THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2013
Vol. 66, No. 6, 1113–1130, http://dx.doi.org/10.1080/17470218.2012.730533

Causal essentialism in kinds

Woo-kyoung Ahn[1], Eric G. Taylor[1], Daniel Kato[2], Jessecae K. Marsh[3], and Paul Bloom[1]

[1]Department of Psychology, Yale University, New Haven, CT, USA
[2]Columbia University, New York, NY, USA
[3]Department of Psychology, Lehigh University, Bethlehem, PA, USA

The current study examines causal essentialism, derived from psychological essentialism of concepts. We examine whether people believe that members of a category share some underlying essence that is both necessary and sufficient for category membership and that also causes surface features. The main claim is that causal essentialism is restricted to categories that correspond to our intuitive notions of existing kinds and hence is more attenuated for categories that are based on arbitrary criteria. Experiments 1 and 3 found that people overtly endorse causal essences in nonarbitrary kinds but are less likely to do so for arbitrary categories. Experiments 2 and 4 found that people were more willing to generalize a member's known causal relations (or lack thereof) when dealing with a kind than when dealing with an arbitrary category. These differences between kinds and arbitrary categories were found across various domains—not only for categories of living things, but also for artefacts. These findings have certain real-world implications, including how people make sense of mental disorders that are treated as real kinds.

Keywords: Concepts; Essentialism.

Some categories, like "robins" and "chairs", are intuitively more appropriate than other categories, like "things that weigh about 2 pounds, are white, and have a smooth surface" and "a group of mental disorder patients whose last names began with F" (see, e.g., Bloom, 2004; Gelman, 2003; Macnamara, 1986; Markman, 1989; Murphy & Medin, 1985). What are the assumptions that people have about these natural categories—or *kinds*? In this paper, we argue that it is the belief in causal essences that critically distinguishes between kinds and arbitrary categories. In what follows, we first define what we mean by a causal essence with reference to theories of psychological essentialism, and then we explain why kinds may elicit beliefs in a causal essence more so than arbitrary categories.

Psychological essentialism

It has been argued that many concepts are not mere collections of correlated features, but rather are groupings based on shared causal mechanisms anchored in a category's essence (e.g., Medin & Ortony, 1989; Murphy & Medin, 1985). For instance, people's concept of birds may involve

Correspondence should be addressed to Woo-kyoung Ahn, Department of Psychology, Yale University, 2 Hillhouse Avenue, New Haven, CT 06520, USA. E-mail: woo-kyoung.ahn@yale.edu

naïve theories about how having wings, flying, laying eggs, and so on are causally related and how some bird essence (e.g., "bird DNA") must be responsible for these features. From this perspective, essences are believed to make categories what they are and cause their surface features (Bloom, 2004; Gelman, 2003; Locke, 1894/1975; Medin & Ortony, 1989). This belief is referred to as psychological essentialism (Medin & Ortony, 1989).

Results from many previous studies are consistent with psychological essentialism. For example, the fact that natural kinds are categorized together based on internal, deeper features, despite differences in surface features (Ahn, 1998; Gelman & Wellman, 1991; Keil, 1989) supports the importance of essences/essential features. Essentialism has been credited with explaining why people will endorse an object as a piece of art regardless of its appearance as long as the artist's intention was to create art, calling upon a deeper underlying feature of the category of art (Gelman & Bloom, 2000). Furthermore, inferences that would follow from belief in an essence (e.g., biological bases, immutability, inductive potency, etc.) have been documented with domains as varied as social categories and mental disorder categories (e.g., Haslam, Rothschild, & Ernst, 2000; see Dar-Nimrod & Heine, 2011, for review). This variety of evidence has been used to support the claim that psychological essentialism is a pervasive cognitive belief. (But see also Kalish, 2002; Strevens, 2000, for evidence against psychological essentialism.)

The current study focuses on one specific aspect of psychological essentialism—namely, whether people explicitly endorse a causal essence (Gelman, 2003). We examine whether people believe that members of the same kind share something that causes surface features of the kind, and that this thing is necessary and sufficient for category membership. We call this specific claim causal essentialism in order to distinguish it from other claims made under psychological essentialism. While belief in a causal relationship between an essence and surface features, or causal essentialism, is considered one of the most crucial elements of an essence (Dar-Nimrod & Heine, 2011; Medin & Ortony, 1989), few previous studies have directly measured whether people explicitly believe across a variety of categories that what causes surface features is an essence.

Causal essentialism in kinds

The main claim of the current study is that causal essentialism is restricted to *kinds*—to collections of individuals that fit with people's intuitive notions of natural and nonarbitrary categories (see also Prasada, Hennefield, & Otap, 2012). For instance, a type of viral infection in humans and birds that leads to flu symptoms would have been considered an arbitrary category before 1997. However, the discovery of a common viral cause resulted in the establishment of this grouping as a true kind called avian flu. We postulate that since avian flu came into existence, people now believe that it has a particular causal essence, which causes its surface features and is necessary and sufficient for an instance of the flu to be referred to as avian flu.

We also suggest that even when people do not know precisely what these essences are (Medin & Ortony, 1989), they may serve as "placeholder" explanations for why kinds are coherent and real. That is, a person could infer that a certain category must have a causal essence simply because it is accepted as a kind; this essence, whatever it is, is necessary for explaining its coherence. For instance, in the early 1900s it might have appeared to be a pure coincidence that a group of people tended to display "unstable interpersonal relationships, affective distress, marked impulsivity, and unstable self-image". Yet, this is now a category listed in the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., *DSM–IV*; American Psychiatric Association, APA, 2000) as borderline personality disorder. In the case of borderline personality disorder, however, the underlying cause is still unknown. But the fact that it has been accepted in an official manual might be sufficient to encourage people to assume that there must be a shared causal essence. The current study tests this possibility by experimentally manipulating whether

categories are true kinds or arbitrary categories and observing how such manipulations influence causal essentialism. We predict that simply believing that a certain category is a kind would be sufficient to trigger causal essentialism.

The current study also examines whether the assumption that a category is a kind leads to belief in shared causal mechanisms among category members. Recent studies (e.g., Ahn, Kim, Lassaline, & Dennis, 2000; Hadjichristidis, Sloman, Stevenson, & Over, 2004; Lassaline, 1996) have demonstrated that causal relations between category features determine which features are considered more central and immutable and which features are more projectable. A crucial premise behind these studies is that members of the same category share causal relations, and it is these shared causal relations that determine feature centrality or property projection. Surprisingly, however, few studies have verified the psychological validity of this premise.

We propose, then, that people believe that an underlying essence governs how the features of a category are causally connected to each other, and because category members share a causal essence, they would also share similar causal relations. While previous theorists have argued that essentialized categories are perceived to be homogeneous with respect to individual features (Gelman, 2003; Medin & Ortony, 1989; see also Gelman & Markman, 1986; Haslam et al., 2000; Yzerbyt, Corneille, & Estrada, 2001; for empirical demonstrations), few have assumed that they would be also homogeneous in terms of causal structures. The basis of our proposal is the idea that surface features caused by the different essences may be perceived to have different causal implications. For example, people may believe that vegetables that are genetically modified lead to fewer health benefits and taste worse than vegetables that are naturally produced. Even if both types of vegetables are in fact chemically identical, their causal implications may be thought to differ if they have different causal essences (for supporting evidence, see Rozin et al., 2004). Conversely, surface features derived from the same causal essences may lead to beliefs in shared causal structures among surface features.

Manipulating kinds versus arbitrary categories

Having explained the core distinction between kinds and arbitrary categories, and the implications of that distinction for causal essentialism and beliefs in shared causal mechanisms, we now discuss how kinds and arbitrary categories are experimentally manipulated in our study. This discussion will also clarify features that would or would not distinguish between kinds and arbitrary categories.

In our experiments, we developed a number of artificial categories consisting of three features (see Figure 1 for sample stimuli), which are used for both kinds and arbitrary categories. To manipulate a certain category as a kind, we simply indicated it using a known superordinate kind. For instance, participants were told, "There is a kind of *animal* called an egoogle", "There is a *mental disorder* called BLV", where italics, although not used in the experimental materials, indicate the known superordinate kinds. In contrast, arbitrary categories are not generally accepted among people and could have been constructed by a person on an idiosyncratic basis. In order to make this feature explicit in our manipulations for Experiments 1 and 2, we used the same category features as the ones used for kinds, but designated a nonexpert of a domain (someone who is highly unlikely to possess generally accepted or valid knowledge of the domain) as an inventor of the category. For example, we stated for an arbitrary category, "A high school student was searching for animals on the web using the Google search engine. He labeled a group of animals displayed on the even-numbered pages 'egoogles'". For Experiments 3 and 4, we indicated arbitrariness of categories by noting that the shared features are mere statistical co-occurrence (see Item 4 of Figure 1).

Note that some dimensions do not necessarily distinguish between kinds and arbitrary categories. First, as shown in Figure 1, both kinds and arbitrary categories can be constructed based on correlated

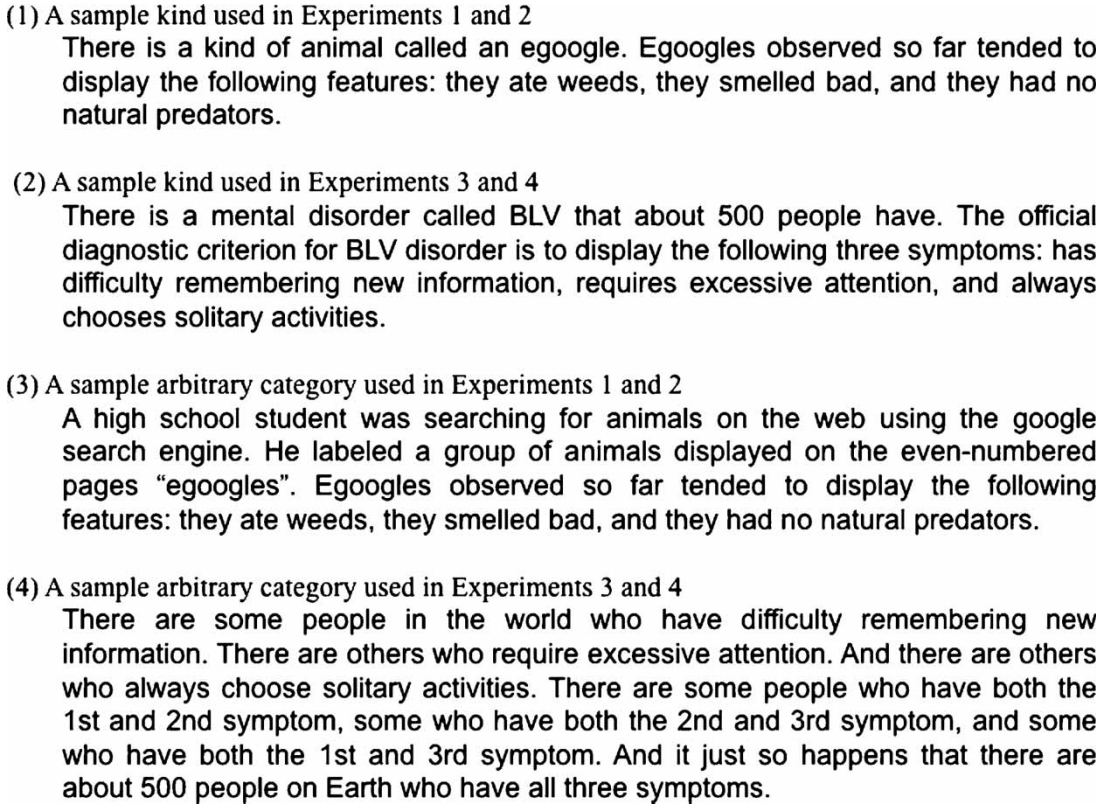

Figure 1. *Sample stimuli used in Experiments 1–4.*

features. Correlated features can lead people to infer the existence of a common cause (a causal essence) because such featural co-occurrence would otherwise be too much of a coincidence (e.g., Gelman, 2003; Markman, 1989). Still, in order to demonstrate that the differences in causal essentialism can be obtained merely due to differences in the kind versus arbitrary category distinction, we used the same correlated features for both kinds and arbitrary categories.

Second, we used labels (e.g., "egoogles") for both kinds and arbitrary categories in Experiments 1 and 2. Past research has shown effects of labels on essentialist beliefs (e.g., Yamauchi, 2005). As shown in Figure 1 (Items 1 and 3), however, the labels were used in two different ways so that the core distinctions between the two types of categories would be intact. For arbitrary categories the labels (e.g., egoogles) were shorthand for the criteria (e.g., even number of pages of Google search) so that we can retain the arbitrariness of the categories. For nonarbitrary kinds the labels were purely nominal and have no meaning as in other existing kinds. We predict that labels per se would not be responsible for differences in causal essentialism between the two types of categories, so the predicted differences would be obtained even though labels are used for both versions. (But see the General Discussion for further discussion of effects of different types of labels.)

Third, in previous research natural kinds tend to be equated with real categories that need to be discovered in nature, whereas artefacts or conventionally established categories tend to be equated with nominal kinds (e.g., Schwartz, 1979). Yet, the distinction between kinds and arbitrary categories is not

necessarily based on whether categories or category members are natural or human-made (except that by definition arbitrary categories must be invented based on arbitrary criteria). For instance, arbitrary categories can consist of animals (e.g., animals displayed on the even-numbered pages of Google search results). Also, nonarbitrary kinds can consist of human-made objects (e.g., chairs) and can be socially constructed (e.g., bachelors). For these reasons, in Experiments 1 and 2 we used a variety of domains for both kinds and arbitrary categories.

Predictions and overview of experiments

To recapitulate, we argue that people's notions of kinds include beliefs in shared causal essences and causal mechanisms. Thus, we predict that merely knowing that a category of certain things is a kind is sufficient to increase people's endorsement of causal essences and generalization of causal mechanisms across members of the category.

Throughout four experiments, we asked participants to make judgements about kinds and arbitrary categories. Arbitrary categories were constructed by indicating that nonexperts invented those categories based on arbitrary criteria (Experiments 1 and 2) or statistical coincidence (Experiments 3 and 4). Then, we measured the extent to which people explicitly endorsed a causal essence (Experiments 1 and 3) and the extent to which they generalized the presence or absence of causal relationships found in a category member to all other category members (Experiments 2 and 4).

We predicted that causal essentialism would be observed in kinds and would be greatly attenuated for arbitrary categories even though they shared the same labels (Experiments 1 and 2) and correlational structures (Experiments 1–4). We predicted that this finding would be true across various domains.

EXPERIMENT 1

Novel categories were developed from each of four domains: mental disorders, medical disorders, living things, and artefacts. For each category within a domain, we developed two versions: a nonarbitrary kind version and an arbitrary category version. The goal of Experiment 1 was to empirically validate our intuition by asking participants to judge the extent to which they ascribe a causal essence to each category. That is, we tested whether merely mentioning a category as a kind is sufficient to trigger causal essentialism, whereas describing a category as constructed using a nonexpert's criterion is not.

Method

Thirty-one Yale University undergraduates participated in this experiment in partial fulfilment of an introductory psychology course's requirements or for monetary compensation. Of these, 16 participated only in this experiment, and the other 15 participated in this experiment after completing other categorization experiments. There was no significant interaction effect involving these two groups of participants, so all data are collapsed in the following analyses.

Two categories, each consisting of three characteristic features, were developed from each of four domains: mental disorders, medical disorders, living things, and artefacts (see Appendix A for a complete list of features for eight categories). The eight categories were divided into two sets (i.e., Set 1 and Set 2), with each set containing one category from each of the four domains. From Appendix A, Set 1 contained FFL (mental disorder), SS7 (medical disorder), egoogle (living things), and notodd (artefact), and Set 2 contained the rest of the categories in Appendix A. Half of the participants received kind versions of Set 1 and arbitrary category versions of Set 2, whereas the other half received kind versions of Set 2 and arbitrary category versions of Set 1. Presentation of kind/arbitrary category versions was blocked and counterbalanced such that half of the participants rated kinds first, and the other half rated arbitrary categories first. Within each block, the order of the four categories was randomized across participants.

For each category, participants were asked whether they believed that there was a causal essence (i.e., defining feature that also causes the other features of the category members), whether or not they knew what that essence might be. For instance, for FFL, they were asked, "*Do you think there is something that is shared by all and only FFL patients that also causes the other features of FFL patients (whether or not we know what that thing is)*?"[1]

Participants answered each question by typing a number on an 8-point scale ranging from 1 (strongly disagree) to 8 (strongly agree). The experiment was programmed using the RSVP computer program (Williams & Tarr, 2001) and run on Macintosh computers. Before starting the task, participants were instructed, "Please read the descriptions of each given category carefully before you make a judgement. There are no right or wrong answers in this task; we are simply interested in your own thoughts. When you arrive at a decision, please enter your response in the box at the top of the screen". Participants proceeded through the experiment at their own pace.

Results and discussion

There was no effect of presentation order so all results are reported with data collapsed across the two orders. As predicted and shown in Figure 2, participants more strongly endorsed causal essences for the kind versions ($M = 5.77$, $SD = 1.07$) than for the arbitrary category versions ($M = 3.65$, $SD = 1.88$), although both versions contained identical sets of characteristic features. A 2 (category type: kinds versus arbitrary category) × 4 (domains) repeated measures analysis of variance (ANOVA) found a significant main effect of category type, $F(1, 30) = 41.32$, $p < .01$, $\eta^2 = .58$. There was no significant main effect of domain, $p > .50$.

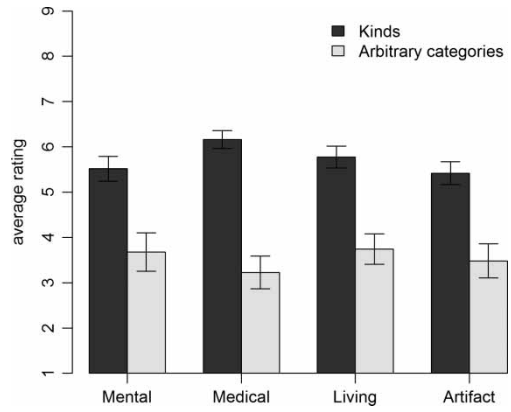

Figure 2. *Mean essentialism ratings of kind and arbitrary category versions across domains. Error bars are standard errors of the mean.*

The interaction between domain and category type was significant, $F(3, 90) = 3.86$, $p < .05$, $\eta^2 = .11$. Pairwise t tests found that this interaction effect was because the effect of category type was larger in the medical domain than in the other three domains, all ps $< .05$, presumably because of lay medical knowledge about essences (e.g., viruses). No other between-domain differences were found. In fact, the difference between the kind and arbitrary category conditions was in the same direction in all four domains, with all ps $< .001$. Furthermore, the mean ratings for the kind versions of all four domains were significantly greater than the midpoint of the scale (4.5), all ts $>$ 3.87, all ps $< .001$, indicating that participants endorsed causal essences for all four domains.

One limitation of Experiment 1 is that it is not clear whether the significant differences between the two types of categories were obtained because kinds are believed to be homogeneous, as we have argued, or because the arbitrary categories' criteria could have discounted the likelihood of an essence. To explain, note that for arbitrary categories the common possession of an arbitrarily selected, nominal criterion determines membership

[1] One may be concerned that asking about judgements of defining features and causality may be too demanding for participants, and they may have paid attention only to the first part of the questions (i.e., judgements of defining features). If so, this would work against our hypothesis because both kinds and arbitrary categories contain defining features (labels in this experiment). In addition, Ahn, Flanagan, Marsh, and Sanislow (2006) used separate questions to specify different aspects of essentialism (albeit with less controlled stimuli) and found results similar to the current study.

and, thus, explains why all members are included in the category, regardless of what features members happen to have in common. The explanation provided by the arbitrary defining criterion for why the category includes the members it does could have ruled out the need to posit the existence of some common, underlying causal essence. Thus, the mechanism here is akin to the well-known discounting principle where a known cause (e.g., being invented based on a certain criterion) discounts the likelihood that an alternative cause (e.g., an essence) is also true (Kelley, 1972).

On one hand, the use of arbitrary criteria to explain the grouping for arbitrary categories was practically unavoidable because participants would be highly unlikely to believe that a group of, say, animals, sharing three features are not the same species of animals. That is, it was methodologically needed in order to make the arbitrary categories truly arbitrary. Yet, to further ensure the generality of the effects of kinds, Experiments 3 and 4 use different stimuli for arbitrary categories, which do not involve explicitly stated arbitrary criteria.

EXPERIMENT 2

Experiment 2 used the same stimuli as those from Experiment 1 and tested the generalizability of causal mechanisms observed in a category member (either a causal chain or lack of any causal relations) across all other members in the same category. For instance, given the egoogle description shown in Passages 1 and 3 in Figure 1, participants in the causal condition were further told that in a particular egoogle, it was found that the three features formed a causal chain (e.g., "the fact that it eats weeds causes it to smell bad, and because it smells bad, it has no natural predators"). Participants in the noncausal condition were told that these features are completely unrelated to each other. Then, all participants were asked to estimate the likelihood that other members of the same category would display the same pattern of causal relationships or lack of causal relationships. We predicted that participants would generalize causal mechanisms more with true kinds than with arbitrary categories.

If generalization of causal structure depends on the type of categories, it would provide novel insights into the causal learning literature. Traditionally, covariation has been considered one of the most important cues to causality (e.g., Cheng, 1997). Given that both kind and arbitrary categories in our experiments contain identical patterns of correlation among features, there is no reason why a known causal structure would be more or less likely to be generalized to other exemplars simply because of the category that the exemplars belong to. Nonetheless, as mentioned in the introduction, we predict that causal essentialism, more likely to be present in kinds, would determine people's willingness to generalize causal structure within a category. The reason for such a prediction is that features have different causal implications depending on what caused them. As alluded to earlier, naturally produced foods are believed to be safer, healthier, and tastier than foods that are created or modified through human intervention, even if people are told that the natural food is chemically identical to the artificial one (Rozin et al., 2004). Intended crimes are treated differently from the identical acts of crime that were not intended (Weiner, 1986), most likely because of differences in the causal implications for the future. Thus, if surface features are believed to share the same underlying causal essences, they may also be believed to share other causal relations as well.

Method

Ninety-six people from Amazon's Mechanical Turk website (https://www.mturk.com) participated for $1.50. The benefits and reliability of experimental data collected from Mechanical Turk have been previously documented (Paolacci, Chandler, & Ipeirotis, 2010).

The two versions (kind versus arbitrary category) of the eight categories used in Experiment 1 were used. As in Experiment 1, for each scenario, participants first learned whether or not a category is a kind or an arbitrary category (e.g., the first row of light-grey boxes in Figure 3), followed by the three characteristic features of the category (e.g.,

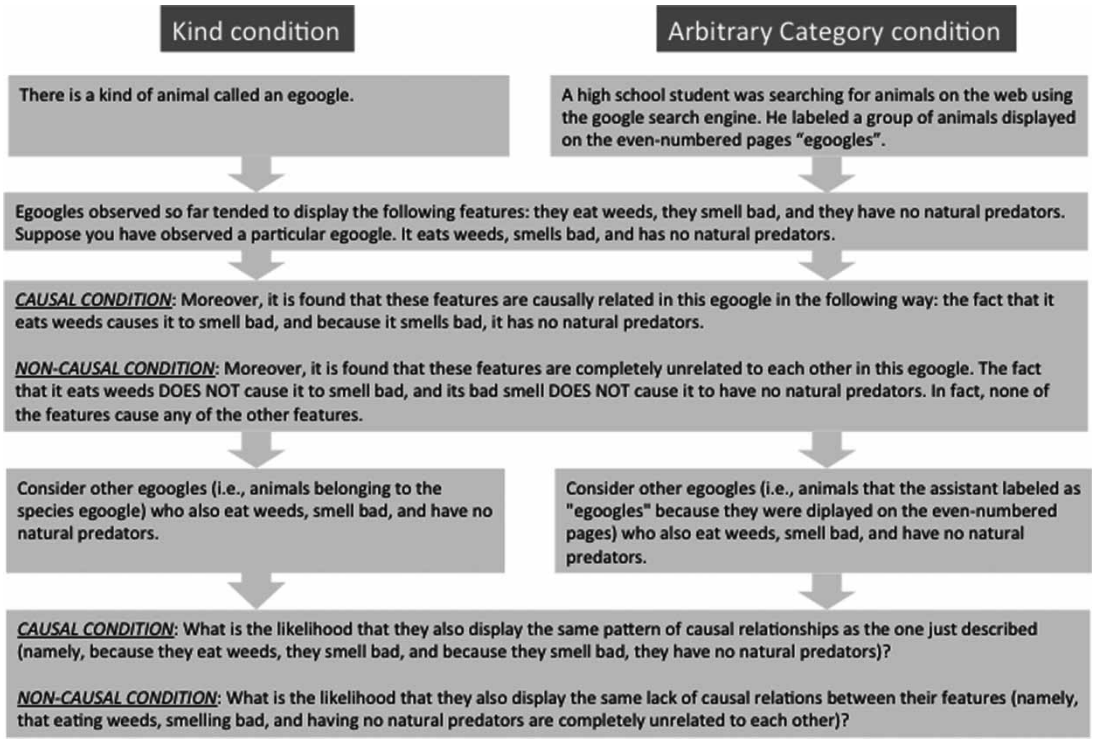

Figure 3. *Each of the four versions of the egoogle category used in Experiment 2. The versions differed in (a) whether the category was a kind or an arbitrary category, and (b) whether or not the features of the category member singled out were described as causally related. Versions of the other categories differed similarly.*

the second light-grey box in Figure 3). Then, one member from the group who had all three symptoms was singled out, and its causal mechanism was described (e.g., the third light-grey box in Figure 3). In the "causal" versions, this member's features were described as forming a causal chain. In the "noncausal" versions, the features were described as causally unrelated to one another. Participants were then told to consider other members of the same category (e.g., the fourth row in Figure 3), and to judge the likelihood that they would also display the same causal pattern or the same lack of causal relations (e.g., the last row in Figure 3) on a 9-point scale (1: *very unlikely* to 9: *very likely*). Thus, the design of Experiment 2 was a 2 (category type: kind or arbitrary category) $\times$ 2 (causal versus noncausal conditions). Although all four domains of categories were used across participants, this factor was not the main interest in this study and was not fully crossed with the two main independent variables, as explained below.

As in Experiment 1, the eight categories were divided into Set 1 and Set 2. Half of the participants received the kind versions of Set 1 and the arbitrary category versions of Set 2, whereas the other half received the kind versions of Set 2 and the arbitrary category versions of Set 1. As in Experiment 1, presentation of kind/arbitrary category versions was blocked and counterbalanced.

Within each block/set, two of the items were causal descriptions, and two were noncausal. The presentation of items within a set was blocked by the causal versus noncausal factor. Half of the participants received the causal description for the medical disorders and living things domains and the noncausal description for the mental disorders

and artefact domains. The other half received the opposite description type for each domain. To clarify with an example, one participant received the causal versions of the medical disorder FFL and the animal *egoogle*, but the noncausal versions of the mental disorder SS7 and the artefact *notodd*. Another participant received the opposite description type for each of these categories.

Results and discussion

The results differed as a function of block, so we restricted our analyses to the data from the first block only. As a result, the kind versus arbitrary category became a between-subjects variable.

Figures 4a and 4b show the mean generalization ratings separated by domain, for the causal and noncausal items, respectively. As predicted, participants were more likely to generalize causal relations for kinds ($M = 8.07$, $SD = 1.15$) than for arbitrary categories ($M = 6.35$, $SD = 2.16$). They were also more likely to generalize the *lack* of causal relations for kinds ($M = 6.48$, $SD = 2.00$) than for arbitrary categories ($M = 5.21$, $SD = 2.20$).

A 2 (category type: kind versus arbitrary category) × 2 (causality: causal versus noncausal conditions) mixed ANOVA was conducted with category type as a between-subjects variable and causality as a within-subjects variable. The main effect of category type was significant, $F(1, 94) = 32.40$, $p < .01$, $\eta^2 = .26$, supporting the hypothesis that nonarbitrary kinds would lead to more causal generalization than arbitrary categories. The main effect of causality was also significant, $F(1, 94) = 59.06$, $p < .01$, $\eta^2 = .38$, suggesting that people were overall more likely to generalize the causal chain explanation ($M = 7.17$, $SD = 1.94$) than the lack of any causal relations ($M = 5.82$, $SD = 2.20$). The interaction between category type and causality was not significant, $F(1, 94) = 1.61$, $p = .21$, $\eta^2 = .01$.

Note that our design allows for a within-subjects test of the main effect causal versus noncausal, but not a within-subjects test of the interaction between causality and domain. An item analysis testing the interaction between causality and domain would also be limited since there are only a small number of items. We simply note that the differences between kinds and arbitrary categories are all in the same direction across all four domains as shown in Figures 4a and 4b.

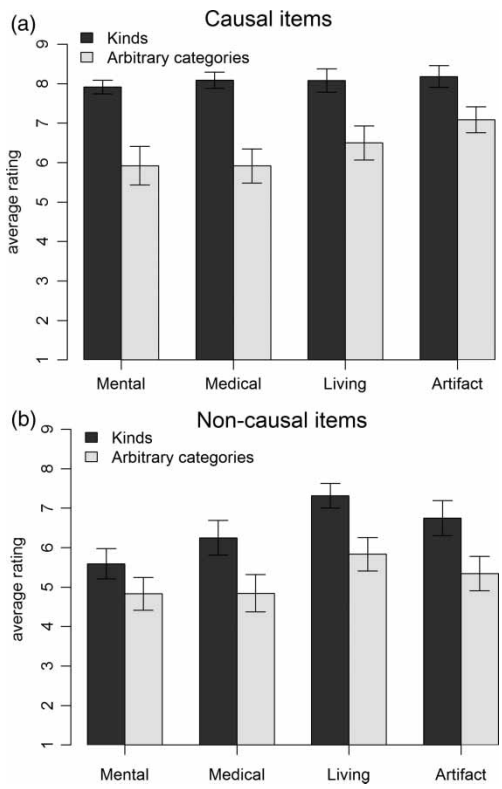

Figure 4. *Mean generalization ratings of kind and arbitrary category versions for the four domains, separately for the (a) causal and (b) noncausal items. Error bars are standard errors of the mean.*

EXPERIMENT 3

So far, we have found that people are more willing to endorse a causal essence and generalize a causal mechanism of a single member to other members in a kind than in an arbitrary category. As discussed earlier, however, one may argue as we do that these differences indicate an effect of nonarbitrary kinds (i.e., a kind being more homogeneous in terms of causal structure), or alternatively, an effect of the arbitrary categories we used (i.e., the presence of

arbitrary category criterion, such as last name beginning with an F) reducing the homogeneity of categories in terms of causal structure. In order to rule out the second possibility, we used a different strategy in Experiments 3 and 4. To explain this new strategy, we first explain the methodological difficulties in developing arbitrary categories below.

The reason why an arbitrary category criterion was specified in the arbitrary category conditions of Experiments 1 and 2 was that without such an arbitrary criterion, people may automatically assume that the category is a true kind, since it is indeed too much of a coincidence that a group of things from a common superordinate category shares three features. Such a tendency would be particularly strong for living things or artefacts, since surface similarities strongly imply underlying shared essences (Medin & Ortony, 1989). In order to undermine such automatic inferences, we spelled out the arbitrary category criterion for participants.

In Experiments 3 and 4, we instead restricted our stimuli to the mental disorder domain, because this domain appears to be a plausible case in which a group of people can share three characteristic features without necessarily having to share an actual known category membership. That is, while people may easily believe that two plants that both *have sticky leaves, absorbed iron,* and *produced red flowers* are the same kinds of plants, two people who both *have difficulty remembering new information, require excessive attention,* and *always choose solitary activities* may not necessarily suffer from the same disorder because each of these symptoms could be caused by different disorders or no disorder at all. Thus, the domain of mental disorders allows us to create arbitrary categories without resorting to describing arbitrary category criteria. Furthermore, restricting the stimuli to the domain of mental disorders does not appear to greatly compromise the generality of the results, as we did not find significant domain differences between the mental disorder categories and living things or artefacts in Experiments 1 and 2.

Figure 1 shows a sample stimulus used in this experiment. In order to make the arbitrary categories as arbitrary as possible, we did not use a category label. Also, to make it clear that sharing three symptoms does not necessarily mean having the same mental disorder, we created the passage emphasizing that some people have only one of the three symptoms, some have two of the three symptoms, and finally, there may be a group of people who happen to have all three symptoms. In addition, we specified the sample size for both kinds and arbitrary categories to be 500 people so that the potential inferred category size was not a confound.

The symptoms used in Experiment 3 (as well as Experiment 4) were developed by Ahn, Novick, and Kim (2003) and are shown in Appendix B. Within each category, the three symptoms are taken from three different existing mental disorders according to the *DSM–IV* (APA, 2000) so that they are highly unlikely to activate concepts of existing mental disorders. This is an important control for the arbitrary category condition. To further ensure the arbitrariness of the collection of the three symptoms for the manipulation of arbitrary categories, the symptoms we used were previously judged to be unlikely to have causal relations among them (Ahn et al., 2003). In particular, this second measure was crucial in avoiding ceiling effects for Experiment 4 where generalizability of causal relations among symptoms was judged.

Method

Twenty people from Amazon's Mechanical Turk website (https://www.mturk.com) participated for $1. For stimuli, four sets of three characteristic symptoms were taken from the implausible causal condition of Experiment 1 in Ahn et al. (2003; see Appendix B for the symptoms). Each set of symptoms was used to create two versions, which differed in whether or not the characteristic symptoms were described as being diagnostic of a known mental disorder (see the bottom of Figure 1 for an example).

All items began with a description of the characteristic symptoms. As before, in the kind versions, individuals were described as sharing a known mental disorder (e.g., "There is a mental disorder

called BLV that about 500 people have"). In the arbitrary category versions, the symptoms were described but there was no mention of a mental disorder. Instead, participants were walked through how there can be a group of people, all of whom happen to share three symptoms (see Figure 1). After participants read the description of each category, they made a causal essence judgement. Specifically, they were asked, "What is the likelihood that there is a single cause underlying these three symptoms that all and only [these individuals] have (whether or not we know what that cause is)?". Ratings were made on a scale of 1 (highly unlikely) to 9 (highly likely).

Each participant viewed all four sets, but which version was given for each set was counterbalanced. Ten participants received the kind versions of BLV and YNA shown in Appendix B and the arbitrary category versions of the other two sets. The remaining 10 participants received the opposite versions of each disorder. Within these groups, half of the participants viewed the kind items first, and the other half viewed the arbitrary category items first. Within each block (i.e., the kind or the arbitrary category), the order of the items was determined randomly for each participant.

Results

The results did not differ as a function of block, so the results were collapsed across the two orders. As predicted, participants were more likely to attribute a causal essence to individuals if the category was a kind ($M = 4.98$, $SD = 1.67$) than to individuals with the same symptoms but without any indication that these individuals share a proper mental disorder ($M = 3.83$, $SD = 1.71$), $t(19) = 2.94$, $p < .01$, $d = 0.66$.

EXPERIMENT 4

Using the stimuli developed for Experiment 3, Experiment 4 tested whether participants were more likely to generalize the causal structure (either the causal chain or no causal relations) observed in a single category member across the entire category as in Experiment 2.

Method

+Seventy-four people from Amazon's Mechanical Turk website (https://www.mturk.com) participated for $1. The same mental disorder symptoms and categories as those from Experiment 3 were used. As in Experiment 2, the design was a 2 (category type: kinds versus arbitrary category) × 2 (causal versus noncausal conditions). The manipulation of the category type was the same as that in Experiment 3. The manipulation of causal versus noncausal conditions was the same as that in Experiment 2. Thus, after reading about a given category, participants read about an individual who belonged to the category. The individual's symptoms were either causally related, forming a causal chain (causal condition), or were not related to each other at all (noncausal condition; See Figure 5 for sample stimuli). Then, participants rated the likelihood that others with same three symptoms displayed the same pattern of interfeature causal relationships (either the causal chain or the lack of any causal relations). Ratings were made on a scale of 1 (highly unlikely) to 9 (highly likely).

Each participant viewed all four categories, but each with a different version. Thirty-eight participants received the kind versions of BLV and YNA and the arbitrary category versions of the other two sets. The remaining 36 participants received the opposite versions of each disorder. Within these groups, half of the participants received the causal version of BLV and FFL and the noncausal version of YNA and SSJ. The other half received the opposite versions of each disorder.

The items were blocked by the category type factor. Thirty-eight participants received the kind items first and the arbitrary category items second. The remaining 36 participants received the items in the reverse order. Items within blocks were always presented in the following order: BLV before FFL, and YNA before SSJ, such that half of the participants received the

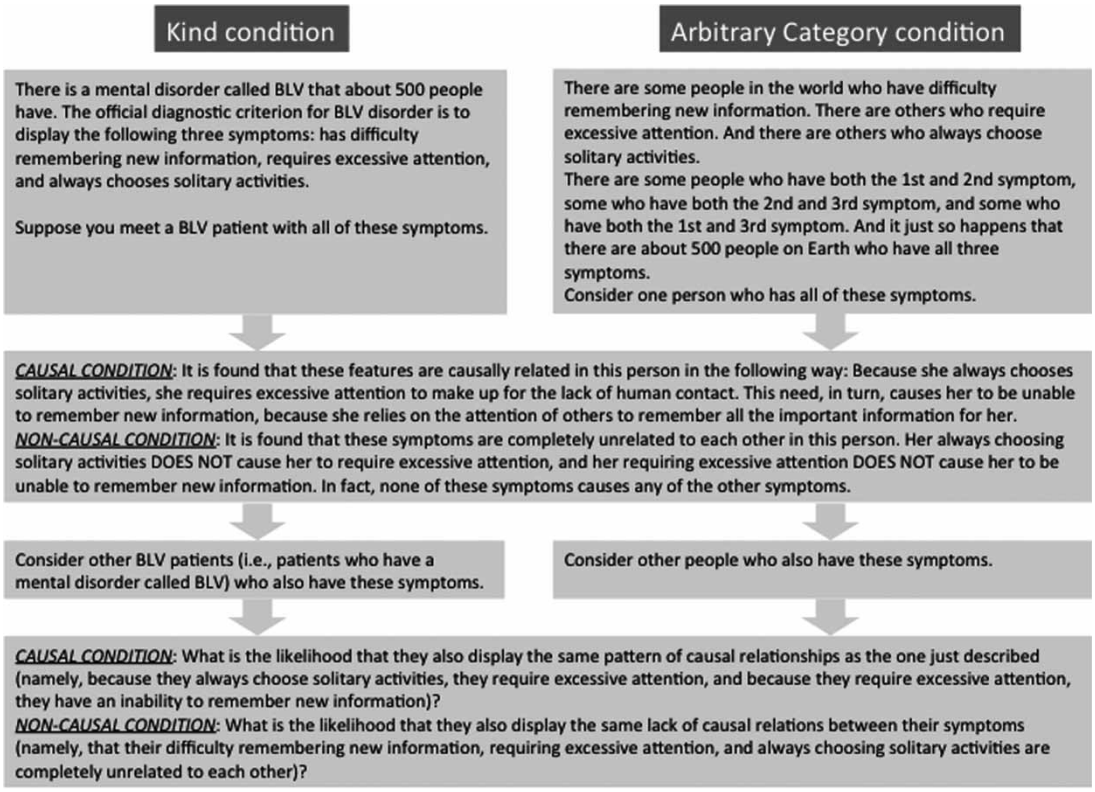

Figure 5. *Each of the four versions of the BLV category. The versions differed in (a) whether or not the individuals were described as sharing a known mental disorder, and (b) whether or not the symptoms of the individual singled out were described as causally related. Versions of the other categories differed similarly.*

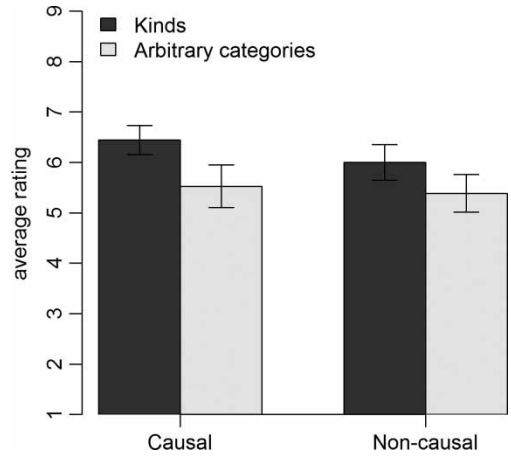

Figure 6. *Mean generalization ratings for the descriptions in Experiment 4. Error bars are standard errors of the mean.*

causal items first, and the other half received the noncausal items first.

Results and discussion

The results differed as a function of block, so we restricted our analyses to the data from the first block only, making category type a between-subjects factor. As predicted and shown in Figure 6, participants were more likely to generalize causal relations for individuals with a common mental disorder ($M = 6.45$, $SD = 1.77$) than for individuals with the same symptoms but no known mental disorder ($M = 5.53$, $SD = 2.56$). They were also more likely to generalize the *lack* of causal relations for individuals with a common mental disorder

($M = 6.00$, $SD = 2.18$) than for individuals with no known disorder ($M = 5.39$, $SD = 2.23$).

A 2 (category type) × 2 (causal vs. non causal) mixed ANOVA with category type as a between-subjects variable and causality as a within-subjects variable revealed a significant main effect of category type, $F(1, 72) = 4.16$, $p = .045$, $\eta^2 = .05$, supporting the hypothesis that generalization would be higher for individuals with a known mental disorder. There was no significant main effect of causality, $F(1, 72) = 0.74$, $p = .39$, $\eta^2 = .01$, and no significant interaction, $F(1, 72) = 0.20$, $p = .66$, $\eta^2 = .003$, meaning that the effect of category type held for both generalizations of causal structure and the lack of causal structure.

GENERAL DISCUSSION

Throughout four experiments, we found that when categories are described as true kinds, people are more willing to endorse causal essentialism—that members share some underlying essence that is both necessary and sufficient for category membership and that causes surface features. In addition, they were more willing to generalize a member's known causal relations or lack thereof to other members of the same kind. Such inferences were much weaker for arbitrary categories. These differences between kinds and arbitrary categories were found across various domains.

Implications for psychological essentialism

These findings lend support to psychological essentialism by providing the first empirical evidence that people do explicitly ascribe causal essences to kinds across a variety of domains. These findings challenge Strevens's (2000) minimal hypothesis. Strevens has argued that previous findings that people categorize based on deeper features such as intentionality or heritage (e.g., Gelman & Bloom, 2000; Gelman & Wellman, 1991) do not necessarily require postulation of an essence. Instead, he claims that these results could be fully explained by a minimal hypothesis that there is something that causes surface features, which does not have to be an essence in that it does not have to be a defining feature of the category. (See also, Hampton, Estes, & Simmons, 2007; Pothos & Hahn, 2000.) Contrary to this claim, however, our participants did overtly endorse causal essences for kinds.

We also found such endorsement across various types of kinds—not only with living kinds, but also with artefacts. Previous studies (e.g., Diesendruck & Gelman, 1999) showed that categorization of artefacts is more graded than that of living kinds, suggesting that artefacts are believed to lack defining features. (See also, Hampton et al., 2007; Pothos & Hahn, 2000.) However, our results show that people seem to believe that many types of kinds, including artefacts, have a necessary and sufficient essence that determines surface features (see also Bloom, 1996, 2004). This apparent contradiction is also present in the findings by Brooks, Squire-Graydon, and Wood (2007), where people endorsed defining features for family-resemblance categories, which clearly lacked defining features.

One cautionary remark is that although our findings replicated across several domains, we are not claiming that all kinds are equally essentialized. There are at least two reasons for this precaution.

First, psychological essentialism is a multifaceted claim. For instance, Haslam et al. (2000) measured five dimensions associated with essentialism (naturalness, stability, discreteness of category boundaries, immutability, necessity of category features) and found large variance among social categories (e.g., gender, age, occupation, politics) along these measures. We only measured belief in causal essences defined as a common feature that causes other surface features, and we found similar effects across four domains often discussed in the literature. However, there may be differences in our essentialism scale for other domains that we did not examine, or there may be differences along other dimensions of essentialism that Haslam et al. (2000) described.

Second, we used artificial categories, which allowed us to strictly equate the number of shared features within each category. We found that when controlling for surface similarities,

information about domains did not seem to have an effect powerful enough to lead to differences in inferring causal essences (with an exception of the results from the medical disorders in Experiment 1). Nonetheless, real-life categories may vary in homogeneity in terms of shared features, and differences in surface similarities can lead to different degrees of essentialism across different domains.

Real-life implications

In addition to making novel empirical contributions in terms of theories of psychological essentialism, the current findings have numerous real-life implications. In a nutshell, the current studies demonstrate the assumptions that people would make upon learning that a new, real category has come into understanding and has become lexicalized, which is a strong sign that a category is broadly accepted as appropriate. For instance, nearly 100 new words are added to Merriam-Webster Collegiate Dictionary each year, including such words as "mouse potato" and "avian influenza" for the year 2006, and "green-collar" and "webisode" for the year 2009. As a category transforms from an arbitrary category status to a true kind, users may spontaneously infer that the category has a causal essence and that members share similar causal structures. (See Dar-Nimrod & Heine, 2011, for a recent review of consequences of "genetic essentialism" on prejudice and stereotypes involving social categories and mental illnesses.)

To make this implication more concrete, consider the classification system of mental disorders, the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*). Since the manual came out in 1952, the number of mental disorders grew from about 60 in the first version (APA, 1952) to over 400 today (APA, 2000; see also Houts, 2002). For instance, posttraumatic stress disorder was not officially recognized as a mental disorder until 1980 (*DSM–III*; APA, 1980). The next version of the DSM is expected to come out in 2013, and there are proposals for more new mental disorders, such as premenstrual dysphoria, attenuated psychosis syndrome, and catatonic disorder, to name a few (American Psychiatric Association, n.d.).

When a new mental disorder is added, people may subsequently make assumptions about the disorder and members of that category. The current study suggests that people are more likely to believe that the symptoms result from a single cause, that all patients with the disorder have this cause, and that the causal relations among symptoms are similar among patients with these disorders. If so, whether these assumptions are actually empirically warranted is an important issue to be considered by the committee that determines what mental disorders are to be included in the manual.

Issues for future research

There are a number of open questions remaining for future research. In the current Experiments 1 and 2, we used labels for both arbitrary categories and kinds, but as discussed earlier, the relationship between the labels and the categories differed depending on the condition. The label–category relationship was strictly nominal for kinds such that the labels were not linked to any specific features of the categories, whereas for arbitrary categories, the labels were a kind of abbreviation of the arbitrary criteria used for constructing the categories (e.g., egoogle for even number pages of Google search). The reason why we chose this approach for arbitrary categories is that if we had used nominal labels, it may have signalled to the participants that the categories were kinds.

This issue leads to a more theoretically interesting question of what factors would turn a certain grouping into a kind. As discussed above, nominal labels may signal that a grouping is a kind. Another factor that was incorporated in the current experimental manipulation was the expertise of the inventor of the categories; had the creator of the categories been a true expert of the domain, even arbitrary categories may have been thought of more so as kinds. Future research can examine more systematically which factors that were manipulated in the current study to

distinguish between kinds versus arbitrary categories are necessary or sufficient for the effect, and whether there are any other factors that would activate the belief that a grouping is a kind.

Relatedly, correlated features have been suggested as one of the most powerful cues to a common cause (an essence), and, thus, highlighting correlated structures may readily induce the belief that a grouping is a kind. In the current study where all categories contained correlated features, we were concerned about the possibility of inference to kinds based on correlated features so that we ended up with longer descriptions for arbitrary categories than for kinds, as we chose to explicate at length the arbitrariness of a category in order to prevent the inference. While we do not believe that the differences in lengths would necessarily have been a confound responsible for the observed differences in causal essentialism, it would be interesting to examine the validity of our original concern—namely, whether the extent to which features are correlated may moderate not only beliefs in causal essences, but also beliefs in kinds.

Another open question is which categories—arbitrary categories or kinds—serve as a baseline for causal essentialism. It is possible that people as a default believe that categories have essences until they are provided evidence to the contrary. In this way, the differences we found between kinds and arbitrary categories in causal essentialism ratings may be because the use of an arbitrary distinction lowers essentialism beliefs. Alternatively, people may believe by default that categories lack a causal essence. In this case, providing the kind labels may elevate essentialism-like responding over baseline. Finally, it is possible that our design moved participants in both directions from a more neutral baseline, with the description of kinds elevating ratings and the description of arbitrary categories lowering ratings. This is an issue for future research to determine how people's default essentialism beliefs interact with the arbitrary status of categories.

In addition, future research can examine the relationship between two sets of results we found with kinds; people believe that members in a true kind tend to share a causal essence and also causal structures. The exact mechanism between these two inferences is unclear. For instance, does the generalization of causal structures depend on inferences to causal essences or does generalizing a causal structure create the inference of a causal essence? Alternatively, these two inferences may be independent of each other and not, as we have assumed in the current research, two characteristics of the same construct, causal essentialism.

Finally, while the current study found that making a category a kind led to causal essentialism, it is possible that inferring a causal essence may make a category more likely to be believed to be a kind. That is, people may sometimes infer causal essences of a category first, which in turn causes them to treat the category as a kind. Similarly, people may notice that members of a certain category share similar causal relations, which may then cause people to believe that the category is a kind. For instance, upon observing that two different groups of patients with somewhat different symptoms respond to the same kind of treatment, one may postulate that these patients actually share the same underlying disorder. This way, causal knowledge would play a critical role in determining what counts as kinds.

Original manuscript received 31 May 2012
Accepted revision received 9 September 2012
First published online 25 October 2012

REFERENCES

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. *Cognition*, *69,* 135–178.

Ahn, W., Flanagan, E., Marsh, J. K., & Sanislow, C. (2006). Beliefs about essences and the reality of mental disorders. *Psychological Science*, *17*, 759–766.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41,* 1–55.

Ahn, W., Novick, L., & Kim, N. S. (2003). Understanding it makes it more normal. *Psychonomic Bulletin and Review*, *10,* 746–752.

American Psychiatric Association (1952). *Diagnostic and statistical manual of mental disorders* (1st ed.). Washington, DC: Author.

American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association. (n.d.). *DSM-5: The future of psychiatric diagnosis*. Retrieved from http://www.dsm5.org/Pages/Default.aspx

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, *60,* 1–29.

Bloom, P. (2004). *Descartes' baby*. New York, NY: Norton.

Brooks, L. R., Squire-Graydon, R., & Wood, T. J. (2007). Diversion of attention in everyday concept learning: Identification in the service of use. *Memory & Cognition*, *35,* 1–14.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104,* 367–405.

Dar-Nimrod, I., & Heine, S. J. (2011). Genetic essentialism: On the deceptive determinism of DNA. *Psychological Bulletin*, *137,* 800–818.

Diesendruck, G., & Gelman, S. A. (1999). Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. *Psychonomic Bulletin and Review*, *6,* 338–346.

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

Gelman, S. A., & Bloom, P. (2000). Young children are sensitive to how an object was created when deciding what to name it. *Cognition*, *76,* 91–103.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23,* 183–209.

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

Hadjichristidis, C., Sloman, S. A., Stevenson, R. J., & Over, D. E. (2004). Feature centrality and property induction. *Cognitive Science*, *28,* 45–74.

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

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

Houts, A. C. (2002). Discovery, invention, and the expansion of the modern diagnostic and statistical manuals of mental disorders. In L. E. Beutler & M. L. Malik (Eds.), *Rethinking the DSM. A psychological perspective* (pp. 17–68). Washington, DC: American Psychological Association.

Kalish, C. W. (2002). Essentialist to some degree: Beliefs about the structure of categories. *Memory & Cognition*, *30,* 340–352.

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

Kelley, H. H. (1972). *Causal schemata and the attribution process*. New York, NY: General Learning Press.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22,* 754–770.

Locke, J. (1975). *An essay concerning human understanding*. Oxford, UK: Oxford University Press. (Original work published 1894).

Macnamara, J. (1986). *A border dispute: A place of logic in psychology*. Cambridge, MA: MIT Press.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–196). Cambridge, MA: Cambridge University Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92,* 289–316.

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

Pothos, E. M., & Hahn, U. (2000). So concepts aren't definitions, but do they have necessary or sufficient features? *British Journal of Psychology*, *91,* 239–250.

Prasada, S., Hennefield, L., & Otap, D. (2012). Conceptual and linguistic representations of kinds and classes. *Cognitive Science*, *36,* 1224–1250.

Rozin, P., Spranca, M., Krieger, Z., Neuhaus, R., Surillo, D., Swerdlin, A., & Wood, K. (2004). Natural preference: Instrumental and ideational/moral motivations, and the contrast between foods and medicines. *Appetite*, *43,* 147–154.

Schwartz, S. P. (1979). Natural kind terms. *Cognition*, *7,* 301–315.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition*, *74,* 149–175.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York, NY: Springer-Verlag.
Williams, P., & Tarr, M. J. (2001). *RSVP: Experimental control software for MacOS*. Retrieved June 15, 2000.
Yamauchi, T. (2005). Labeling bias and categorical induction: Generative aspects of category information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 538–553.
Yzerbyt, V., Corneille, O., & Estrada, C. (2001). The interplay of subjective essentialism and entitativity in the formation of stereotypes. *Personality and Social Psychology Review*, *5*, 141–155.

APPENDIX A

Stimuli used in Experiments 1 and 2

Domain	*Versions*	*Descriptions*	*Characteristic features*
Mental Disorder Set 1	Kind	There is a mental disorder called FFL.	chronic feelings of emptiness, excessive devotion to work, and social isolation
	Arbitrary category	An assistant was alphabetizing files of patients with mental disorders. He labeled one group of patients "FFL" because these patients' last names began with F.	
Mental Disorder Set 2	Kind	There is a mental disorder called END-3	hallucinations, fear of impending death, and insomnia
	Arbitrary category	An assistant was organizing the insurance information for patients with mental disorders. He labeled one group of files "END-3" because these patients' insurance number ended with 3.	
Medical Disorder Set 1	Kind	There is a disease called SS7	constricted blood vessels, very high temperature, and runny nose
	Arbitrary category	An assistant was sorting files of a doctor's patients by social security number. She labeled one group of patients "SS7" because these patients' social security numbers ended with 7.	
Medical Disorder Set 2	Kind	There is a disease called YNA	stomach acid buildup, vomiting, and sore throat
	Arbitrary category	An assistant was alphabetizing files of patients with medical diseases. He labeled one group of patients "YNA" because these patients' last names began with Y.	
Living Kind Set 1	Kind	There is a kind of animal called an egoogle.	they ate weeds, they smelled bad, and they had no natural predators
	Arbitrary category	A high school student was searching for animals on the web using the google search engine. He labeled a group of animals displayed on the even-numbered pages "egoogles".	

(*Continued overleaf*)

Appendix A. Continued.

Domain	*Versions*	*Descriptions*	*Characteristic features*
Living Kind Set 2	Kind	There is a kind of plant called a starta.	they had sticky leaves, absorbed iron, and produced red flowers
	Arbitrary category	A high school student is looking through a plant database. She labeled a group of plants "starta" because their biological names started with A.	
Artefact Kind Set 1	Kind	There is a kind of tool called a notodd.	they rotated, generated heat, and emitted light.
	Arbitrary category	A high school student is studying tools from around the world. He added together the digits of each tool's model number and grouped together all the tools whose sum came to an even number. He labeled these "notodds".	
Artefact Kind Set 2	Kind	There is a kind of tool called a scrapsix.	they contained magnets, vibrated, and produced a whistling noise.
	Arbitrary category	A high school student was sorting out tools from a landfill. He grouped together the tools that had a serial number whose digits summed together to be greater than sixty and labeled this group scrapsix.	

APPENDIX B

Stimuli used in Experiments 3 and 4

Disorder names used for kinds	*Characteristic features*
BLV	has difficulty remembering new information, requires excessive attention, always chooses solitary activities
YNA	uncontrollably shouts words at random occasions, pulls their hair out on a frequent basis, lacks the ability to produce facial expressions
FFL	is unable to discard worthless objects, believes that thoughts are placed into their head by others, is unable to concentrate
SSJ	believes that complete strangers are in love with them, has periods of extremely elevated mood, is physically cruel towards animals