COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2016) 1–21

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12364

Normative Judgments and Individual Essence

Julian De Freitas,[a] Kevin P. Tobia,[b] George E. Newman,[c] Joshua Knobe[b,d]

[a]*Department of Psychology, Harvard University*
[b]*Department of Philosophy, Yale University*
[c]*School of Management, Yale University*
[d]*Program in Cognitive Science, Yale University*

Received 28 April 2015; received in revised form 29 December 2015; accepted 5 January 2016

Abstract

A growing body of research has examined how people judge the persistence of identity over time—that is, how they decide that a particular individual is the same entity from one time to the next. While a great deal of progress has been made in understanding the types of features that people typically consider when making such judgments, to date, existing work has not explored how these judgments may be shaped by *normative* considerations. The present studies demonstrate that normative beliefs do appear to play an important role in people's beliefs about persistence. Specifically, people are more likely to judge that the identity of a given entity (e.g., a hypothetical nation) remains the same when its features improve (e.g., the nation becomes more egalitarian) than when its features deteriorate (e.g., the nation becomes more discriminatory). Study 1 provides a basic demonstration of this effect. Study 2 shows that this effect is moderated by individual differences in normative beliefs. Study 3 examines the underlying mechanism, which is the belief that, in general, various entities are essentially good. Study 4 directly manipulates beliefs about essence to show that the positivity bias regarding essences is causally responsible for the effect.

Keywords: Concepts; Essentialism; Normative factors; Identity; Persistence; True self; Morality

1. Introduction

This paper is concerned with how people determine the persistence of identity over time—that is, how they decide that an individual entity at t_0 is the same individual at t_1. For example, suppose we start out with a specific scientific paper, Paper X, and then begin revising it in a way that changes some properties of the paper. Certain properties will be regarded as inessential, and if we change those aspects, people should think that

Correspondence should be sent to Julian De Freitas, Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, MA 02138. E-mail: defreitas@g.harvard.edu

the resulting document is still Paper X. In contrast, other aspects of the paper will be regarded as absolutely essential, and if we change those, people should think that the resulting document is not truly Paper X at all.

Clearly, these judgments will be affected by a variety of different considerations, but the aim of the present studies is to explore one particular effect: Specifically, we demonstrate that people's intuitions in cases like this are influenced by the *normative* status of the properties themselves. That is, when reasoning about many different types of entities, people are inclined to view the properties of an entity that they regard as normatively good as most essential. In turn, if those normatively good properties are changed in some way, observers are more likely to judge that the entity has ceased to exist compared to when analogous changes occur to properties that are seen as normatively bad.

2. Individual essence

While there have been several theories proposed in metaphysics about what *should* constitute identity (in a normative sense), here our focus is descriptive. In other words, we focus on the criteria that everyday people use when making persistence judgments. Consider an example from the ancient thought experiment of Plutarch (2001), known as the "Ship of Theseus." Readers are asked to imagine a ship that, throughout its voyage, begins to decay. Each plank is replaced with a new one, until eventually, the ship is entirely composed of new parts. The question is whether this ship is still the "Ship of Theseus."

At first blush, one might think that a relatively straightforward way of approaching questions such as this would simply be to tally up all of the features at t_0 and all of the features at t_1 and assess the overall degree of similarity between them—presumably, if the two objects are sufficiently similar, an observer might conclude that they are, in fact, *the same*.

However, a growing literature within psychology and philosophy suggests that people's judgments in these cases are much more nuanced (for a review, see Rips, Blok, & Newman, 2006). One fact that emerges from this literature is that people do not seem to treat all features equivalently. Instead, they seem to prioritize deep, sometimes unobservable, characteristics over more surface attributes (e.g., Blok, Newman, & Rips, 2005; Hall, Waxman, Brédart, & Nicolay, 2003; Newman, Bartels, & Smith, 2014). Consider a study by Blok, Newman, Behr, and Rips (2001). One set of participants read about a male accountant Jim, who underwent plastic surgery to resemble Marsha, a female actress. The other set of participants read a similar story in which Jim's brain was replaced with Marsha's. Both groups then reported whether the individual was still Jim or had become Marsha after the surgery. A significantly greater proportion of participants in the Brain Transplant group believed Jim's identity had changed than in the Plastic Surgery group (45% and 15%, respectively). Although intuitive, these results highlight the fact that people will often overlook perceptual similarity in favor of "deeper" characteristics when making judgments of persistence.

A similar effect is observed among young children. For example, Hall et al. (2003) introduced 3- and 4-year-olds to an entity that was described in terms of a particular surface feature (a red-colored character named "Mr. Red"). Children were then told that the entity underwent a transformation that eliminated this very feature (it was painted green). However, despite this feature change, the majority of children reported that the entity was still Mr. Red, indicating that like adults, children seem to assign identity to something other than an entity's appearance.

One explanation for these patterns is that when assessing the persistence of individual entities, adults and young children alike tend to prioritize features that are seen as causally central (Rips & Hespos, 2015; Rips et al., 2006). For example, when reasoning about the persistence of persons, people weight the continuity of the brain and one's memories (Blok et al., 2005; see Johnson, 1990, for a similar result with school-age children); when reasoning about animals like lions and tigers, people prioritize the continuity of the animal's internal *stuff* (Newman, Herrmann, Wynn, & Keil, 2008; Rips et al., 2006); and when reasoning about physical objects like icebergs and rivers (Rips & Hespos, 2015; Rips et al., 2006), people incorporate relevant causal knowledge about those entities (e.g., how rivers flow and plausible changes in direction).

In the present studies, however, we explore an altogether different type of feature that may influence people's identity judgments. Specifically, we explore whether *valence*—that is, whether valuing certain traits as *good* versus *bad*—similarly influences persistence judgments. For example, are people more likely to conclude that an object is the "Ship of Theseus" when the ship becomes normatively better (e.g., it improves) than when it becomes normatively worse (e.g., it deteriorates)? It might seem strange to even ask whether valence can influence something as concrete as whether an object is considered the same individual. And yet, a growing body of research suggests that people's value judgments can actually influence their intuitions about all sorts of matters, which, on the surface, appear to have nothing to do with values (see Knobe, 2010).

3. Individual essence and normative belief

One insight into the role of normative judgments comes from existing work on the way people think about the essences of human beings. Recent studies on people's intuitions about individual human beings suggest that people tend to believe that the most essential properties of humans are their moral properties (Strohminger & Nichols, 2014). For example, in one study, participants were asked to consider a person named John who had various properties. When participants were told that he lost non-moral properties (certain preferences, perceptual capacities, etc.), they tended to conclude that he was still John. In contrast, when he lost moral properties, observers concluded that the resulting person was no longer John (Strohminger & Nichols, 2014).

Within the existing literature, this effect has been characterized by saying that people regard moral traits as lying at the *essence* of the self (Strohminger & Nichols, 2014). We follow that terminology here. Thus, we will be using the word "essence" to pick out

something that people attribute to individuals (rather than to categories) and that explains the special priority they assign to moral traits in persistence judgments. (In the General discussion, we explore the question as to whether this notion might be related in some way to the phenomenon of psychological essentialism found in representations of categories [e.g., Medin & Ortony, 1989].)

Additional research has shown that beyond considering a person's moral qualities, people tend to say that the most essential properties are those that are normatively *good* (De Freitas et al., unpublished data; Newman, Bloom, & Knobe, 2014; Newman, De Freitas, & Knobe, 2015). For example, when participants are told about a human being who has both good and bad moral properties, they tend to say that the good properties constitute the human being's "true self" (De Freitas et al., unpublished data; Newman, Bloom et al., 2014). In turn, this positivity regarding the essence of others appears to have a host of downstream consequences for judgments about questions such as whether the person is truly happy, or whether the person has shown weakness of will (Newman et al., 2015).

To date, these effects have largely been understood as resulting from the way in which people are inclined to think about human beings specifically. For example, one explanation for this effect is that people believe that, deep down, other people are fundamentally good because this belief serves an adaptive function in encouraging cooperation (Strohminger & Nichols, 2014). Believing the best about others' essential nature would seem to encourage a host of prosocial behaviors. Clearly, however, people do not have to cooperate with entities such as ships or science papers, so for that reason it is certainly plausible that the positivity bias regarding essence arises only for judgments about individual human beings.

Yet an alternative possibility is that this phenomenon reflects a more general role of normative evaluation in beliefs about individual concepts—one that extends even to entities other than individual persons. For example, just as people think that the morally good parts of a human being are the most essential, they might think that the scientifically good parts of a paper are the most essential. Note that if this is indeed true, then it suggests that *moral* traits per se are not necessarily the only traits that are seen as essential. Rather, it would simply be that moral traits are the most relevant positive traits in the case of human beings, but that other kinds of positive traits are more relevant in the case of other entities.

In support of this prediction, developmental work has found that children think that properties such as poor eyesight or a missing finger will spontaneously improve over time (Lockhart, Chang, & Story, 2002), which might reflect an early tendency within folk biology to regard the good properties of the body as most essential. The real test of this hypothesis, however, is whether people's intuitions about ubiquitous entities other than individual human beings—such as institutions, groups, and texts—actually do show the same basic pattern observed for intuitions about individual human beings. For example, are people more inclined to say that the identity of a scientific paper changes when it loses its most scientifically valuable sections? Is a rock band less likely to be considered "the same" when it stops performing the songs that are regarded as esthetically good?

4. The present studies

Study 1 demonstrates a basic asymmetry effect in which observers perceive that removing good properties is more disruptive to identity than removing bad properties. Study 2 tests whether this asymmetry is based on people's own particular values about what constitutes the good versus bad properties of an entity, suggesting that the effect does not arise only for certain types of entities or particular transformations. Study 3 confirms that these asymmetric judgments based on valence are driven by beliefs about the entity's "essence," while ruling out an alternative possibility based on whether the entity is viewed as continually belonging to the same category. Study 4 directly manipulates beliefs about essence to show that this normative essentialism does in fact *causally* affect persistence judgments.[1]

5. Study 1: Basic effect

5.1. Methods

Three hundred and twenty participants were recruited using Amazon's Mechanical Turk, and 86 participants were excluded for failing to answer comprehension questions correctly, yielding a final sample of 234 participants (M_{age} = 30, 32% female). However, for all studies, including all participants does not alter the results. Participants were assigned to one of ten conditions in a 2 (valence: improvement vs. deterioration) × 5 (vignette) design. The vignettes described an entity that either improved or deteriorated (see the Supplementary file for all materials). To ensure that people thought the good and bad properties were equally intended, we included explicit information about intentionality and described the conditional change as going from a majority good (bad) to a majority bad (good). For example, "*In the majority of its regions the local government intentionally teaches people to express their opinions freely in public... Now, in the majority of regions the local government intentionally teaches people to discriminate against one another for being different*." The different vignettes served merely as a robustness check, and they included a band, science paper, nation, university, and conference.

Note that although some of the vignettes described the physical replacement of an entity's parts, these changes were always held constant across the two conditions, with the only difference between conditions being the direction of valence change (i.e., good to bad vs. bad to good). Furthermore, in the nation vignette there was no physical replacement of parts at all (the government merely changed its teachings in some regions). Finally, another way in which the current changes were different from the kind of change employed in Ship of Theseus-like thought experiments is that they involved replacing the current parts with completely different parts (i.e., bad vs. good), rather than with newer versions of the old parts (Plutarch, 2001).

Participants were then asked to rate, using a 1–7 scale, the extent to which they agreed with a question about identity persistence (1 = Completely disagree, 4 = Neither agree nor disagree, 7 = Completely agree):

The [Bellshore] after the changes is not really the same country as the [Bellshore] before the changes.

Participants also responded to a second counterbalanced item:

Person A thinks that [Bellshore] after the changes is still the same [country] as [Bellshore] before the changes.

Person B thinks that it makes more sense to say that [Bellshore] is no longer the same [country] it used to be. The way he sees it, the original [Bellshore] no longer exists.

Who do you agree with more, Person A or Person B?

They rated their agreement using a 1–7 scale (1 = Person A, 4 = Equally agree with both persons, 7 = Person B). We used a gradable measure rather than a forced-choice measure to gain the required sensitivity to compute mediation and moderation analyses in later experiments. Notice that what was gradable was just the level of participants' *agreement* with the persistence statement. In other words, if a participant gives a rating of 4 to the persistence statement, it would not be that this participant necessarily thinks that an object is *only persisting to a certain degree*. Rather, it could be that they are just *agreeing to a certain degree* with the claim that the object persists (e.g., it could be that they are certain that the object either persists or does not persist, but they are just not sure which.) Therefore, this measure can still be interpreted as tapping into intuitions about identity, with different ratings reflecting different degrees of agreement with the provided identity statements.

Finally, participants completed two comprehension checks (see the Supplementary file for both questions).

5.2. Results and discussion

The two items measuring persistence intuitions showed high internal consistency ($\alpha = 0.81$) and were averaged to produce a single measure. As predicted, a 2 (valence: improvement vs. deterioration) × 5 (vignette) ANOVA revealed that participants were more likely to agree that the entity's identity changed when it deteriorated ($M = 5.67$, $SD = 1.18$) than when it improved ($M = 4.99$, $SD = 1.38$), $F(1, 224) = 19.24$, $p < .001$, $\eta_p^2 = .079$. There was also a main effect of vignette, $F(4, 224) = 4.86$, $p < .01$, $\eta_p^2 = .080$. However, this factor did not interact with valence, $F(4, 224) = 1.23$,

p = .301, η_p^2 = 021. All vignettes were directionally consistent with our hypothesis (see Fig. 1).

6. Study 2: The role of values

The results of Study 1 were consistent with the notion that even for entities such as institutions, groups, and texts, people tend to view the normatively good properties as most essential. As a result, they are more likely to judge that the entity has remained the same when the entity's features improve than when the features deteriorate.

People, however, sometimes have very different values about the same thing. This observation naturally gives rise to another prediction from our theory: When people value opposing characteristics in the same entity, they should exhibit correspondingly different views about whether an increase or decrease in each characteristic constitutes an improvement or a deterioration, which in turn, should influence their intuitions about persistence. Note that we are predicting different identity judgments about the very same scenarios, depending only on differences in participants' own values. Thus, beyond demonstrating this phenomenon in a new way, this study provides an important test that the asymmetries observed in Study 1 are not unique to, for example, certain types of entities or particular transformations.

To test this, Study 2 exploited differences in values between liberals and conservatives (Graham, Haidt, & Nosek, 2009). We predicted an interaction effect whereby liberals would be more likely to say that the entity's identity had changed when it acquired "con-

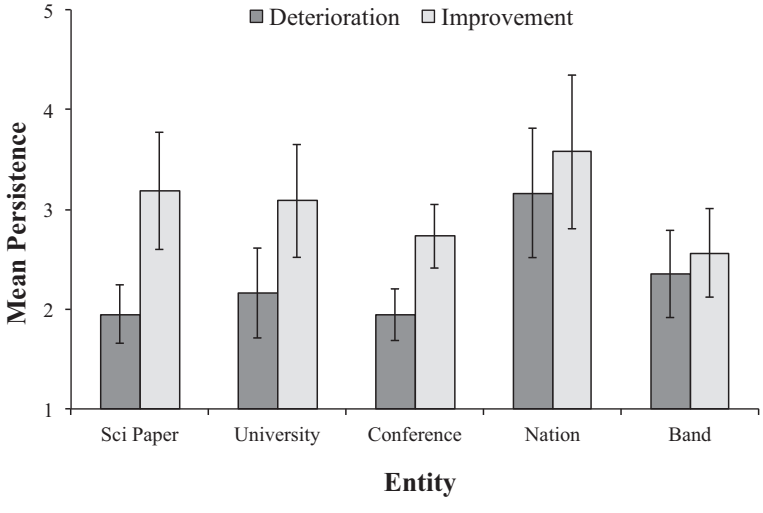

Fig. 1. Persistence ratings for each condition of each vignette in Study 1. Scores are reverse coded such that higher values indicate higher identity persistence. Error bars show 95% CI.

servative" properties, while conservatives would be more likely to say that the entity's identity had changed when it acquired "liberal" properties.

6.1. Methods

Three hundred and twenty participants were recruited from Amazon's Mechanical Turk, and 76 participants were excluded for failing to answer comprehension questions correctly, yielding a final sample of 244 participants (M_{age} = 46, 38% female). This study used a mixed-model design with change (toward conservative vs. toward liberal) as a between-subject factor and vignette as a within-subject factor. So each participant judged two vignettes, which served as a robustness check, and the two vignettes were always in the same condition—either changing toward liberal or conservative. For the two conditions of a particular vignette, the first portion of the vignette was always exactly the same. Then, the entity was described as changing in either a more liberal direction or a more conservative direction (see the Supplementary file for all materials).

Participants then received the same identity measure from Study 1, in which two people disagree about whether the entity is still the same after the changes. They also indicated their political orientation (1 = liberal, 7 = conservative) and answered two comprehension questions about each vignette.

6.2. Results and discussion

For our initial analyses, we used the mean of the responses to the two vignettes, yielding a single persistence score for each participant. We then conducted a linear regression analysis with condition, political orientation, and the interaction between condition and political orientation as factors. This analysis indicated a significant interaction between condition and political orientation, $\beta = -.24$, $SE = 0.07$, $p < .001$, such that political orientation moderated the effect of condition on persistence judgments. After running this first regression, we then ran two separate regressions, in which we regressed identity judgments on political orientation for each of the valence conditions, respectively. In the "change toward liberal" condition, conservatives were significantly more likely than liberals to say the identity was lost, $r = .25$, $p = .016$, while in the "change toward conservative" condition, liberals were significantly more likely than conservatives to say the identity was lost, $r = -.24$, $p = .010$. We also observed a main effect of condition whereby participants were more likely to say identity changed when the organization became more conservative versus more liberal, $\beta = .32$, $SE = 0.11$, $p = .004$. This result was likely due to a liberal-leaning sample ($M = 3.09$, $SD = 1.61$; midpoint = 4). There was no main effect of political orientation, $\beta = .001$, $SE = 0.07$, $p = .984$.

As a robustness check, we repeated the same analyses for each of the two vignettes separately. For the conference vignette, we again observed a significant interaction between condition and political orientation, $\beta = -.19$, $SE = 0.09$, $p = .031$, a significant main effect of condition, $\beta = .46$, $SE = 0.15$, $p = .002$, and no main effect of political orientation, $\beta = .03$, $SE = 0.09$, $p = .740$. Furthermore, in the "change toward liberal"

condition, conservatives were marginally more likely than liberals to say the identity was lost, $r = .22$, $p = .103$, while in the "change toward conservative" condition, liberals were marginally more likely than conservatives to say the identity was lost, $r = -.16$, $p = .155$. For the scout club vignette, we also observed a significant interaction between condition and political orientation, $\beta = -.31$, $SE = 0.11$, $p = .005$, but no significant main effect of condition, $\beta = .20$, $SE = 0.17$, $p = .234$, and no main effect of political orientation, $\beta = -.03$, $SE = 0.11$, $p = .764$. Furthermore, in the "change toward liberal" condition, conservatives were marginally more likely than liberals to say the identity was lost, $r = .28$, $p = .074$, while in the "change toward conservative" condition, liberals were significantly more likely than conservatives to say the identity was lost, $r = -.34$, $p = .025$.

7. Study 3: Essence mechanism

Studies 1 and 2 assumed, in line with previous work, that persistence judgments about an entity can be used to reveal beliefs about what constitutes its essence: If the removal of X properties disrupts identity judgments more so than the removal of Y properties, then X properties must be more essential (Blok et al., 2005; Hall et al., 2003; Newman, Bartels et al., 2014; Rips et al., 2006; Strohminger & Nichols, 2014). Study 3 sought to provide direct evidence that essentialism is driving the valence asymmetry in persistence judgments observed in Studies 1 and 2—that is, that the valence effect is explained by a tendency to view positive properties as more essential to identity than negative properties. We also sought to rule out an alternative possibility, inspired by research demonstrating the importance of category membership for persistence judgments (e.g., Rhemtulla & Xu, 2007). For example, people may think that the very definition of what it is to belong to the *category* of a nation, band, etc. is spelled out in positive (rather than negative) properties. As a result, participants may have been less likely to say the entity was the same when it possessed negative properties simply because it was no longer seen as satisfying the definition of the category to which it previously belonged. To address this, Study 3 asked about both category membership and essence.

7.1. Methods

Three hundred and twenty new participants were recruited from Amazon's Mechanical Turk, and 104 participants were excluded for failing to answer comprehension questions correctly, yielding a final sample of 216 participants ($M_{age} = 29$, 30% female). The design was nearly identical to that of Study 1, except that we included two additional measures about category membership and essence. Following the persistence measure, participants rated their agreement (1 = completely disagree, 7 = completely agree) with the following statements (order counterbalanced across subjects):

Plainly speaking, if you had to categorize [Bellshore] after the changes, you would say that it is a [country] (as opposed to another category, such as an animal, car, fruit, etc).

[Bellshore] after the changes no longer reflects the true essence of the original [Bellshore].

Previous studies have successfully employed this kind of wording to probe intuitions about essence (e.g., Newman, Bloom et al., 2014; Newman et al., 2015). Finally, participants completed comprehension questions about the vignette.

7.2. Results and discussion

A 2 (valence: improvement vs. deterioration) × 5 (vignette) ANOVA again revealed that participants were significantly more likely to agree that the entity's identity had changed when it deteriorated (M = 5.58, SD = 1.45) than when it improved (M = 4.58, SD = 1.87), $F(1, 206) = 19.95$, $p < .001$, $\eta_p^2 = .088$. There was no effect of vignette, $F(4, 206) = 1.98$, $p = .099$, $\eta_p^2 = .037$, and no valence × vignette interaction, $F(4, 206) = 2.56$, $p = .456$, $\eta_p^2 = .017$.

To determine whether beliefs about essence explain the effect of valence on identity ratings—that is, to determine the extent to which condition influences identity judgments through the mediator variable, essence—we conducted a multiple mediation bootstrap analysis (Hayes, 2012; Preacher & Hayes, 2008), with condition as the independent variable, ratings of identity persistence as the dependent variable, and measures of essence (M = 4.51, SD = 1.66; midpoint = 4) and category membership (M = 6.51, SD = 1.03; midpoint = 4) as potential mediators. This analysis indicated that essence did indeed significantly mediate the effect of valence on identity judgments (95% CI = −1.2 to −.57), whereas category membership did not (95% CI = −.05 to .12; see Fig. 2, and Table 1 for variable means). Therefore, intuitions about essence appear to explain the asymmetric effect of valence on judgments of object identity. Tests indicated that multicollinearity between the identity and essence variables was not a concern ($r < .8$).

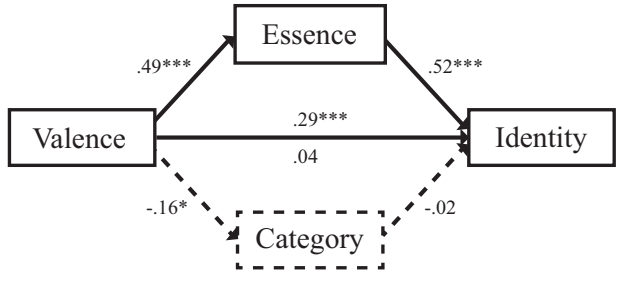

Fig. 2. Mediation results from Study 3, showing standardized coefficients. Scores are reverse coded such that higher values indicate higher identity persistence. *$p < 0.5$; ***$p < .001$.

8. Study 4: Causal influence of essence

The final study sought to test whether essence is *causally* responsible for the effects observed here. We reasoned that even though people assume by default that the essence of an entity is good, if they are directly told that the essence of an entity is bad, then this should lead to a reversal in their intuitions about whether identity persists as a result of deterioration versus improvement. In particular, a bad essence entity that deteriorates should be seen as staying in line with its true essence, and so people should be more inclined to say its identity is the same after the changes; by contrast, a bad essence entity that improves should be seen as deviating from its true essence, and so people should be more inclined to say its identity is not the same after the changes.

Note that this study also provides an especially direct test of our hypothesis in that we are only manipulating whether the entity's essence is initially described as good or bad. Further, such a design addresses two alternative explanations of our results. The first is that people are more inclined to say that identity changes after deterioration to communicate their disapproval of the negative resulting characteristics. Note that our prediction for this study is that if an entity is described as having a bad essence, then people should be more inclined to say exactly the opposite: that its identity no longer exists after it *improves*.

The second alternative explanation is that the valence asymmetry can be explained by the tainting quality of bad properties (Reeder, 1993; Reeder & Brewer, 1979). On this account, any traits that are initially bad will irrevocably "contaminate" an entity's

Table 1
Results for Experiments 3 and 4. Mean scaled ratings (*SD*s in parentheses) and variable correlations.

Experiment 3

	Scaled Ratings		Correlations		
Variables	Deterioration	Improvement	1	2	3
1. Identity	5.58 (1.45)	4.58 (1.87)	–		
2. Essence	5.36 (1.37)	3.73 (1.51)	0.510**	–	
3. Category	6.34 (1.16)	6.67 (.853)	0.052	−0.048	–

Experiment 4

	Scaled Ratings				Correlations			
	Good Essence		Bad Essence					
Variables	Deterioration	Improvement	Deterioration	Improvement	1	2	3	4
1. Identity	6.00 (1.29)	4.45 (1.87)	4.04 (2.08)	5.46 (1.39)	–			
2. Essence	5.74 (1.40)	3.58 (1.86)	3.20 (1.86)	5.58 (1.41)	0.701**	–		
3. Good Traits	1.97 (1.54)	6.31 (1.36)	1.62 (1.31)	6.32 (1.14)	0.006	0.018	–	
4. Bad Traits	6.34 (1.34)	1.75 (1.49)	6.55 (1.18)	1.69 (1.26)	−0.011	−0.026	−0.926**	–

***p < .01.

essence, such that subsequent improvements will not be able to shake the impression that the essence remains contaminated. As such, participants in our experiments may have been saying that an entity's identity was still the same after improving because they viewed the entity as being irrevocably tainted from the start, rather than as always having a good essence. Yet notice again that we are making exactly the opposite prediction for the following study: If an essence is explicitly described as having a bad essence, it will be viewed as losing its identity if it *improves*; by contrast, the tainting account predicts that improvement will not change identity judgments for a tainted entity.

8.1. Methods

A total of 640 participants (M_{age} = 34, 246 female) were recruited using Amazon's Mechanical Turk, and assigned to one of four conditions in a 2 (essence: good vs. bad) × 2 (valence: improvement vs. deterioration) design. Participants were shown a vignette about a fictional educational institute that existed during the Nazi regime, named the "Iserlohn Institute." The institute was explicitly described as having either a good essence (teaching traditional academic ideals) or bad essence (teaching Nazi ideology):

> *During the Nazi regime, some educational institutions taught a mixture of courses on traditional academic subjects (science, literature, etc.) and courses in Nazi ideology (often with strong anti-Semitic messages). But the Iserlohn Institute was different. Even though it taught a mix of these two kinds of courses, everyone who enrolled could tell that the real essence of the institution was its focus on academic subjects like science and literature [Nazi ideology and anti-Semitism]. The material they taught on Nazi ideology [traditional academic subjects] was just a thin veneer over this more essential part of the curriculum.*

To determine whether we had successfully convinced participants that the essence of the institute was good or bad, we then asked them a manipulation check question, *Based on this information, how would you characterize Iserlohn Institute's "true essence"?* (1 = Fundamentally bad, 7 = Fundamentally good).

Next, participants read that the institute either deteriorated or improved:

> *. . .Then, after a number of years, there was a sudden administrative change. The rector of the institute was replaced by a new rector who decided to shake things up in certain ways. Specifically, the new rector decided to completely eliminate all courses on traditional academic subjects (science, literature, etc.) [Nazi ideology and anti-Semitism]. Instead, from that day onwards the institute always taught courses in just Nazi ideology [traditional academic subjects].*

Participants then completed a measure of identity (two people disagreeing about whether the entity was still the same institute after the changes). Afterward participants were asked three questions (counterbalanced between subjects, with each question

presented on separate pages) that we intended to use as potential mediator variables. The essence mediator asked participants how much they agreed (1 = Completely disagree, 7 = Completely agree) with the statement *Iserlohn Institute after the changes no longer reflected the true essence of the original Iserlohn Institute.* Since, based on the tainting account, the amount of good versus bad traits should predict people's judgments (i.e., a tainted entity should be viewed as having more bad traits than good traits), we also included two other potential mediators that asked about the amount of bad traits and good traits remaining after the changes: ***After** the changes, how much of the curriculum at Iserlohn Institute contained **Nazi ideals [traditional academic pursuits]**?* (1 = None at all, 7 = A great deal). We predicted that the essence variable would mediate the effects, while these other variables would not. We did not include any comprehension questions, since this study was already fairly long compared to the others, and in the previous studies excluding participants based on comprehension questions did not have a qualitative effect on the overall statistical results.

8.2. Results and discussion

Results from the manipulation check question indicated that we successfully manipulated essence judgments: Participants were significantly more likely to rate the institute as good when they read the "good" essence stem (M = 4.73, SD = 1.54) than the "bad" essence stem (M = 1.86, SD = 1.33), $t(638) = 25.26$, $p = .002$.

Mean ratings on the persistence statement for each condition are displayed in Fig. 3. A 2 (essence: good vs. bad) × 2 (valence: improvement vs. deterioration) ANOVA indicated a significant interaction between essence and valence, $F(1, 636) = 123.29$, $p < .001$, $\eta_p^2 = .16$. Consistent with our hypothesis, when the institute was described as having a good essence, participants were significantly more likely to agree that the entity was no longer the same entity after it deteriorated (M = 6.00, SD = 1.29) than after it improved (M = 4.45, SD = 1.87), $t(319) = 8.67$, $p < .001$. Conversely, when the entity was described as having a bad essence, participants were more likely to agree that the entity's identity was no longer the same after it improved (M = 5.46, SD = 1.39) than after it deteriorated (M = 4.04, SD = 2.08), $t(320) = 7.17$, $p < .001$.

To determine whether the interaction effect is explained by beliefs about essence—that is, to determine the extent to which the condition × essence interaction influences identity judgments through the essence mediator variable—we conducted a multiple mediation bootstrap analysis (Hayes, 2012; Preacher & Hayes, 2008), with the interaction as the independent variable, ratings of identity persistence as the dependent variable, and the potential mediators (a) essence, and (b) the average of the amount of positive traits and (reverse coded) negative traits, since answers to these questions were highly correlated (see Table 1). This analysis indicated that essence did indeed significantly mediate the effect of the interaction on identity judgments (95% CI = −.86 to −.63), whereas the effect was not mediated by the amount of good and bad traits after the changes (95% CI = −.003 to .01; see Fig. 4, and Table 1 for variable means). Thus, we find no evidence in favor of the contamination account, while accumulating even more evidence for

the essence account. Tests indicated that multicollinearity between the identity and essence variables was not a concern ($r < .8$).

Taken together, these results indicate that attributions of individual essence have a causal impact on persistence judgments, and that individual essence attributions mediate the impact of valence on persistence judgments.

9. General discussion

Across four studies, people showed a bias toward believing that the core properties of entities are positive, and therefore they were more likely to say that an entity persisted if its properties improved (becoming normatively better) than if those properties deteriorated

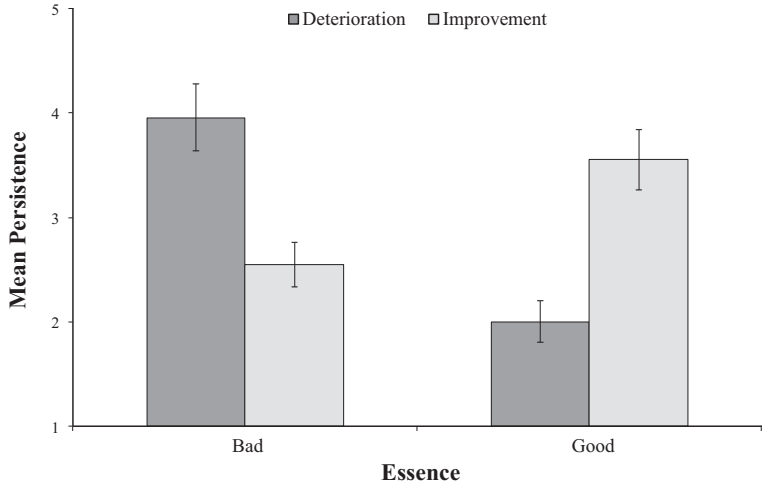

Fig. 3. Means by condition in Study 4. Scores are reverse coded such that higher values indicate higher identity persistence. Error bars show 95% CI.

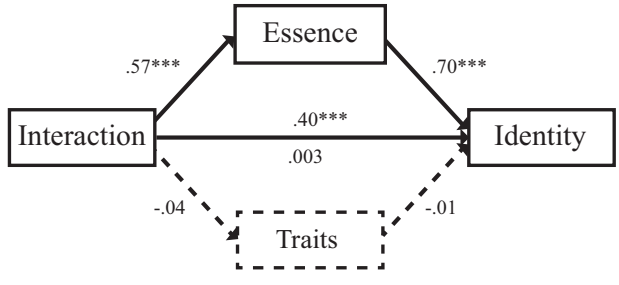

Fig. 4. Mediation results from Study 4, showing standardized coefficients. Scores are reverse coded such that higher values indicate higher identity persistence. ***$p < .001$.

(becoming normatively worse). Study 1 demonstrated this basic effect for many different types of entities; Study 2 showed that these persistence judgments are moderated by individual differences in values; Study 3 provided direct evidence for the essence mechanism; and Study 4 demonstrated that manipulating beliefs about essence causally changes the nature of identity judgments.

9.1. Relation to judgments about human beings

Recall that earlier research indicates that normative considerations affect judgment about the essences of *human beings* (Newman, Bloom et al., 2014; Newman et al., 2015). At first, it might seem that this effect is unique to judgment about persons (e.g., Sears, 1983; Tobia, 2015). However, the present results suggest that this effect actually extends to other entities like institutions, groups, and texts. The good aspects of humans are seen as more essential, but so too are the good aspects of bands, science papers, universities, conferences, scout camps, and nations. This finding suggests that the effect of normative beliefs on persistence judgments is not something specific to beliefs about individual human beings. Instead, such normative considerations seem to reflect a more general mechanism.

One might wonder whether the effect of valence on identity judgments is *stronger* for entities that are more human-like or person-like. This, however, does not appear to be the case. For example, we conducted a posttest in which 64 participants rated how similar (1 = Not at all, 7 = Very Similar) each entity was to a person, for example: *How similar are [universities] to people?* These anthropomorphism values were then correlated with the positivity biases (good > bad) for each of the entities in Experiments 1 and 3. If the effect was stronger for person-like entities, we would expect to find a positive correlation. Yet we found a significant *negative* correlation for Experiment 1 ($r = -.982$, $p = .003$), and no significant correlation for Experiment 3 ($r = -.751$, $p = .144$).

One question that has arisen from recent research is *why* people tend to believe the essence of other humans is good (Newman, Bloom et al., 2014; Newman et al., 2015). The present studies—and specifically, the analogous findings across humans, institutions and groups—provide insights regarding this question. For instance, one might initially suppose that we judge other people's good traits as most essential because this has the specific function of fostering cooperation. But it is less plausible that people's tendency to perceive entities such as bands and scout camps as essentially good is to be explained in terms of a tendency to cooperate with those entities. Similarly, we might long to believe that people are good "deep down" perhaps for optimistic, religious, or spiritual reasons, but it is less plausible that we hold out dearly for the good true soul of science papers.

In light of the present studies, one plausible explanation is that the role of normative considerations is not merely a symptom of our desires, but instead, represents a basic fact about the cognitive processes involved in how we represent individual entities. In other words, in conceiving the core features of entities we may simply be prone to automatically

apply our own normative judgments, such that our very understanding of an entity's identity is that it is consists of those traits that we value as good.

9.2. Scope of the effect

The current results suggest that previous findings—whereby positive moral traits are ascribed to the deepest aspects of individual human beings—also extend to judgments about entities of other types, such as group agents and institutions. But now one might wonder just how general this intuition is. Would it apply to entities of all types, or is it limited to certain kinds of entities?

For example, suppose we had looked at judgments of ordinary artifacts, such as cars, chairs, or houses. Would we still have found an effect such that these entities were judged to persist more in the improvement condition than in the deterioration condition? What if we had looked at biological kinds, such as animals or plants? Or what if we had looked at physical objects, such as rocks?

We suspect that the effect observed here would arise for objects of some kinds but not others. This raises the question as to whether a more general theory can be developed to specify the conditions under which observers are likely to incorporate their own normative beliefs into their persistence judgments. While we view this as an important next step for future work (and one that would require focused empirical studies), we can offer a tentative proposal.

Specifically, it may be that the types of normative effects observed here are most likely to arise in cases where there is believed to be some purpose or teleology of a particular entity. Some things quite obviously have a telos or purpose (e.g., the purpose of a fork is to help in eating), but research suggests people apply teleology more broadly (Kelemen, 1999; Kelemen & Rosset, 2009). Perhaps people believe that there is some sense in which the purpose of bands is to make meaningful music, the purpose of physics papers is to make valuable scientific contributions, and the purpose of human beings is to be morally good.

If this hypothesis turns out to be correct, it would give us a way of understanding the scope of the effect observed here. Specifically, it might be that the effect arises only for entities that are seen as having a deeper purpose in this relevant sense.

9.3. Existing research on identity

In the Introduction, we discussed a body of research on the way people reason about the persistence of individuals over time (Blok et al., 2005; Hall et al., 2003; Newman, Bartels et al., 2014; Strohminger & Nichols, 2014). A key insight from this work is that not all features are equally relevant in identity judgments. Instead, people seem to prioritize features of specific types.

As we noted above, existing work suggests that people prioritize features that they deem to be *causally central*—for example, a person's brain, a lion' innards, etc. (Blok et al., 2001, 2005; Johnson, 1990; Newman et al., 2008; Rips & Hespos, 2015;

Rips et al., 2006). The present studies, however, examined a very different sort of phenomenon. People appear to have a bias toward believing that core properties of entities are *normatively good* (see especially Study 2). A question now arises about how to understand the relationship between these two effects. Are these simply unconnected phenomena that both just happen to impact people's intuitions about identity, or is there some way to unify them in a more general theory?

While the existing data are not sufficient to answer this question, we think that there is, in fact, reason to believe that these phenomena are importantly related. In a forthcoming paper, Rips and Hespos (2015) suggest that people understand the identity of objects in a way that involves going beyond their superficial features and searching for a deeper principle that allows one to make sense of these features. As they note, existing work shows that this can be done by identifying the causal forces that "unify and shape the object" (Rips & Hespos, 2015, p. 9), but perhaps it can also be done by identifying a normative ideal that the object to some degree approximates. Clearly, these two notions are different in their specific content, but they appear to involve a similar sort of structure. Both involve a deeper principle that in one way or another allows us to make sense of the superficial features of the object and to unify them in a more coherent understanding of the object as a whole. We return to this issue briefly below.

9.4. Relation to category essence

The present studies explored people's judgments about individual entities, and we therefore drew on the existing literature about how people understand such entities (Hall et al., 2003; Rips et al., 2006; Strohminger & Nichols, 2014). Within this literature, the features that are seen as required for the persistence of an individual entity are sometimes referred to as that entity's "essence" (Strohminger & Nichols, 2014).

It should be noted, however, that the majority of existing work on psychological essentialism has been concerned with a somewhat different topic. This work has been concerned not with the essences of *individual* entities (e.g., the essence of an individual nation, the essence of an individual scientific paper), but rather with the essences of *categories* of entities (e.g., the essence of a species, the essence of a gender). For example, it has investigated how category membership is determined by beliefs about deep, causally central features (e.g., Keil, 1989) as well as similarity and physical appearance (e.g., Hampton, Estes, & Simmons, 2007; Hampton, Storms, Simmons, & Heussen, 2009). A question naturally arises therefore about whether the phenomena explored here are related in some way to questions about category essence.

To begin with, we can easily rule out the most direct sort of relation. It would not be at all plausible to suggest that people think the essence of an individual entity simply *is* the essence of one of the corresponding categories. For example, our hypothetical physics paper "Atom Dynamics" (from Study 1) is a member of various categories, including the category *physics papers* and the superordinate category *scientific papers*. Yet it would not be at all plausible to suggest that people see its individual essence (the essence of "Atom Dynamics") as simply being identical to the essence of one of these categories (e.g., the

essence of scientific papers). First, in Study 3, we find an effect that is mediated by judgments of individual essence but not by judgments of category essence. Second, the two notions seem to come apart quite clearly in cases in which an entity undergoes radical change. For example, suppose that someone removed almost all of the material from the paper "Atom Dynamics" and gradually turned it into a paper about a completely different question in physics. Then the resulting entity might still retain the essence of scientific papers, but it would not retain the individual essence of this one particular paper.

A more plausible view would be that the study of individual essence is connected to the study of category essence at a more indirect level. Perhaps there are certain principles about the way people attribute essences, and these principles apply both to attributions of individual essence and to attributions of category essence. Then, if we discover something about the way people understand individual essences, it might be that the very same thing will apply to the way people understand category essences.

Existing results do provide at least some indication that this may be the case. As we noted above, people's intuitions about individual identity tend to prioritize features that are *causally central* (Blok et al., 2001, 2005; Johnson, 1990; Newman et al., 2008; Rips & Hespos, 2015; Rips et al., 2006). Research on category essence has arrived at a parallel finding, showing that people's intuitions about category essence tend to involve features that are causally central (e.g., Ahn et al., 2001; Bloom, 2004; Gelman, 2003; Keil, 1989). Similarly, the present studies show that intuitions about individual identity tend to prioritize features that are deemed *normatively good*. Intriguingly, some work points to a parallel effect for intuitions about categories, indicating that people show a tendency to regard normatively good features of categories as especially essential (Barsalou, 1985; Knobe, Prasada, & Newman, 2013; Lynch, Coley, & Medin, 2000).

In short, there is at least some potential for an even greater degree of theoretical unification here. We noted in the previous section that it might prove possible to develop a more abstract account that would give a unified explanation of the different effects observed for individual entities (causal centrality and normative goodness). The claim now is that it might prove possible in turn to unify these effects with the parallel effects observed for judgments about categories. We are pursuing this topic in ongoing research.

9.5. Conclusion

At the beginning of this paper we posed a question about how people determine the persistence of an entity over time. Specifically, we asked whether this judgment is influenced by one's *normative* views about the entity's properties. Indeed, we found that people viewed normatively good properties as more essential than bad properties, such that removing the good properties was more likely to lead to the impression that the entity ceased to exist. Subsequent experiments confirmed that this asymmetric pattern of identity judgments was influenced by people's individual values (rather than by superficial factors), and that intuitions about essence (rather than about categorization or contamination) were causally responsible for the effect.

Extending previous work, these findings show that normative essentialism is applied not only to other humans, but also to institutions, groups, and texts, influencing identity judgments for all these various entities. In short, people appear to have a much more general tendency to believe that the essence of an object is good.

Acknowledgment

We thank Christina Starmans and Jillian Jordan for helpful comments.

Note

1. Two pilot studies are not reported in the current manuscript. The first found the same significant results as Study 1, but it was not included since the vignettes were less controlled; for example, they did not explicitly state that intentionality was constant across conditions, and the proportion of traits before and after the described changes was somewhat ambiguous. The second study replicated the basic pattern of Study 4, but not as convincingly: When participants were directly told that the essence of the entity was bad, then the direction of identity judgments did reverse (as predicted), but not significantly so (although we did still find a significant interaction between condition and the essence manipulation). We believe this may be because the phenomenon observed in the current experiments—that people have a default tendency to attribute positive traits to essences—worked against the manipulation in this study, which should have used "badder" traits.

References

Ahn, W. K., 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.

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.

Blok, S., Newman, G., Behr, J., & Rips, L. J. (2001). Inferences about personal identity. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 80–85). Mahwah, NJ: Erlbaum.

Blok, S. V., Newman, G., & Rips, L. J. (2005). Individuals and their concepts. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the lab* (pp. 127–149). Washington, D.C.: American Psychological Association.

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

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

Graham, J., Haidt, J., & Nosek, B. (2009). Liberals and conservatives use different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046.

Hall, D. G., Waxman, S. R., Brédart, S., & Nicolay, A. C. (2003). Preschoolers' use of form class cues to learn descriptive proper names. *Child Development*, *74*, 1547–1560.

Hampton, J. A., Estes, Z., & Simmons, S. (2007). Metamorphosis: Essence, appearance, and behavior in the categorization of natural kinds. *Memory & Cognition*, *35*, 1785–1800.

Hampton, J. A., Storms, G., Simmons, C. L., & Heussen, D. (2009). Feature integration in natural language concepts. *Memory & Cognition*, *37*, 1150–1163.

Hayes, A. F. (2012) PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www.afhayes.com/ public/ process2012.pdf

Johnson, C. N. (1990). If you had my brain, where would I be? Children's understanding of the brain and identity. *Child Development*, *61*, 962–972.

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

Kelemen, D. (1999). The scope of teleological thinking in preschool children. *Cognition*, *70*, 241–272.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*, 138–143.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*, 315–329.

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

Lockhart, K. L., Chang, B., & Story, T. (2002). Young children's beliefs about the stability of properties: Protective optimism? *Child Development*, *73*, 1408–1430.

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.

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

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

Newman, G. E., Bloom, P., & Knobe, J. (2014). 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.

Plutarch (2001). *Plutarch's Lives* (Vol. 1). H. Clough Ed. & J. Dryden, Trans. (Eds.), New York, NY: Random House LLC. (Original work published AD 46 – AD 127).

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

Reeder, G. D. (1993). Trait-behavior relations and dispositional inference. *Personality and Social Psychology Bulletin*, *19*, 586–593.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*, 61–79.

Rhemtulla, M., & Xu, F. (2007). Sortal concepts and causal continuity: Comment on Rips, Blok, and Newman (2006). *Psychological Review*, *114*, 1087–1094.

Rips, L. J., Blok, S., & Newman, G. (2006). Tracing the identity of objects. *Psychological Review*, *113*, 1–30.

Rips, L., & Hespos, S. (2015). Divisions of the physical world: Concepts of objects and substances. *Psychological Bulletin 141*, 786–811.

Sears, D. O. (1983). The person positivity bias. *Journal of Personality and Social Psychology*, *44*, 233–250.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*, 159–171.

Tobia, K. (2015). Personal identity and the Phineas Gage Effect. *Analysis*, *75*, 396–405.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Data S1. Stimuli Presented in Studies 1 and 3.