RUNNING HEAD: GOOD TRUE SELF

Consistent Belief in a Good True Self in Misanthropes and Three Interdependent Cultures

Julian De Freitas[1], Hagop Sarkissian[2], George E. Newman[3], Igor Grossmann[4], Felipe De Brigard[5], Andres Luco[6], Joshua Knobe[3]

1 –Harvard University
2 –The City University of New York, Baruch College
3 –Yale University
4 –University of Waterloo
5 –Duke University
6 –Nanyang Technological University

Address for correspondence	:	Julian De Freitas William James Hall 33 Kirkland Street Cambridge, MA 02138
Email address	:	defreitas@g.harvard.edu
Phone	:	+1-(626)-559-6401
Word Count	:	9240 (main text)

Abstract

People sometimes explain behavior by appealing to an essentialist concept of the self, often referred to as the true self. Existing studies suggest that people tend to believe that the true self is morally virtuous, i.e., that, deep inside, every person is motivated to behave in morally good ways. Is this belief particular to individuals with optimistic beliefs or people from Western cultures, or does it reflect a widely held cognitive bias in how people understand the self? To address this question, we tested the good true self theory against two potential boundary conditions that are known to elicit different beliefs about the self as a whole. Study 1 tested whether *individual differences* in misanthropy — the tendency to view humans negatively — predict beliefs about the good true self in an American sample. The results indicate a consistent belief in a good true self, even among individuals who have an explicitly pessimistic view of others. Study 2 compared true self-attributions across *cultural groups*, by comparing samples from an independent country (USA) and a diverse set of interdependent countries (Russia, Singapore, and Colombia). Results indicated that the direction and magnitude of the effect are comparable across all groups we tested. The belief in a good true self appears robust across groups varying in cultural orientation or misanthropy, suggesting a consistent psychological tendency to view the true self as morally good.

Keywords: concepts, social cognition, moral reasoning, true self, culture, misanthropy

Consistent Belief in a Good True Self in Misanthropes and Three Interdependent Cultures

People sometimes explain their own and others' behaviors by appealing to a deeper, more essential conception of the self. One consistent finding from studies of this concept is that people tend to think of the true self as something that is morally good (De Freitas & Cikara, 2017; Newman, Bloom, & Knobe, 2014). We refer to this tendency as the good true self bias.

How should we interpret this belief in a good true self? To date, research on the true self has not assessed the extent to which individual differences in optimism toward humans plays a role in these beliefs, and the studies have also been conducted exclusively with adults in the United States. Thus, it is possible that a positive view of the true self, and perhaps even the very notion of a true self, is specific to individuals with highly optimistic views of others, and/or the middle-class American culture. For instance, it is possible that these findings reveal some semantic fact about how people in the U.S. use the term 'true self', or more specifically, the tendency of middle-class white American college students to focus on the self as an individuated entity, and to have a generally more optimistic outlook than other sub-populations. Alternatively, it may be that the true self studies reveal something more fundamental about how people in general tend to conceive of the self. The present studies investigated these alternative possibilities, exploring whether beliefs in a good true self arise in two important comparison samples: individuals who vary in how pessimistically they view others, and cultures that vary in how the self is construed.

The Good True Self

People often explain behavior by appealing to the notion of a 'real', or 'genuine', or 'true' self. This concept also appears to have an impact in people's lives.

For example, beliefs about the true self have been shown to influence attributions about behavior (Johnson & Boyd, 1995; Johnson, Robinson, & Mitchell, 2004; Landau et al., 2011; Newman, Bloom & Knobe, 2014b; Newman, De Freitas, & Knobe, 2015; Sripada, 2010), assessments of others' lives (Newman, Lockhart, & Keil, 2010), beliefs about the meaning of life (Schlegel, Hicks, Arndt, & King, 2009; Schlegel, Hicks, King, & Arndt, 2011), decision-making (Schlegel, Hicks, Davis, Hirsch, & Smith, 2013), and general measures of well-being (Kernis & Goldman, 2004; Schimel, Arndt, Pyszczynski, & Greenberg, 2001).

One consistent finding from this research is that people tend to attribute to the true self the traits that they themselves regard as good (Bench et al., 2015; Haslam, Bastian, Bissett, 2004; Molouki & Bartels, in press; Newman, Bloom, & Knobe; 2014). More specifically, a host of studies suggest that people are especially inclined to equate an agent's true self with those aspects of the agent they regard as *morally* good (Bench, Schlegel, Davis, & Vess, 2015; Christy et al., 2016; De Freitas & Cikara, 2017; De Freitas, Tobia, Newman, & Knobe, 2016; Newman, Bloom, & Knobe, 2014b).

One way to explore people's intuitions about the self is to present participants with vignettes in which an agent loses some of her traits and then ask whether that agent's identity has been fundamentally altered. Studies using this technique find that people's conception of the essence of the self is tied more to moral traits than to other mental faculties, including personality, memory, perception, and preferences (Prinz & Nichols, in press; Strohminger & Nichols, 2014), and is tied more to morally good traits than to morally bad ones (Tobia, 2016).

Another method is to directly ask participants about the true self. Existing studies show that if an agent changes from bad behavior to good behavior,

participants are more likely to report that the behavioral change reflects the emergence of the agent's true self, compared to if the agent undergoes an analogous change from good to bad (Newman, Bloom, & Knobe, 2014b). Notably, these effects depend on the moral values of the participants themselves. For instance, in one study participants were told about an agent who believed that homosexuality was wrong, yet had the desire to sleep with other men. Liberals were more likely to attribute the homosexual desire to the true self, whereas conservatives were more likely to attribute the anti-homosexual belief (Newman, Bloom, & Knobe, 2014b; De Freitas, Tobia, Newman, & Knobe, 2016). Results like this one suggest that participants tend to attribute to other people's true selves the traits that they themselves believe to be good.

Testing the Resilience of Belief in a Good True Self: Two Boundary Cases

To reach a better understanding of the cognitive processes underlying true self attributions, it may be helpful to examine the pattern of people's intuitions about the true self across two dimensions of difference, namely, individual differences in personality and cross-cultural differences. Existing work has found that other kinds of judgments about the self vary systematically across both of these dimensions. The present studies ask whether judgments about the true self show this same pattern or whether the tendency to see the true self as morally good arises even among participants who differ with regard to other kinds of judgments about the self.

Individual Variation in Beliefs about the True Self?

Existing work on other kinds of judgments about the self has systematically explored the relationship between judgments about the self and individual difference variables. Of direct relevance to the present topic, individuals can vary in how positively they view other people or human nature in general (Hanson, 1975;

Rosenberg, 1956). One might therefore ask whether people who are generally higher in misanthropy also have more negative views of the true self.

A common measure that has been used to detect individual differences in misanthropy is the 'faith in people' scale, which asks people to indicate the extent to which they agree with statements such as, "You cannot be too careful in your dealing with people" and "If you do not watch yourself, people will take advantage of you" (Rosenberg, 1957). Such items seem to assess negative views of humanity and the belief that, in general, humans are untrustworthy, unfair, and unhelpful. Furthermore, misanthropic traits are also predictive of people's pessimistic beliefs toward other issues, such as politics (Rosenberg, 1956) and international affairs (Rosenberg, 1957).

One might reasonably suspect that previous studies on the true self have been skewed toward a baseline *optimistic* view, but that if we were to detect those individuals who were most misanthropic, then we would find that these individuals actually have a more negative view of the true self than does the average participant.

Cultural Variation in Beliefs about the True Self?

Existing work also shows that U.S. Americans differ from people of many other cultures in their conception of the self. U.S. Americans tend to be relatively more independent in their social orientation – i.e. they are more inclined than members of interdependent cultures to focus on private, inner attributes that make a person seem unique (e.g., Geertz, 1975; Grossmann & Na, 2014; Heine, 2001; Hofstede, 2001; Markus & Kitayama, 1991; Varnum et al. 2010), and to view emotions as reflecting the inner self of the person, i.e., originating from within (Uchida, Townsend, Markus, & Bergsieker, 2009). Conversely, many other cultures (including various societies from East Asia, Eastern Europe and Latin America) embrace a more interdependent social orientation than US Americans – i.e., they

conceive of the self as interdependently connected within a web of social relationships (Cousins, 1989; De Vos, 1985; Grossmann & Varnum, 2011; Kanagawa, Cross, & Markus, 2001; Markus & Kitayama, 1991; Morris & Peng, 1994; Rhee, Uleman, Lee, & Roman, 1995; Triandis, 1995), and view emotions and actions as originating through interactions between people and their environments (Greenfield, 2009; Kashima, Siegal, Tanaka, & Kashima, 1992; Uchida et al., 2009).

Relatedly, American individuals are also more inclined to optimistically enhance their self-esteem (e.g. Baumeister, Tice, & Hutton, 1989; Crary, 1966; Dunning, Meyerowitz, & Holzberg, 1989; Grossmann et al., 2014; Zuckerman, 1979; Taylor & Brown, 1988), and to construe situations in manners that make them less threatening to their personal self-image (Chang, 1996; Sherman & Cohen, 2006). Conversely, individuals from certain other cultures tend to view the world in a less positive light (Grossmann, Huynh, & Ellsworth, 2016), including engaging in moderately positive or even critical self-reflection, since they are motivated to self-improve as a means to pursue goals associated with interdependence, e.g., fulfilling their designated role within a given social network (Grossmann, Ellsworth, & Hong, 2012; Grossmann & Kross, 2010; Heine, Lehman, Markus, & Kitayama, 1999).[1]

In short, the tendency of Western individuals both to focus on the self as an individuated entity and to view the self more optimistically relative to other cultures

[1] The cultural dimension of independence-interdependence (Markus & Kitayama, 1991) has been referred to as ego- vs. socio-centrism (Shweder & Sullivan, 1993), individualism-collectivism (Hofstede, 1980; Triandis, 1989), and Gesellschaft vs. Gemeinschaft (Greenfield, 2009; Tönnies, 1887). Despite subtle distinctions between these constructs, on the cultural level of analysis, we view them as generally overlapping (Grossmann & Na, 2014; Kitayama, Park, Sevincer, Karasawa, & Uskul, 2009; Na et al., 2010; Varnum, Grossmann, Kitayama, & Nisbett, 2010).

may be associated with the kinds of behaviors that have been observed in true self studies. It may be that these same studies administered to members of interdependent cultures would show a much more reduced tendency to attribute positive behaviors rather than negative behaviors to the true self, either because the notion of a true self is not considered relevant in the first place, and/or because there is a counter-tendency to not attribute positive traits to the self.

Universal Beliefs about the True Self?

At the same time, there is also reason to suspect that belief in a good true self may be more consistent across these potential boundary cases. In particular, although existing studies find that both personality and culture predict certain kinds of judgments involving the self, there is reason to suspect that judgments about the true self are not the product of the same cognitive processes that generate those other judgments. Instead, true self judgments may be generated by a distinct type of cognitive process that is more robust across personality and cultural differences.

More specifically, true self judgments may be the product of people's *psychological essentialism*. Psychological essentialism is the tendency to conceive of entities as having a deeper essence that is not readily observed (Ahn et al., 2001; Bloom, 2004; 2010; Gelman, 2003; Dar-Nimrod & Heine, 2011; Keil, 1989; Medin & Ortony, 1989; Xu, & Rhemtulla, 2005). Existing work provides evidence that psychological essentialism is a surprisingly robust tendency. It emerges early in development (Gelman, 2003; Keil, 1989; Newman, Herrmann, Wynn, & Keil, 2007) and arises in strikingly similar forms across cultures (Atran, 1993, 1998; Brown, 1991; Gil-White, 2001a, 2001b; Hirschfeld, 1998; Sousa, Atran, & Medin, 2002; Sperber, 1996). One hypothesis would be that people's belief in a true self is simply a manifestation of this more general psychological tendency. That is, just as people

posit an unobservable essence for a variety of other entities, they may posit an unobservable essence of the self, yielding the notion of a 'true self.' If this does indeed turn out to be the case, there would be reason to expect true self beliefs to be just as robust as other manifestations of essentialism.

Results from recent studies do provide evidence that people's true self beliefs may be an instance of a broader essentializing tendency. True self beliefs show hallmarks signatures of psychological essentialism, including immutability consistency, and inherence (e.g., Christy, Schlegel, & Cimpian, 2016; Haslam, Bastian, & Bissett, 2004). Further, the more that a personal characteristic is believed to be a part of the true self, the more it is viewed as immutable, discrete, and inherent, and the higher it is rated in these features compared to other self concepts (e.g. the "everyday" or superficial self; Christy et al., 2016).

Moreover, existing work suggests that the tendency to regard the true self as good may not reflect something unique to people's understanding of the self but may instead be a manifestation of a more general fact about psychological essentialism. Specifically, participants not only believe in a good true self for human beings, but also exhibit a similar tendency to ascribe normatively positive traits to entities such as science papers, nations, bands, and universities (De Freitas et al., 2016). For example, in one experiment, participants were told about a nation in which some regions had morally good laws while others had morally bad laws. Participants tended to conclude that the morally good laws reflected the true essence of the nation in a way that the morally bad laws did not (De Freitas et al., 2016). A similar effect is found when people reason about certain artifacts, like artwork, and even certain abstract concepts, like poetry (Barsalou, 1985; Knobe, Prasada, & Newman, 2013). Results like these

suggest that belief in a good true self may be just one instance of a more general essentializing bias that is applied to individual identities of a variety of types.

In sum, there is at least some reason to suspect that belief in a good true self is not the result of people's idiosyncratic understandings of the self in particular but is instead the product of a more general fact about people's cognition. If this latter hypothesis turns out to be correct, one might expect belief in a good true self to be robust across differences in both personality and culture.

The Current Studies

To explore the generalizability of beliefs in a good true self, we test two key boundary conditions that may influence how the self is viewed. Specifically, Study 1 investigates whether individual differences in misanthropy (the general dislike of people) moderate the impact of moral status on true self attributions. We addressed this possibility using both a direct measure in which we explicitly asked about the true self (Study 1a) and an indirect measure in which we asked about a judgment that is known to depend on intuitions about the true self (Study 1b). Study 2 asks whether good true self beliefs differ between the U.S., an independent culture in which beliefs about the good true self have been studied before, and three interdependent cultures, Colombia, Singapore, and Russia.

It is important to emphasize that the different accounts of people's true self judgments make specific different predictions. If the true self is consistent across individual differences in pessimism about humans and also across cultures, then the extent to which people attribute moral improvements vs. deteriorations to the true self should not differ among these individual differences or cultures. If the true self is not consistent, however, then variation across individuals or cultures should manifest in different relative attributions of improvements vs. deteriorations to the true self. For

example, the attributions of improvements to the true self may be reduced relative to attributions of deteriorations, or in a more dramatic outcome the attributions of improvements vs. deteriorations could even be equaled or reversed. This could be either because the true self is simply not considered a relevant cause of behavior, or because certain individuals or cultures do not tend to attribute positive traits to the self in general.

Thus, we reasoned that if people hold the same belief in a good true self across these various cultures and individual differences, this would provide evidence for the claim that existing findings on people's tendency to see the true self as morally good actually are revealing something fundamental about the cognitive processes people use to make sense of the self.

Study 1: True Self Beliefs Among Misanthropes

Study 1 tested whether belief in a good true self is equally present among individuals who have an explicitly misanthropic, rather than optimistic, view of others, since these individuals might be less likely to show a biased attribution of good rather than bad traits to the true self. To this end, we conducted two studies — one using an explicit measure of the true self (Study 1a) and one using a more indirect measure (Study 1b). We reasoned that any potential influence of misanthropy on true self beliefs should be observable within a single culture, by directly measuring individual variation in misanthropy and then testing whether variation in misanthropy predicts true self beliefs.

Study 1A: Explicit Measure

Method. 280 U.S. American participants (M_{age} = 31 years, 104 female) were recruited from the online labor crowdsourcing platform, Amazon Mechanical Turk (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Goodman,

Cryder, & Cheema, 2012; Ipeirotis, 2010; Paolacci, Chandler, & Ipeirotis, 2010). The study design was inspired by Newman et al. (2014b), who used 130 participants in their first study; since the current study included an additional dependent variable presented between-subjects, we aimed for double this sample size.

Participants were presented with 6 of the vignettes from that study which described a behavioral change from morally good to morally bad, or from morally bad to morally good. Matched pairs were presented between participants, such that each participant saw either 6 morally good changes or 6 morally bad changes. For each scenario, participants were told: "Imagine an individual named [agent's name]. [Agent's name] is different from you in almost every way—he has a different occupation and prefers different things than you." Participants were then told that this person used to engage in Behavior X but now engages in Behavior Y, with the direction of change (X to Y, or Y or X) counterbalanced between participants

After each scenario, participants received either a *prediction question* that asked how much they agreed that the agent would revert to his previous behavior, e.g., "It is likely that Omar will treat ethnic minorities with respect [mistreat ethnic minorities] again." (1 = not at all, 9 = very much so) or a *true self question,* e.g., "Now that Omar treats minorities with respect [mistreats minorities], to what extent is he being true to the deepest, most essential aspects of his being?" (1 = not at all, 9 = very much so).

We expected that misanthropes would be more likely to agree that the agent would revert to moral badness, since this was a pessimistic prediction that should fall out of their pessimistic outlook toward humans in general. Therefore, the study employed a 2 (valence: good, bad) by 2 (question: true self, prediction) between-subject design. Finally, all participants completed the 'faith in people scale'

(Rosenberg, 1957), which measures individual differences in misanthropy by asking participants how much they agree with 5 different statements about humans (1 = completely disagree, 7 = completely agree; see Supplementary Materials for all the questions).

Results. We excluded 18 participants for incorrectly answering an attention check at the beginning of the study (see Supplementary Materials for exact wording). We first report the results for the prediction question, in order to confirm that our individual differences measure of misanthropy is in fact predictive of the kinds of judgments people make. We then test whether or not, despite this influence, individual differences in misanthropy have any influence on beliefs about the true self.

Prediction question. We conducted a regression analysis with valence, misanthropy, and the interaction as factors. This analysis found a main effect of valence, $\beta = .39$, $p < .001$, a main effect of misanthropy, $\beta = .21$, $p = .044$, and a significant interaction whereby misanthropy moderated the effect of valence on prediction judgments, $\beta = .36$, $p = .001$. Higher misanthropy scores predicted higher agreement that the agent would revert to a previous morally bad behavior, $\beta = .57$, $p = .001$, but not to a previous morally good behavior, $p = .241$. These results confirm that those who score highly on a measure of misanthropy are more likely to make negative predictions about others.

True Self Question. A regression analysis with valence, misanthropy and the interaction as factors found a main effect of valence, $\beta = .90$, $p < .001$, and, strikingly, no main effect of misanthropy nor an interaction, (*ps* = .413 and .875). In other words, even though individual differences in misanthropy impacted predictions, they had no effect on the tendency to say that moral improvements are caused by the true self

more so than moral deteriorations. As an even stronger test of this possibility, we then focused on just those participants who had the highest misanthropy scores (>5/7 on the scale). Remarkably, even these participants showed the good true self bias (moral improvement: M = 5.92, SD = 1.53 vs. moral deterioration: M = 4.10, SD = 2.20), $t(37) = 3.06$, $p < .01$, $\eta_p^2 = 0.20$ (see Figure 1).

In short, participants who scored high in misanthropy were more likely to predict that the agent would revert to an immoral action, yet misanthropy scores did not predict how likely participants were to show the pattern characteristic of belief in a good true self. Remarkably, even those who have a negative view of others nonetheless continue to believe that the *essence* of a person is good.

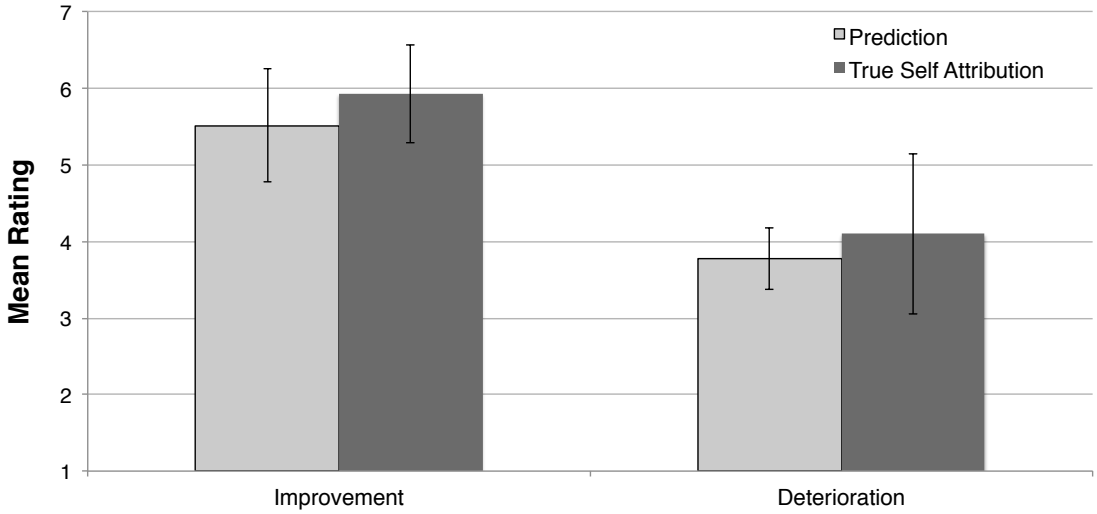

Figure 1. Mean predictions (that the agent would revert to their previous behavior despite their recent improvement or deterioration) and true self judgments for the subset of extreme misanthropes in Study 1A. Error bars represent Means ± 95% CIs.

Study 1B: Indirect Measure

Study 1B tested the robustness of the effect found in Study 1A using a more indirect measure. Specifically, instead of being asked directly about the true self,

participants were asked about whether an agent showed ‘weakness of will.’ Consider an agent whose reasoning tells her to do something morally good (e.g., to stop stealing), but whose emotions are drawing her to do something morally bad (e.g., to steal). In such a case, if the agent acts on her emotions and thereby goes against her reasoning, people tend to say that she has demonstrated weakness of will (May & Holton, 2012; Sousa & Mauro, 2015). But why do people privilege reasoning over emotion in this way? What is it about going against reasoning in particular that makes people attribute weakness of will? Existing research suggests that this intuition arises in part because people think that in this specific type of case the agent's reasoning constitutes her true self. When one switches to cases in which people think that the agent's true self is reflected in her emotions rather than in her reasoning, people no longer show this same weakness of will intuition (Newman et al., 2015). For this reason, the patterns of people's intuitions about weakness of will can give us an indirect measure of their intuitions about the true self.

Among American participants, there is a general tendency to identify the agent's true self with the part of her mind that is drawing her to be morally good. Thus, American participants show a greater tendency to attribute weakness of will when the agent goes against the part of her mind drawing her to be morally good than when she goes against the part of her mind drawing her to be morally bad (May & Holton, 2012; Sousa & Mauro, 2015), and this tendency is mediated by judgments about the true self (Newman et al., 2015).

We can now use this tendency to ask whether misanthropes have a different understanding of the true self. If misanthropes view the true self negatively, they should show a different pattern of weakness of will intuitions. In particular, if an agent is drawn by her emotions to do something morally good, they should be more

likely to say that an agent has shown weakness of will, since, to them, this action should seem to diverge from what they see as the agent's *bad* true self. In contrast, if the good true self is actually a more stable way in which people, even misanthropes, understand the very notion of a self, then misanthropes should show the usual pattern, being more inclined to attribute weak will when an agent is drawn by emotion to do something bad than when she is drawn by emotion to do something good.

Method. 280 U.S. American participants (M_{age} = 31 years, 97 female) were recruited from Amazon Mechanical Turk. The study design was inspired by Newman et al. (2015), who used 139 participants in their study on weakness of will and the true self (Study 3); since the current study included an additional dependent variable presented between-subjects, we aimed for double this sample size.

Vignettes were taken from Sousa & Mauro (2015). Each participant read about either an assassin or a robber who is torn between a morally good and morally bad action, e.g. "John is a professional assassin. He has started to think about quitting this profession because he feels that it is wrong to kill another person. However, he is strongly inclined to continue with it because of the financial benefits." Then, depending on the question type, participants either:

(1) Read the rest of the vignette from Sousa & Mauro (2015), in which the agent decides to continue [or end] his current career, but the next day goes against his decision:

"John is in conflict, but after considering all aspects of the matter, he concludes that the best thing for him to do is to continue with [quit his] profession. Accordingly, he decides that the next day he will kill another person for money [look for a job that does not involve violence].

The next day, while still completely sure that the best thing for him to do is to

kill another person for money [look for a job that does not involve violence], John is swayed by the feeling that it is wrong to kill another person for money [the financial benefits]. Against what he decided, he looks for a job that does not involve violence [kills another person for his money]."

They were then asked to indicate how much they agreed that the agent showed weakness of will (1 = strongly disagree, 7 = strongly agree), e.g. "John displays weakness of will when, the next day he looks for a job that does not involve violence [kills another person for money]."

(2) Or instead of being presented with the rest of the vignette, participants were immediately asked how likely it was that the agent would engage in either the morally good or morally bad behavior (1 = extremely unlikely, 7 = extremely likely), e.g. "If the next day John had the chance to look for a job that does not involve violence [kill another person for money], how likely do you think it is that he would do it?" We predicted that misanthropes (as measured by the faith in people scale) would find morally bad actions especially likely.

Therefore, this study used a 2 (valence: morally good, morally bad) by 2 (question type: weakness of will vs. prediction) by 2 (character: assassin, robber) design, with all variables manipulated between participants.

Results. We excluded 20 participants for incorrectly answering the attention check at the beginning of the study.

Prediction Question. We conducted a regression analysis with valence, misanthropy, and the interaction as factors. This analysis found a main effect of valence, $\beta = 1.03$, $p < .001$, a main effect of misanthropy, $\beta = .32$, $p = .013$, and no interaction, $p = .880$. Despite the lack of an interaction, we still assessed the simple effects, given our prior results for Study 1A. As in Study 1A, misanthropy

significantly predicted likelihood ratings for the morally bad action, $\beta = .30$, $p = .024$, but not the morally good action, $p = .136$.

Weakness of Will Question. A regression analysis with valence, misanthropy and the interaction as factors found only a main effect of valence, $\beta = 1.70$, $p < .001$, but, strikingly, no main effect of misanthropy nor an interaction (*ps* = .850 and .993). In other words, even though individual differences in misanthropy impacted predictions, they had no effect on the good true self bias (the tendency to say that moral improvements are caused by the true self more so than moral deteriorations). As an even stronger test of this possibility, we focused on just those participants who had the highest misanthropy scores (>5/7 on the scale). Remarkably, even these participants showed the good true self bias, (moral improvement: $M = 6.31$, $SD = 1.14$ vs. moral deterioration: $M = 2.46$, $SD = 2.07$), $t(27) = 6.38$, $p < .001$, $\eta_p^2 = 0.60$ (see Figure 2).

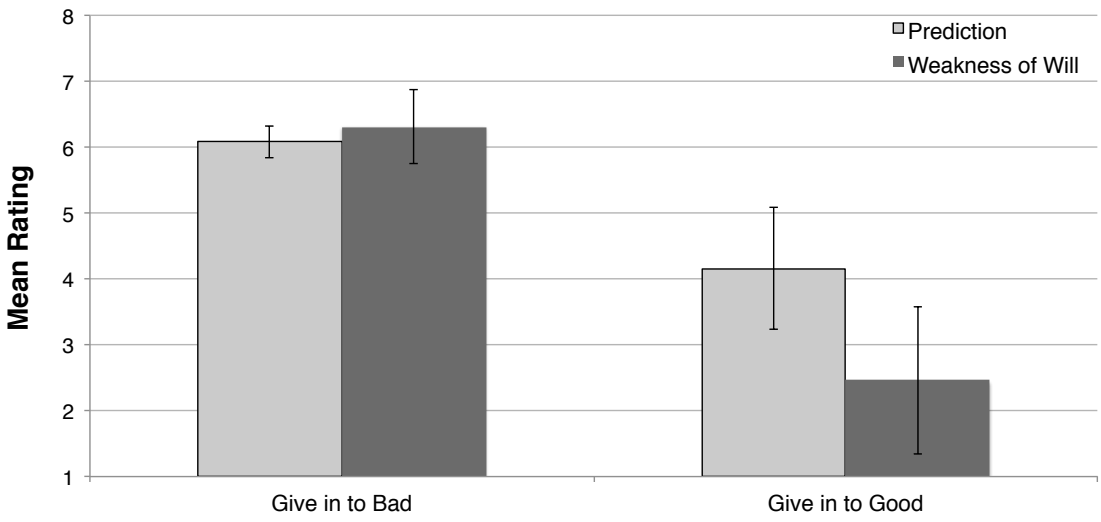

Figure 2. Mean predictions and weakness of will ratings for the subset of extreme misanthropes in Study 1B. Error bars represent Means ± 95% CIs.

Discussion

Across both explicit and indirect measures, we found that participants who scored high in misanthropy tended to predict that agents would act on their morally bad desire (although the directional nature of this effect was more convincing in Study 1A). Yet, remarkably, across both studies misanthropes showed a good true self bias. Taken together, the studies are more convincing than either study alone: the implicit study addresses a potential issue with task demands surrounding the term "true self", and nonetheless still finds an asymmetry in a judgment previously shown to be mediated by belief in a good true self. The explicit study confirms that participants are indeed willing to make these judgments about the true self when it is explicitly referred to, showing that the two effects are related.

What is surprising about these results is that the very same participants who say that most human beings are awful also appear to hold the belief that human beings are fundamentally good deep down in their true selves. This result suggests that whatever cognitive processes are at work in people's true self judgments, these processes are remarkably unaffected by individual differences in judgments about other aspects of the self.

These findings also lead us to reconsider what it actually means to be misanthropic. Specifically, these studies suggest that misanthropy may be described as a belief about how others will succumb to negative worldly ways of living that lead them astray; that they will fail to realize their inner goodness just beneath the surface. Nonetheless, when people do behave in a morally good manner, even misanthropes' first intuition is that such changes must have been caused by the true self. In some sense, therefore, misanthropy only goes skin deep. This result is also consistent with recent findings that people believe that even members of disliked outgroups

ultimately have a morally good true self (De Freitas & Cikara, 2017).

Finally, it is worth emphasizing that although the current findings show that people's true self beliefs can come apart from their predictions about future behavior, existing work suggests that true self beliefs have a strong impact in numerous other domains, including, attributions about behavior (Johnson, Robinson, & Mitchell, 2004; Newman, Bloom & Knobe, 2014b; Newman, De Freitas, & Knobe, 2015), assessments of personal character (Newman, Lockhart, & Keil, 2010), beliefs about the meaning of life (Schlegel, Hicks, Arndt, & King, 2009), satisfaction with major life decisions (Schlegel, Hicks, Davis, Hirsch, & Smith, 2013), general measures of well-being (Schimel, Arndt, Pyszczynski, & Greenberg, 2001), a variety of everyday judgments (Newman, De Freitas, & Knobe, 2015), and intergroup behavior (De Freitas & Cikara, 2017).

Study 2: True Self Beliefs Across Cultures

Study 2 turns to the question of whether the good true self bias is robust across cultural differences. One possible view would be that the effect observed among Western participants only arises because of certain idiosyncratic facts about Western culture. Specifically, numerous studies find that people from Western cultures have a strikingly independent conception of the self, whereas people from many other cultures show a more interdependent conception (for reviews, see Cross, Hardin & Gersek-Swing, 2011; Varnum et al., 2010). One hypothesis would be that the effect observed in previous studies only arises among people who share this distinctively Western conception; another would be that the effect is the product of a more fundamental fact about human cognition, which might be expected to arise even in people who have quite different conceptions. To decide between these hypotheses, we asked whether good true self beliefs differ between the U.S. and three interdependent

cultures, Colombia, Singapore, and Russia.

Although our primary interest was in the distinction between independent and interdependent conceptions of the self, we also chose these particular samples because they systematically vary in their dominant belief systems and social economy. In particular, there is a further possibility that US participants in previous true self studies were influenced by a predominantly Protestant Christian belief system or by the experience of living in a developed market economy. Looking at our country samples, Colombia is predominantly Catholic and an emerging economy, and so is similar to America religiously (47% Protestant), but not economically. In contrast, Singapore has a highly developed market economy like the US, but the majority of Singaporeans (44%) is either Buddhist (a religion that emphasizes that there is no self that is independent from the universe) or Taoist. Finally, unlike the US, Russia has a post-Soviet market economy, and the majority of Russians is either atheist or considers themselves Russian Orthodox. In short, by also ensuring that we focused on samples varying in economic systems and dominant beliefs, Study 2 served the further function of exploring whether economic systems and historically dominant belief systems play a role in the prevalence of beliefs that the true self is good.

Method

Participants. Our goal was to recruit at least 130 participants from each country, to match the sample size used by Newman, Bloom, & Knobe (2014b); however, we expected some people might have missing data or might not complete the survey carefully, so we oversampled from each population to the extent it was possible to do so in a comparable data collection time frame. Since participants were collected from multiple locations, however, some samples ended up with more participants than others (see Table 1).

Table 1. Demographics and recruitment site for each country.

Country	**N**	M_{age}	**Sex**	**Source**
USA	133	36	68 f	Amazon Mechanical Turk
Colombia	266	25	143 f	Public Universidad Nacional de Colombia, Bogotá
Singapore	170	21	134 f	Public Nanyang Technological University
Russia	143	21	92 f	6 Public universities in Tomsk, Omsk, Novosibirsk

Materials. For Russia and Columbia respectively the materials were translated into Spanish and Russian, then back-translated by the third and fourth authors to ensure accuracy, and the authors compared the original and back-translated versions to ensure semantic equivalence (Grossmann & Na, 2014).

Independence-interdependence. As a strong test of independence - interdependence, we focused on behavioral reactions, assessed via analysis of spontaneous open-ended self-descriptions, which have a number of advantages over survey-based measures of self-construal (Kitayama, 2002; Kitayama et al., 2009). Specifically, at the beginning of the survey, participants were given the twenty statements test (TST; Kuhn & McPartland, 1954), in which they complete twenty sentences beginning with the words "I am". In the case of the Russian and Colombian participants, these statements were completed in their native languages, and then translated into English by independent, hypothesis-blind coders. Next, two separate independent, hypothesis-blind, condition-blind coders coded the full set of English responses along independence-interdependence dimensions, following the exact procedure recommended by Rhee et al. (1995); all country identifiers were removed for coding purposes. Inter-coder reliability was high ($\kappa = .80$).

Good true self. The study design was taken in full from Newman et al. (2014b). Participants were presented with 12 different scenarios in randomized order, each of which described a different agent who underwent one of three types of behavioral changes: morally good to morally bad ("bad"), morally bad to morally good ("good"), or a non-moral change ("neutral") in preferences (see Supplementary Materials for all vignette materials). The neutral vignettes enabled a baseline against which we could compare people's judgments about moral improvements and deteriorations. For each scenario, participants were told: "Imagine an individual named [agent's name]. [agent's name] is different from you in almost every way—he has a different occupation and prefers different things than you." For the interdependent samples we translated the name of the agents into similar local equivalents, in order to avoid any potential intergroup biases. Participants were then told that this person used to engage in Behavior X but now engages in Behavior Y, with the direction of change (X to Y, or Y or X) counterbalanced between participants. Each participant thus saw one of two blocks consisting of four "good" vignettes, four "bad vignettes, and four "neutral vignettes". However, the corresponding matched-item pairs were always presented between participants. This produced a 3 (valence: good, bad, neutral) by 2 (block 1 vs. block 2) mixed-model design.

After each scenario, participants answered a forced-choice question about what caused the agent's behavior: e.g., "In your opinion, what aspect of Omar's personality caused him to treat ethnic minorities with respect [mistreat ethnic minorities]?" The answer options were "(a) His 'true self' (the deepest, most essential aspect of his being), (b) His 'surface self' (the things that he learned from society or others),' (c) None of the above." The third option included a space in which

participants could explain their choice. A second question asked participants about the agent's behavior in relation to his/her true self, e.g., "Now that Omar treats minorities with respect [mistreats minorities], to what extent is he being true to the deepest, most essential aspects of his being?" (1 = not at all, 9 = very much so).

Results

Some participants were excluded for incorrectly answering an attention check at the beginning of the study: 14 (USA), 62 (Colombia), 68 (Singapore), and 17 (Russia). Analyses were conducted on the remaining participants.

Independence-interdependence. We calculated interdependence vs. independence scores for each participant (% of interdependent statements – % independent statements), then compared this score across countries. A univariate ANOVA found a significant main effect of country, $F(3, 547) = 21.06$, $p < .001$, $\eta_p^2 = .104$. This main effect also holds when controlling for age, gender, education, and socio-economic status, $F(3, 525) = 17.19$, $p < .001$, $\eta_p^2 = .089$.

Replicating prior research (Grossmann & Varnum, 2011; Realo & Allik, 1999; Varnum et al., 2010), U.S. participants scored significantly lower ($M = -.06$, $SD = .49$), i.e. more independent, than each of the other countries: Colombia, $M = .16$, $SD = .46$, $t(321) = 4.03$, $p < .001$, $\eta_p^2 = .05$, Singapore, $M = .15$, $SD = .44$, $t(219) = 3.36$, $p = .001$, $\eta_p^2 = .05$, and Russia, $M = .41$, $SD = .47$, $t(243) = 7.70$, $p < .001$, $\eta_p^2 = .20$ (see Figure 1). Furthermore, Sidak-corrected post-hoc comparisons indicated that Russian participants showed a significantly greater degree of interdependence than participants from Colombia, $z = 4.79$, $p < .001$, and Singapore, $z = 4.19$, $p < .001$, whereas the samples from Colombia and Singapore did not significantly differ from each other, $z = 0.12$, *ns*.

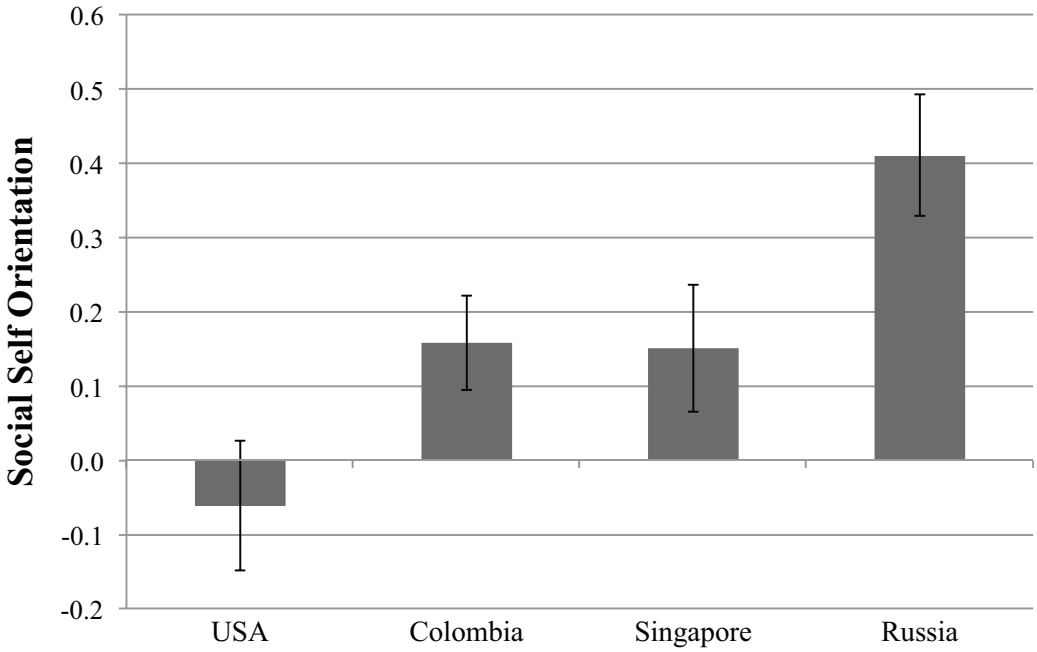

Figure 3. Social self orientation (% interdependent traits - % independent traits) for each of the country samples in Study 2, as measured by the twenty statements test. Error bars represent Means ± 95% CIs.

True self forced-choice items. In line with Newman et al. (2014b), we recoded the forced-choice item as a binary response with "true self" response as "1" and "surface self" or "other" responses as "0". These scores were then summed across the morally good, morally bad, and neutral items to produce three scores for each participant ranging from 0 (no endorsement of the true self) to 4 (endorsement of the true self for all items of that valence). We ran a 3 (valence: good, bad, neutral) by 2 (block 1 vs. block 2) mixed-model ANOVA for each of the samples. The ANOVA found a main effect of valence for all samples, [USA] $F(2, 116) = 35.98$, $p < .001$, $\eta_p^2 = .383$; [Colombia] $F(2, 201) = 40.43$, $p < .001$, $\eta_p^2 = .287$; [Singapore] $F(2, 99) = 18.26$, $p < .001$, $\eta_p^2 = .269$; [Russia] $F(2, 123) = 24.65$, $p < .001$, $\eta_p^2 = .286$. For all samples, participants were more likely to say the true self caused a change that was morally good than morally bad or neutral (see Fig. 2a).

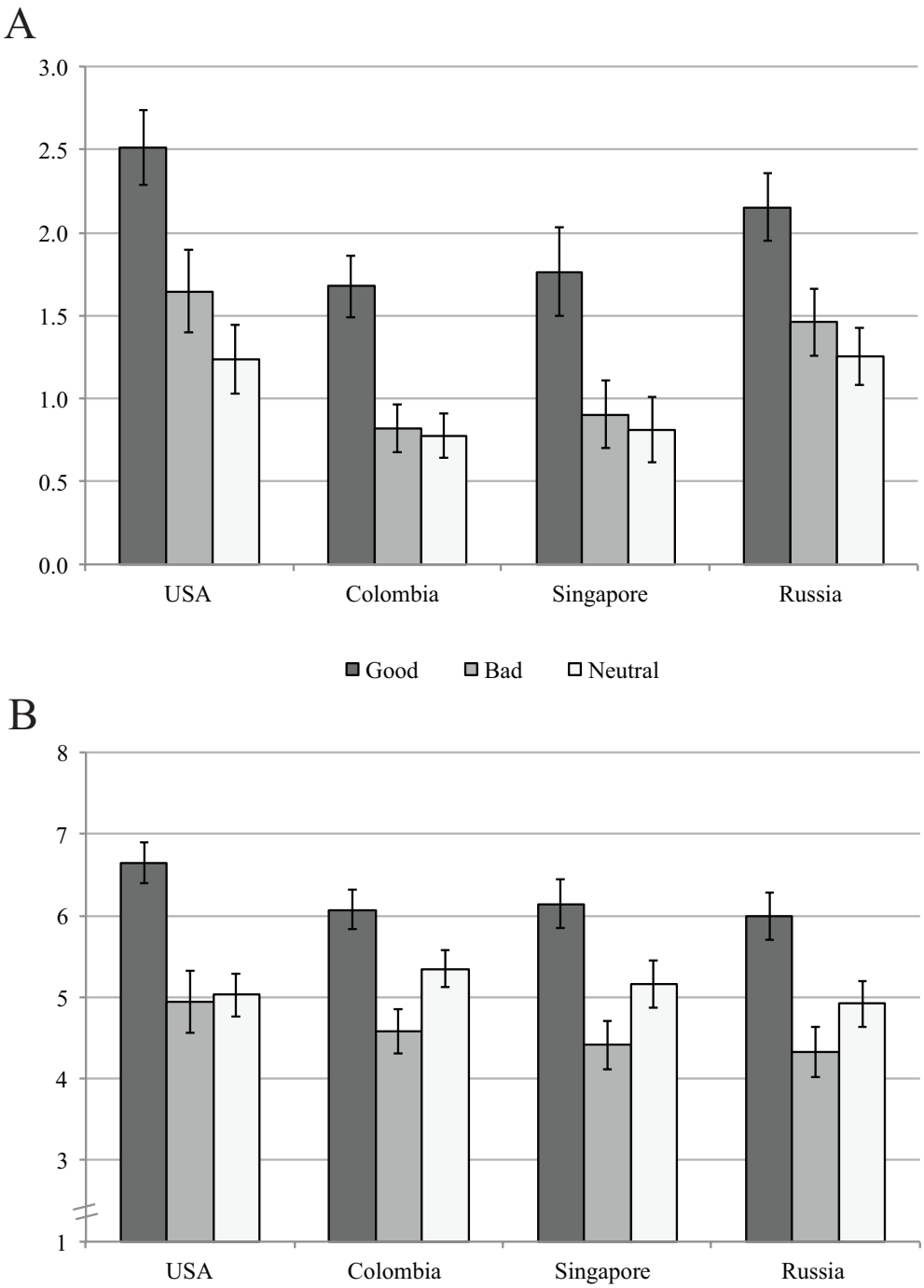

Figure 4. Study 2 results of beliefs about the true self across cultures. Panel A. Results from the forced-choice question – Belief that the change in behavior was caused by the agent's "true self". Panel B. Results from the scaled question – Agreement that the change in behavior was true to the "deepest, most essential aspects of his being". Error bars represent Means ± 95% CIs.

True self scaled items. For all samples, a 3 (valence: good, bad, neutral) by 2 (block 1 vs. block 2) mixed-model ANOVA found a main effect of valence, [USA] $F(2,116) = 55.01$, $p < .001$, $\eta_p^2 = .487$; [Colombia] $F(2, 201) = 48.11$, $p < .001$, $\eta_p^2 = .324$; [Singapore] $F(2, 99) = 31.43$, $p < .001$, $\eta_p^2 = .388$; [Russia] $F(2,123) = 40.30$, $p < .001$, $\eta_p^2 = .396$. For all samples, participants were more likely to say that the agent's current behavior reflected their true self when it was morally good than morally bad or neutral (see Fig. 2b).

Comparing the good true self bias across countries. The difference in true self judgments for the morally good and bad conditions was then calculated for both the forced-choice and scaled questions, and this difference was compared among the different country samples in order to test our main question of interest. Since we calculated the difference in this way before comparing the difference across countries, this is effectively equivalent to measuring a valence by country interaction. Remarkably, there was no significant difference among samples in the good true self bias, either in the forced-choice item or scaled rating (*ps* = .823 and .778). Therefore, despite various differences among these countries — including a measurable difference in interdependence versus independence — all participants showed the good true self bias. We did not analyze any differences in the neutral vs. bad items, since we did not have strong a priori predictions about these. Although the different questions could have led to slight differences in judgments about neutral and bad items, the main observation is that the difference between good and bad was remarkably consistent.

Finally, an inspection of the results suggested that although there was no cross-cultural difference in the degree to which the morally good condition differed from the morally bad condition, there might be a cross-cultural difference in the

morally good condition considered alone. We therefore ran an additional exploratory analysis in which we looked only at the morally good condition. For this condition, the countries differed both on the forced-choice item, $F(3, 547) = 11.99$, $p < .001$, $\eta p^2 = .062$; [USA] = ($M = 2.51$, $SD = 1.25$), [Colombia] = ($M = 1.68$, $SD = 1.37$), [Singapore] = ($M = 1.77$, $SD = 1.39$), [Russia] = ($M = 2.15$, $SD = 1.16$), and the scaled ratings, $F(3, 547) = 4.26$, $p = .005$, $\eta p^2 = .023$; [USA] = ($M = 6.65$, $SD = 1.41$), [Colombia] = ($M = 6.07$, $SD = 1.73$), [Singapore] = ($M = 6.14$, $SD = 1.56$), [Russia] = ($M = 5.99$, $SD = 1.65$). Specifically, the US scored the highest on both measures.

Discussion

We tested whether there is a difference in the good true self bias across samples spanning the U.S., Colombia, Russia, and Singapore. Prior research found a greater degree of interdependent self-construal in Colombia, Russia, and Singapore, as compared to the U.S. (Cousins, 1989; De Vos, 1985; Grossmann & Varnum, 2011; Kanagawa, Cross, & Markus, 2001; Markus & Kitayama, 1991; Rhee, Uleman, Lee, & Roman, 1995; Triandis, 1995). We replicated that effect, but we also found that participants from all samples behaved in a manner consistent with belief in a good true self.

Our results indicated a noticeable cultural difference in the degree to which people attributed morally good changes to the true self. However, participants in all four countries were more inclined to attribute changes to the true self when those changes made the agent more morally good than when they made the agent more morally bad. Moreover, there was no significant cross-cultural difference in the magnitude of this effect.

Our samples likely differed not only in their social orientation towards

interdependence versus independence, but also in genetic lineage, dominant beliefs, language, geographic mobility, political system, and history of industrialization and democratization. Each sample was also idiosyncratic in various other ways that could have influenced our measurements. As one example, previous theoretical and empirical work has found that Russian participants are relatively more inclined toward cynicism and anti-social punishment (e.g., Gächter & Herrmann, 2015; Hofstede, 2001), suggesting that our Russian sample may have constituted an especially strong test of our hypothesis. Overall, then, the results showed that true self beliefs were strikingly robust across cultural differences.

Notably, these cultures have not yet been systematically compared on the open-ended TST measure employed here, which avoids the biases of self-report scales of individualism/collectivism or independent/interdependent self-construal (Heine et al., 2002). Strikingly, despite differences in social orientation and the inclusion of countries that differ in economic and/or socio-historically dominant belief systems, the present results point to an area of apparent cultural similarity.

General Discussion

At the beginning of this paper we offered two possible interpretations of previous research demonstrating beliefs in a good true self. One was that these findings do not reveal anything particularly general about people's conception of the self, but rather tell us something that is specific to people who have certain personality traits or who come from Western cultures in particular. The second possibility was that these findings point to something more fundamental about how people (in general) conceive of the self. To distinguish these two possibilities, we tested whether the belief in a good true self is robust across individual differences in misanthropy and across four notably different cultures. Strikingly, we found that

participants at all levels of misanthropy and from all four cultures behaved in a manner consistent with the hypothesis that they believe in a good true self.

Implications for research on Cross-Cultural and individual differences

Previous research has explored the ways in which the various kinds of judgments about the self vary across cultures and across individual differences in personality. The present studies suggest that judgments that are specifically about the *true self* do not show the same patterns observed in this previous research.

Existing research has shown that Eastern and Western cultures differ in the degree to which they conceptualize the self as independent versus interdependent (e.g., Chiu, Morris, Hong, & Menon, 2000; Cross et al., 2011; Heine & Lehman, 1997, 1999; Hofstede, 1980; Hong, Morris, Chiu, & Benet-Martinez, 2000; Markus & Kitayama, 1991; Schwartz, 1994; Singelis, 1994; Triandis, 1995; Varnum et al., 2010). The present Study 2 replicated previously established cultural differences in independence-interdependence, showing that U.S. Americans are relatively more independent as compared to Colombians, Russians, and Singaporeans. When it came to judgments about the true self, however, we found a surprising degree of cross-cultural similarity. Across all four of the cultures tested, we found evidence that participants shared an intuition that, deep down, human beings have a true self that is morally good.

Just as differences in social orientation have been used to generate many hypotheses about what psychological factors may differ among cultures (e.g., Grossmann & Kross, 2010; Hamamura et al., 2009; Heine, Lehman, Markus & Kitayama, 1999; Kanagawa et al., 2001; Maddux et al., 2010; Rozin, Lowery, Imada, & Haidt, 1999), the consistency of belief in a good true self may be used to generate predictions about factors that are likely to be consistent across cultures. In particular,

recall that previous work with North Americans found that intuitions about the true self influence people's evaluations of happiness, blameworthiness, strength of will, and valuing (Newman et al., 2015), and that thinking about the true self reduces intergroup bias (De Freitas & Cikara, 2017). One implication of the current cross-cultural results is that all of these same effects should apply across the various cultures we tested. This is a promising avenue for future work. Moreover, we hope that the current theory will be tested in more countries still, including small-scale societies such as hunter-gatherers, pastoralists, and horticulturalists.

A similar point applies to the study of individual differences in personality. Existing work consistently finds that certain aspects of people's intuitions about the self can be predicted by personality variables (e.g. Peterson et al., 1982). Studies 1a and 1b provided further support for this view, showing that misanthropic participants make different predictions about how an agent is likely to behave in the future. However, these studies also found that intuitions about the true self were surprisingly stable across differences in misanthropy. Even the most misanthropic participants tended to think that the true self is morally good. Here again, it should be emphasized that the present studies explored just one personality variable, and future work could examine other dimensions of individual difference. Still, it is striking that judgments about the true self differ in this way from other types of judgments about the self.

In short, we find evidence that people's judgments about the true self are different in fundamental respects from other kinds of judgments about the self. Previous work shows that other kinds of judgments about the self vary systematically across culture and personality. We replicate those existing findings but also find that judgments about the true self are surprisingly stable across those same variables.

Essentialism as an Explanation of True Self Beliefs

A question now arises as to why judgments about the true self appear to be so remarkably stable across culture and personality. This stability is especially surprising in light of the well-documented fact that other sorts of judgments about the self do vary across cultures and across personality variables. How is the robustness of true self judgments across these variables to be explained?

One natural approach would be to suggest that people's judgments about the true self are generated by a type of cognition that differs from the one that generates these other types of judgments about the self. Within this broad approach, the hypothesis that has been most thoroughly explored thus far involves an appeal to existing work on *psychological essentialism* (Ahn et al., 2001; Bloom, 2004; 2010; Gelman, 2003; Dar-Nimrod & Heine, 2011; Keil, 1989; Medin & Ortony, 1989; Xu, & Rhemtulla, 2005). Quite apart from anything about how people understand the self in particular, people show a general tendency to understand various entities by attributing essences (the essence of a nation, the essence of a scientific paper, etc.). Existing studies suggest that this capacity to attribute essences arises early in development (Gelman, 2003; Keil, 1989; Newman, Herrmann, Wynn, & Keil, 2007) and is strikingly similar across different cultures (Atran, 1993, 1998; Brown, 1991; Gil-White, 2001a, 2001b; Hirschfeld, 1998; Sousa, Atran, & Medin, 2002; Sperber, 1996). One hypothesis would be that judgments about the true self involve the application of this general capacity to the understanding of the self (i.e., that people conceptualize the true self as the essence of the self). If this hypothesis turns out to be correct, it would give us some explanation for the surprising robustness we find in people's true self judgments.

One task for future work will be to fill out this hypothesis in more detail and explain why people's general capacity for psychological essentialism would lead to

the specific patterns we find in their true self judgments. Most importantly, such work would need to provide answers to two questions. First, why would people associate the essence of the self with specifically moral aspects of the self? Second, why would people regard the essence of the self as good?

Regarding the first question, existing work has asked whether people associate the essence of the self with specifically moral traits. Instead of asking participants directly about the true self, such work sometimes proceeds by giving participants a case in which an agent has lost certain traits and asking about the degree to which the agent's identity has changed (Chen, Urminsky, & Bartels, 2016; Heiphetz, Strohminger, & Young, 2016; Phillips et al., 2017; Prinz & Nichols, in press; Strohminger & Nichols, 2014). Studies using these methods indicate that people see moral traits as especially essential (Strohminger & Nichols, 2014). It is not yet entirely clear why people show this pattern of judgments, but some studies suggest that the pattern arises because moral traits are seen as especially important for social interaction (Heiphetz, Strohminger, & Young, 2016).

Regarding the second question, existing work has also asked whether people show a general tendency to see the essences of entities as good. When people are asked about nonhuman entities such as nations or scientific papers, they tend to say that the essences of these entities are good (De Freitas et al., 2016). In other words, the pattern observed for judgments about the true self seems also to arise for judgments about individual nonhuman entities. Here again, a difficult question arises about how to explain this broader pattern. This answer to this question is not yet known, but one possible view would be that the effect is ultimately to be explained in terms of people's teleological thinking. Existing research suggests that people show a striking tendency to explain numerous phenomena teleologically, i.e., by ascribing

deeper purposes to them (Kelemen & Rosset, 2009; Rose & Schaffer, 2015). One possibility would be that people tend to think of the essence of an entity in terms of its telos and that they tend to think that the telos of entities is in some way to be good (more specifically, that people think the telos of human beings is to be morally good).

One important task for future work in this area will be to explore the difference between attributions of essence to individuals and attributions of essence to categories. People do not always see categories as having good essences, as in the well-documented tendencies in the way people think of the essences of out-group categories, e.g. 'the essence of Arab immigrants' (Haslam, Rothschild, & Ernst, 2000; Rozin, & Royzman, 2001). Yet, even when thinking about out-groups, people show a tendency to attribute a good essence to the individual members, e.g. 'the essence of Alhadin', or 'the essence of Jafri' (De Freitas & Cikara, 2017). It may be that people have a fundamentally different way of attributing a telos to out-group categories vs. individual human beings, and that this difference explains the normative directions of the beliefs. This possibility remains an important question for future work.

In short, research suggests that the patterns we observe in people's intuitions about the true self reflect more general patterns in people's capacity to attribute essences. This finding provides one natural explanation for the robustness of people's true self intuitions across cultures and personality variables. Research has not yet answered the further question as to why people's capacity to attribute essences shows these patterns, but this question may be addressed in future work.

Conclusion

We found evidence that the belief in a morally good true self is strikingly consistent across individual and cultural differences. The concept is different from

other concepts of the self that have been studied in personality psychology and cross-cultural work, and may be a manifestation of psychological essentialism. These results provide first support for the hypothesis that belief in a morally good true self is a fundamental aspect of people's commonsense understanding of the self, and may thus have reliable and widespread consequences for other aspects of cognition.

Acknowledgments

We thank Olga Krylova and Juan De Brigard for translating, Anna Leontieva and Diana Carolina Rodriguez for both translating and helping us recruit participants, and Judith E. Fan for helpful thoughts on the manuscript.

References

Ahn, W., Kalish, C., Gelman, S. A., Medin, D. L., Luhmann, C., Atran, S., Coley, J. D., Shafto, P. (2001). Why essences are essential in the psychology of concepts. *Cognition*, *82*, 59–69.

Atran, S. (1993). *Cognitive foundations of natural history: Towards an anthropology of science*. Cambridge University Press.

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, *21*, 547–569.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 629–654.

Baumeister, R. F., Tice, D. M., & Hutton, D. G. (1989). Self-presentational motivations and personality differences in self-esteem. *Journal of Personality, 57*, 547–579.

Bench, S. W., Schlegel, R. J., Davis, W. E., & Vess, M. (2015). Thinking about change in the self and others: The role of self-discovery metaphors and the true self. *Social Cognition*, *33,* 169–185.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical turk. *Political Analysis, 20*, 351–368.

Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human.* New York: Basic Books.

Bloom, P. (2010). *How pleasure works*. New Work, NY: Basic Books.

Brown, D. E. (1991). *Human universals* (p. 118). New York: McGraw-Hill.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A

new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6*, 3–5.

Chang, E. C. (1996). Cultural differences in optimism, pessimism, and coping: Predictors of subsequent adjustment in Asian American and Caucasian American college students. *Journal of Counseling Psychology*, *43*, 113–123.

Chen, S. Y., Urminsky, O., & Bartels, D. M. (in press). Beliefs about the causal structure of the self-concept determine which changes disrupt personal identity. *Psychological Science*, *27,* 1398–1406.

Chiu, C., Morris, M., Hong, Y., & Menon, T. (2000). Motivated cultural cognition: The impact of implicit cultural theories on dispositional attribution varies as a function of need for closure. *Journal of Personality and Social Psychology*, *78*, 247–259.

Christy, A., Schlegel, R., & Cimpian, A. (2016). *Why do people believe in true selves? The role of psychological essentialism.* Manuscript under review.

Cimpian, A., & Markman, E. M. (2009). Information learned from generic language becomes central to children's biological concepts: Evidence from their open-ended explanations. *Cognition, 113*, 14–25.

Cousins, S.D. (1989). Culture and selfhood in Japan and the U.S. *Journal of Personality and Social Psychology, 56,* 124–131.

Crary, W. G. (1966). Reactions to incongruent self-experiences. *Journal of Consulting Psychology, 30*, 246–252.

Cross, S. E., Hardin, E. E., & Gercek-Swing, B. (2011). The What, How, Why, and Where of Self-Construal. *Personality and Social Psychology Review*, *15*, 142–179.

Dar-Nimrod, I., & Heine, S. J. (2011). Genetic essentialism: On the deceptive

determinism of DNA. *Psychological Bulletin, 137,* 800–818.

De Freitas J., & Cikara, M. (2017). *Deep down my enemy is good: Thinking about the true self reduces intergroup bias*. Manuscript under review.

De Freitas, J., Tobia, K., Newman, G. E., & Knobe, J. (2016). Normative judgments and individual essence. *Cognitive Science.* Advanced online publication. DOI: 10.1111/cogs.12364

De Vos, G. A. (1985). Dimensions of the self in Japanese culture. In A. J. Marsella, G. De Vos, and F. L. K. Hsu (Eds.), *Culture and self: Asian and Western perspectives* (pp. 141–182). New York: Tavistock.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definition in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*, 1082–1090.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. California: Stanford University Press.

Gächter, S., and Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society, 364*, 791–806.

Geertz, C. (1975). On the nature of anthropological understanding. *American Scientist, 63***,** 47–53.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought.* Oxford, UK: Oxford University Press.

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

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development, 58*, 1532–1541.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition, 38*, 213–244.

Gil-White, F. J. (2001a). Sorting is not categorization: A critique of the claim that Brazilians have fuzzy racial categories. *Journal of Cognition and Culture*, *1*, 219–249.

Gil-White, F. J. (2001b). Are ethnic groups biological "species" to the human brain? *Current Anthropology*, *42*, 515–553.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2012). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*, 213–224.

Greenfield, P. M. (2009). Linking social change and developmental change: Shifting pathways of human development. *Developmental Psychology, 45,* 401–418.

Grossmann, I., Ellsworth, P. C., & Hong, Y. (2012). Culture, Attention, and Emotion. *Journal of Experimental Psychology: General, 141*, 31–36.

Grossmann, I., Huynh, A. C., & Ellsworth, P. C. (2016). Emotional complexity: clarifying definitions and cultural correlates. *Journal of Personality and Social Psychology*,*111,* 895–916.

Grossmann, I., Karasawa, M., Kan, C. & Kitayama, S. (2014). A cultural perspective on emotional experiences across the lifespan. *Emotion*, *14*, 679–692.

Grossmann, I., & Kross, E. (2010). The impact of culture on adaptive versus maladaptive self-reflection. *Psychological Science*, *21*, 1150–1157

Grossmann, I., & Na, J. (2014). Research in culture and psychology: past lessons and

future challenges. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*, 1–14.

Grossmann, I. & Varnum, M.E.W. (2011). Culture, social class, and cognition. *Social Psychological and Personality Science*, *2*, 81–89.

Hamamura, T., Meijer, Z., Heine, S. J., Kamaya, K., & Hori, I. (2009). Approach—Avoidance Motivation and Information Processing: A Cross-Cultural Analysis. *Personality and Social Psychology Bulletin*, *35*, 454–462.

Hanson, D. J. (1975). Dogmatism and Misanthropy. *Psychological Reports*, *36,* 670–670.

Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin*, *30*, 1661–1673.

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology*, *39*, 113–127.

Heine, S. J. (2001). Self as cultural product: An examination of East Asian and North American selves. *Journal of Personality*, *69*, 881–905.

Heine, S., & Lehman, D. (1997). Culture, dissonance, and self-affirmation. *Personality and Social Psychology Bulletin*, *23*, 389–400.

Heine, S., & Lehman, D. (1999). Culture, self-discrepancies, and self-satisfaction. *Personality and Social Psychology Bulletin*, *25*, 915–925.

Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review, 106,* 766–794.

Heiphetz, L., Strohminger, N., & Young, L. L. (2016). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive Science*. DOI: 10.1111/cogs.12354.

Hirschfeld, L. A. (1998). *Race in the making: Cognition, culture, and the child's construction of human kinds*. MIT Press.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage.

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Thousand Oaks, CA: Sage Publications.

Hong, Y., Morris, M., Chiu, C., & Benet-Martinez, V. (2000). Multicultural minds: A dynamic constructivist approach to culture and cognition. *American Psychologist, 55*, 709–720.

Inagaki, K., & Hatano, G. (2002). *Young children's naive thinking about the biological world.* New York: Psychology Press.

Ipeirotis, P. (2010, March 9). The new demographics of Mechanical Turk. Retrieved from http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html.

Johnson, J. T., & Boyd, K. R. (1995). Dispositional traits versus the content of experience: Actor/observer differences in judgments of the "authentic self." *Personality and Social Psychology Bulletin*, *21*, 375–383.

Johnson, J. T., Robinson, M. D., & Mitchell, E. B. (2004). Inferences about the authentic self: When do actions say more than mental states? *Journal of Personality and Social Psychology*, *87*, 615–630.

Kalish, C. W., & Gelman, S. A. (1992). On wooden pillows: Multiple classification and children's category-based inductions. *Child Development, 63*, 1536–1557.

Kanagawa, C., Cross, S. E., & Markus, H. R. (2001). "Who am I?": The cultural psychology of the conceptual self. *Personality and Social Psychology Bulletin,*

27, 90–103.

Kashima, Y., Siegal, M., Tanaka, K., & Kashima, E. S. (1992). Do people believe behaviours are consistent with attitudes? Towards a cultural psychology of attribution processes. *British Journal of Social Psychology*, *31*, 111–124.

Keil, F.C. (1989) *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kelemen, D, and Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111,* 138–43.

Kernis, M. H., & Goldman, B. M. (2004). Authenticity, social motivation, and wellbeing. In J. P. Forgas, K. D. Williams & S. Laham (Eds.), *Social motivation: Conscious and unconscious processes* (pp. 210–227). New York, NY: Cambridge University Press.

Kitayama, S. (2002). Culture and basic psychological processes: Toward a system view of culture. *Psychological Bulletin*, *128*, 189–196.

Kitayama, S., Park, H., Sevincer, A. T., Karasawa, M., & Uskul, A. K. (2009). A cultural task analysis of implicit independence: Comparing North America, Western Europe, and East Asia. *Journal of Personality and Social Psychology, 97,* 236–255.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition, 127,* 242–257.

Kuhn, M. H., & McPartland, T. S. (1954). An empirical investigation of self-attitudes. *American Sociological Review*, *19*, 68–76.

Landau, M. J., Vess, M., Arndt, J., Rothschild, Z. K., Sullivan, D., & Atchley, R. A. (2011). Embodied metaphor and the “true” self: Priming entity expansion and protection influences intrinsic self-expressions in self-perceptions and

interpersonal behavior. *Journal of Experimental Social Psychology*, *47*, 79–87.

Lynch, E. B., Coley, J. D., & Medin, D. L. (2000). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory & Cognition*, *28*, 41–50.

Maddux, W. W., Yang, H., Falk, C., Adam, H., Adair, W., Endo, Y., ... & Heine, S. J. (2010). For whom is parting with possessions more painful? Cultural differences in the endowment effect. *Psychological Science*, *21*, 1910–1917.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: a (surprising) meta-analysis. *Psychological Bulletin*, *132*, 895–919.

Markus, H. R., and Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98,* 224–253.

May, J., & Holton, R. (2012). What in the world is weakness of will? *Philosophical Studies*, *157*, 341–360.

Medin, D. L., and Ortony, A. (1989). Psychological essentialism. In S. Vosniadou and A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge: Cambridge University Press.

Molouki, S., & Bartels, D. M. (in press). Personal change and the continuity of identity. *Cognitive Science.*

Morris, M. W., & Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social Psychology*, *67*, 949–971.

Na, J., Grossmann, I., Varnum, M. E. W., Kitayama, S., Gonzalez, R., & Nisbett, R. E. (2010). Cultural differences are not always reducible to individual differences. *Proceedings of the National Academy of Sciences of the United States of America, 107,* 6192–6197.

Newman, George E. (2013), "The Duality of Art: Body and Soul," [commentary on Bullot and Reber] *Brain & Behavioral Sciences*, *36*, 153.

Newman, G. E., Bartels, D. M., & Smith, R. K. (2014a). Are artworks more like people than artifacts? Individual concepts and their extensions. *Topics in Cognitive Science*, *6*, 647–662.

Newman, G. E., & Bloom, P. (2012). Art and authenticity: The importance of originals in judgments of value. *Journal of Experimental Psychology: General*, *141*, 558–569.

Newman, G. E., Bloom, P., & Knobe, J. (2014b). Value judgments and the true self. *Personality and Social Psychology Bulletin*, *40,* 203–216.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science, 39,* 96–125.

Newman, G. E., Herrmann, P., Wynn, K., & Keil, F. C. (2008). Biases towards internal features in infants' reasoning about objects. *Cognition*, *107*, 420–432.

Newman, G., & Keil, F.C. (2008). 'Where's the essence?': Developmental shifts in children's beliefs about the nature of essential features. *Child Development*, *79,* 1344–1356.

Newman, G. E., Lockhart, K. L., & Keil, F. C. (2010). "End-of-life" biases in moral evaluations of others. *Cognition*, *115*, 343–349.

Opfer, J. E., & Siegler, R. S. (2004). Revisiting preschoolers' living things concept: A microgenetic analysis of conceptual change in basic biology. *Cognitive Psychology, 49*, 301–332.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411–419.

Peterson, C., Semmel, A., Von Baeyer, C., Abramson, L. Y., Metalsky, G. I., &

Seligman, M. E. (1982). The attributional style questionnaire. *Cognitive Therapy and Research*, *6*, 287–299.

Phillips, J., De Freitas, J., Mott, C., Gruber, J. & Knobe, J. (2017). True happiness: The role of morality in the concept of happiness. *Journal of Experimental Psychology: General.*

Prinz, J., & Nichols, S. (in press). Diachronic identity and the moral self. In J. Kiverstein (Ed.), Handbook of the social mind. London: Routledge.

Realo, A., & Allik, J. (1999). A cross-cultural study of collectivism: A comparison of American, Estonian, and Russian students. *The Journal of Social Psychology*, *139*, 133–142.

Rhee, E., Uleman, J. S., Lee, H. K., Roman, R. J. (1995). Spontaneous self-descriptions and ethnic identities in individualistic and collectivistic cultures. *Journal of Personality & Social Psychology, 69*, 142–152.

Rose, D., & Schaffer, J. (2015). Folk mereology is teleological. *Noûs*. doi: 10.1111/nous.12123.

Rosenberg, M. (1956). Misanthropy and political ideology. *American Sociological Review, 21,* 690–695.

Rosenberg, M. (1957). Organizations and values (pp. 25-35). Glencoe, IL: Free Press. This scale was published in: Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). Measures of personality and social psychological attitudes (pp. 404–406). San Diego: Academic Press.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology,* 76, 574–586.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5, 296–320.

Schimel, J., Arndt, J., Pyszczynski, T., & Greenberg, J. (2001). Being accepted for who we are: Evidenced that social validation of the intrinsic self reduces general defensiveness. *Journal of Personality and Social Psychology*, *80*, 35–52.

Schlegel, R. J., Hicks, J. A., Arndt, J., & King, L. A. (2009). Thine own self: True self-concept accessibility and meaning in life. *Journal of Personality and Social Psychology*, *96*, 473–490.

Schlegel, R. J., Hicks, J. A., Davis, W. E., Hirsch, K. A., & Smith, C. M. (2013). The dynamic interplay between perceived true self-knowledge and decision satisfaction. *Journal of Personality and Social Psychology*, *104*, 542–558.

Schlegel, R. J., Hicks, J. A., King, L. A., & Arndt, J. (2011). Feeling like you know who you are: Perceived true self-knowledge and meaning in life. *Personality and Social Psychology Bulletin*, *37*, 745–756.

Schwartz, S. (1994). *Beyond individualism/collectivism: New cultural dimensions of values*. In U. Kim & H. Triandis (Eds.), Individualism and collectivism: Theory, method, and applications (pp. 85-119). Thousand Oaks, CA: Sage.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in Experimental Social Psychology*, *38*, 183–242.

Shweder, R. A., & Sullivan, M. (1993). Cultural psychology: Who needs it? *Annual Review of Psychology, 44,* 497–523.

Singelis, T. (1994). The measurement of independent and interdependent self-construals. *Personality and Social Psychology Bulletin, 20*, 580–591.

Sousa, P., Atran, S., & Medin, D. (2002). Essentialism and folkbiology: Evidence from Brazil. *Journal of Cognition and Culture*, *2*, 195–223.

Sousa, P., & Mauro, C. (2015). The evaluative nature of the folk concepts of weakness and strength of will. *Philosophical Psychology*, *28,* 487–509.

Sperber, D. (1996). *Explaining culture* (pp. 1-2). Oxford: Blackwell Publishers.

Sripada, C. S. (2010). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*, 159–176.

Strohminger, N. and Nichols, S. (2014). The essential moral self. *Cognition, 31,* 159–171.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193–210.

Tobia, K. P. (2016). Personal identity, direction of change, and neuroethics. *Neuroethics*, *1,* 1–7.

Tönnies, F. (1887). *Community and society*. Oxford, UK: Transaction Books.

Triandis, H. C. (1989). The self and social behavior in differing cultural contexts. *Psychological review*, *96*, 506–520.

Triandis, H. (1995). Individualism and collectivism. Boulder, CO: Westview.

Uchida, Y., Townsend, S. S., Markus, H. R., & Bergsieker, H. B. (2009). Emotions as within or between people? Cultural variation in lay theories of emotion expression and inference. *Personality and Social Psychology Bulletin*, *35*, 1427–1439

Varnum, M.E.W., Grossmann, I., Kitayama, S., & Nisbett, R.E. (2010). The origin of cultural differences in cognition: Evidence for the social orientation hypothesis. *Current Directions in Psychological Science, 19*, 9–13.

Xu, F., & Rhemtulla, M. (2005). In defense of psychological essentialism. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 2377-2380). Erlbaum Mahwah, NJ.

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality, 47*, 245–287.