaly—as the unpaid bill, the unanswered letter, or the poorly completed job comprise dynamic aspects of the environment in their own right. (Try leaving a bill unpaid, and observe the amount of trouble it causes!) Self-deception is therefore likely both to produce substantial stress over the long-term, as anomalies continue to accumulate, and to feed back upon itself, such that it will become an increasingly utilized strategy, in an environment that is more and more hostile (Peterson, 1999a, 1999b). It is very interesting to note, in this regard, that subjects classified as repressors (by high scores on a measure of socially desirable responding, the MCSD,¹ in conjunction with low Taylor Manifest Anxiety scores) are characterized by very high levels of cortisol, a stress hormone (Brown, Tomarken, Orth, Loosen, Kalin, & Davidson, 1996). Given that all measures of socially desirable responding tend to be fairly strongly negatively correlated with anxiety and Neuroticism, Brown et al.'s (1996) results suggest (1) that repressors and most self-deceivers may be chronically stressed, physiologically—a finding in keeping with the idea that they respond affectively to anomaly even though they tend not to recategorize because of it, and (2) that this chronic stress has far from trivial costs, as excess levels of cortisol are associated with long-term neurophysiological and physiological damage and with increased risk for psychopathology, including depression (Raber, 1998).

Eventually, a person for whom self-deception has become habit may find even relatively benign anomaly difficult to recategorize. Taylor and colleagues (Taylor, 1989; Taylor & Brown, 1988, 1994) maintain that self-deceivers exhibit the markers of successful life adjustment, and that they remain flexible enough to respond to corrective information (Taylor, Collins, Skokan, & Aspinwall, 1989). However, it is difficult to conceive of a less existentially shattering stimulus than the presentation of an anomalous playing card under experimentally controlled circumstances.

¹ The MCSD appears to assess a combination of both the moralistic and egoistic types of bias (Paulhus, 1991; Paulhus & John, 1998) and is thus related to all subscales of the BIDR.

Nonetheless, individuals high in self-deception take longer to identify an anomalous card correctly. Given the size of our sample, our results should be replicated before this conclusion is accepted categorically. Assuming its validity, however, we find it hard to imagine how a tendency to shy away from adjusting to even trivial environmental change could produce anything but negative results when considered from a reasonably long-term and comprehensive perspective. In fact, the incorporation of new information into existing structures of category and habit seems not only necessary for learning, but perhaps indistinguishable from it.

It may be argued that even short-term reductions in anxiety provide a valuable boost, in the form of increased optimism or positive affect (Taylor & Brown, 1988), and that such optimism or positive affect grants a motivational advantage in routine and adverse contexts (Sackeim, 1983; c.f. Ashby, Isen, & Turken, 1999). Evidence that high self-deceivers perform worse after failure than low self-deceivers suggests otherwise, however (Johnson, Vincent, & Ross, 1997). Real *optimism*—a word that is probably too weak in this context—might be conceived as belief in one's ability to confront and transform anomaly when it arises, rather than as the avoidance-buttressed belief that the world is a safe and benevolent place (Peterson, 1999a). Indeed, the results of clinical studies utilizing guided exposure to the unknown and threatening as a curative mechanism (Hiss, Foa, & Kozak, 1994; Jaycox, Foa, & Morral, 1998) and the subsequent positive generalization with regards to self-efficacy that occurs (Biran & Wilson, 1981; Williams, Doseman, & Kleifield, 1984) suggest strongly that this is the case.

Acknowledgements

Preparation of this article was made possible by a Knox Fund Grant from Harvard University and by support from the Social Sciences and Humanities Research Council of Canada. We thank Daniel Higgins and Adrienne Seiffert for help in programming and modifying this task, and Mike Michaud and Mike McGarry for help in executing the study. We are also grateful to Deborah Yeh, Karen Ruggiero, and several anonymous reviewers for their comments on earlier versions of this article. Finally, we are very grateful for statistical assistance from Robert Rosenthal.

References

- Ashby, F. G., Isen, A. M., & Turken, U. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*, 3, 529–550.
- Baumeister, R. F. (1989). The optimal margin of illusion. *Journal of Social and Clinical Psychology*, 8, 176–189.
- Biran, M., & Wilson, G. T. (1981). Treatment of phobic disorders using cognitive and exposure methods: a self-efficacy analysis. *Journal of Consulting & Clinical Psychology*, 49, 886–899.
- Blanchard, D. J., & Blanchard, D. C. (1989). Antipredator defensive behaviors in a visible burrow system. *Journal of Comparative Psychology*, 103, 70–82.
- Brown, L. L., Tomarken, A. J., Orth, D. N., Loosen, P. T., Kalin, N. H., & Davidson, R. J. (1996). Individual differences in repressive-defensiveness predict basal salivary cortisol levels. *Journal of Personality and Social Psychology*, 70(2), 362–371.
- Bruner, J., & Postman, L. (1949). On the perception of incongruity: a paradigm. *Journal of Personality*, 18, 206–223.
- Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality-social, clinical, and health psychology. *Psychological Bulletin*, 92, 111–135.

- Colvin, C. R., & Block, J. (1994). Do positive illusions foster mental health? An examination of the Taylor and Brown formulation. *Psychological Bulletin*, *116*, 3–20.
- Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: negative implications for mental health. *Journal of Personality and Social Psychology*, *68*, 1152–1162.
- Costa, P. T., & McCrae, R. R. (1992). Reply to Eysenck. *Personality and Individual Differences*, *13*, 861–865.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349–354.
- Damasio, A. R. (1994). *Descartes' error: emotion, reason and the human brain*. New York: Avon Books.
- Davis, P. J. (1990). Repression and the inaccessibility of emotional memories. In J. L. Singer (Ed.), *Repression and dissociation: implications for personality theory, psychopathology, and health* (pp. 387–403). Chicago: University of Chicago Press.
- Digman, J. M. (1997). Higher-order factors of the big five. *Journal of Personality and Social Psychology*, *73*, 1246–1256.
- Dollard, J. C., & Miller, N. E. (1950). *Personality and psychotherapy*. New York: McGraw-Hill.
- Eysenck, S. B., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, *6*, 121–129.
- Fossum, T. A., & Barrett, L. F. (2000). Distinguishing evaluation from description in the personality-emotion relationship. *Personality and Social Psychology Bulletin*, *26*, 669–678.
- Goleman, D. (1985). *Vital lies, simple truths: the psychology of self-deception*. New York: Simon & Schuster.
- Gray, J. A. (1982). *The neuropsychology of anxiety: an enquiry into the functions of the septo-hippocampal system*. Oxford: Oxford University Press.
- Gray, J. A. (1987). *The psychology of fear and stress* (2nd ed.). Cambridge: Cambridge University Press.
- Greenwald, A. G. (1988). Self-knowledge and self-deception. In J. S. Lockard, & D. L. Paulhus (Eds.), *Self-deception: an adaptive mechanism?* (pp. 113–131). Englewood Cliffs, NJ: Prentice-Hall.
- Greenwald, A. G. (1992). New look 3: unconscious cognition reclaimed. *American Psychologist*, *47*, 766–779.
- Greenwald, A. G. (1997). Self-knowledge and self-deception: further consideration. In M. S. Myslobodsky (Ed.), *The mythomanias: the nature of deception and self-deception* (pp. 51–71). Mahwah, NJ: Lawrence Erlbaum Associates.
- Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, *11*, 23–63.
- Hawley, K. J., Johnston, W. A., & Farnham, J. M. (1994). Novel popout with nonsense strings: effects of predictability of string length and spatial location. *Perception & Psychophysics*, *55*, 261–268.
- Hiss, H., Foa, E. B., & Kozak, M. J. (1994). Relaps prevention program for treatment of obsessive-compulsive disorder. *Journal of Consulting & Clinical Psychology*, *62*, 801–808.
- Jaycox, L. H., Foa, E. B., & Morral, A. R. (1998). Influence of emotional engagement and habituation on exposure therapy for PTSD. *Journal of Consulting & Clinical Psychology*, *66*, 185–192.
- Johnson, E. A., Vincent, N., & Ross, L. (1997). Self-deception versus self-esteem in buffering the negative effects of failure. *Journal of Research in Personality*, *31*, 385–405.
- Johnston, W. A., Hawley, K. J., & Farnham, J. M. (1993). Novel popout: empirical boundaries and tentative theory. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 140–153.
- Kelly, G. (1969). The threat of aggression. In B. Maher (Ed.), *Clinical psychology and personality: the selected papers of George Kelly* (pp. 281–288). New York: Wiley.
- Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago: Chicago University Press.
- LeDoux, J. (1996). *The emotional brain: the mysterious underpinnings of emotional life*. New York: Simon and Schuster.
- Levin, D. T., & Simons, D. J. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin and Review*, *4*, 501–506.
- Mele, A. (1997). Real self-deception. *Behavioral & Brain Sciences*, *20*, 91–136.
- Meston, C. M., Heiman, J. R., Trapnell, P. D., & Paulhus, D. L. (1998). Socially desirable responding and sexuality self-reports. *Journal of Sex Research*, *35*, 148–157.
- Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.
- Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart & Winston.
- Oatley, K., & Johnson-Laird, P. N. (1987). Towards a cognitive theory of emotion. *Cognition and Emotion*, *1*, 29–50.

- Ohman, A. (1979). The orienting response, attention and learning: an information-processing perspective. In H. D. Kimmel, E. H. Van Olst, & J. F. Orlebeke (Eds.), *The orienting reflex in humans* (pp. 443–467). Hillsdale, NJ: Erlbaum.
- Ohman, A. (1987). The psychophysiology of emotion: an evolutionary-cognitive perspective. In P. K. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.), *Advances in psychophysiology: a research annual* (Volume 2) (pp. 79–127). Greenwich, CT: JAI Press.
- Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaire* (pp. 143–165). New York: Springer Verlag.
- Paulhus, D. L. (1988). *Assessing self-deception and impression management in self-reports: the balanced inventory of desirable responding*. Unpublished manual. Vancouver, British Columbia, Canada: University of British Columbia.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. Wrightsman (Eds.), *Measures of personality and social-psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.
- Paulhus, D. L. (1998). Interpersonal and intrapsychic adaptiveness of trait self-enhancement: a mixed blessing? *Journal of Personality and Social Psychology*, *74*, 1197–1208.
- Paulhus, D. L. (1999). *BIDR self-deceptive denial (SDD)*. Personal communication.
- Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: the interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66*, 1025–1060.
- Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, *60*, 307–317.
- Peterson, J. B. (1999a). *Maps of meaning: the architecture of belief*. New York: Routledge.
- Peterson, J. B. (1999b). Individual motivation for group aggression: psychological, mythological and neuropsychological perspectives. In L. Kurtz (Ed.), *Encyclopedia of violence, peace and conflict* (pp. 529–545). San Diego: Academic Press.
- Raber, J. (1998). Detrimental effects of chronic hypothalamic-pituitary-adrenal axis activation. From obesity to memory deficits. *Molecular Neurobiology*, *18*, 1–22.
- Roland, P. E., Eriksson, L., Stone-Elander, S., & Widen, L. (1987). Does mental activity change the oxidative metabolism of the brain? *Journal of Neuroscience*, *7*, 2373–2389.
- Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: methods and data analysis* (2nd Ed.). New York: McGraw-Hill.
- Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, & D. E. Rumelhart (Eds.), *Parallel distributed processing* (pp. 7–57). Cambridge, MA: MIT Press.
- Sackeim, H. A. (1983). Self-deception, self-esteem, and depression: the adaptive value of lying to oneself. In J. Masling (Ed.), *Empirical studies of psychoanalytical theories* (pp. 101–157). Hillsdale, NJ: Analytic Press.
- Sackeim, H. A., & Gur, R. C. (1978). Self-deception, self-confrontation, and consciousness. In G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation, advances in research and theory* (Vol. 2) (pp. 139–197). New York: Plenum Press.
- Sackeim, H. A., & Gur, R. C. (1979). Self-deception, other deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology*, *47*, 213–215.
- Schwartz, G. E. (1990). Psychobiology of repression and health: a systems approach. In J. L. Singer (Ed.), *Repression and dissociation: implications for personality theory, psychopathology and health* (pp. 405–434). Chicago: University of Chicago Press.
- Shedler, J., Mayman, M., & Manis, M. (1993). The illusion of mental health. *American Psychologist*, *48*, 1117–1131.
- Taylor, S. E. (1989). *Positive illusions: creative self-deception and the healthy mind*. New York: Basic Books.
- Taylor, S. E., & Brown, J. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*, 193–210.
- Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin*, *116*, 21–27.
- Taylor, S. E., Collins, R. L., Skokan, L. A., & Aspinwall, L. G. (1989). Maintaining positive illusions in the face of negative information: getting the facts without letting them get to you. *Journal of Social and Clinical Psychology*, *8*, 114–129.

- Tranel, D., Logan, C. G., Frank, R. J., & Damasio, A. R. (1997). Explaining category-related effects in the retrieval of conceptual and lexical knowledge for concrete entities: operationalization and analysis of factors. *Neuropsychologia*, *35*, 1329–1339.
- Tulving, E., Markowitsch, H. J., Kapur, S., Habib, R., & Houle, S. (1994). Novelty encoding networks in the human brain: positron emission tomography data. *Neuroreport*, *20*, 2525–2528.
- Williams, S. L., Doseman, G., & Kleifield, E. (1984). Comparative effectiveness of guided mastery and exposure treatments for intractable phobias. *Journal of Consulting & Clinical Psychology*, *52*, 505–518.
- Wittgenstein, L. (1968). *Philosophical investigations* (3rd ed.). (*G.E.M. Anscombe, Trans.*). New York: Macmillan.