Folk Judgments of Normality: Part Statistical, Part Evaluative

Adam Bear (adam.bear@yale.edu)
Department of Psychology, 2 Hillhouse Avenue
New Haven, CT 06520 USA

Joshua Knobe (joshua.knobe@yale.edu)
Department of Philosophy, 344 College Street
New Haven, CT 06511 USA

Abstract

Existing research has emphasized the importance of normality judgments in many aspects of cognition and life (e.g., causal cognition, gradable adjectives, cooperative behavior). Yet little work has explored how people actually come to understand what sorts of things are normal. We argue that people's normality intuitions reflect a mixture of statistical and evaluative considerations. Specifically, we suggest that people's intuitions about what is normal can be influenced by representations both of the *average* and of the *ideal*. We test this idea in three experiments. Experiment 1a demonstrates that explicit judgments of normality reflect this mixture of statistical and evaluative considerations. Experiments 1b and 2 then show that the hybrid notion that comes out in these explicit judgments can also explain people's judgments about gradable adjectives. Taken together, these findings have potential implications not only for normality judgments themselves, but also for the many other mental activities that these judgments impact.

Keywords: normality; moral cognition; experimental philosophy

Introduction

In our daily lives, we frequently need to judge what kinds of things are normal or abnormal. Indeed, existing research has used judgments of normality to explain a diverse array of phenomena (e.g., Egré & Cova, in press; Halpern & Hitchcock, 2014; Rand et al., 2014).

But how do we assess what is normal in the first place? The most obvious view would be that intuitions about normality are based on purely *statistical* considerations. For example, suppose that people are trying to judge what is a normal amount of television for a person to watch in a day. They might simply answer by making an estimate about the average amount of television that people watch in a day.

We propose an alternative view according to which intuitions about normality are driven by not only statistical considerations, but also *evaluative* considerations. That is, people do not assess what is normal by simply considering what is average, but also by considering what is ideal.

An Evaluative Theory of Normality

A large body of research has invoked normality to explain important aspects of cognition and behavior. Philosophers in epistemology have appealed to a framework of "normal worlds" to explain when we are justified in believing certain propositions (Goldman, 1986); behavioral economists have shown how norms of cooperation guide moral behavior (Rand et al., 2014); and linguists have argued that an understanding of normality plays a role in a number of important phenomena, including generics (Nickel, 2008) and progressives (Dowty, 1979).

Nevertheless, most of these research programs have focused on the downstream consequences of cognition about normality rather than the nature of normality itself. Little work has explored how, exactly, the norms themselves are represented in the mind.

One notable exception comes from the study of causal judgment (e.g., Hart & Honoré, 1985). Early work in this area assumed that the relevant notion of normality would be a purely statistical one (Hilton & Slugoski, 1986), but more recently, researchers have argued that one can more accurately predict people's causal judgments if one introduces a notion of normality that is in part evaluative (Halpern & Hitchcock, 2014; Hitchcock & Knobe, 2009; Knobe & Fraser, 2008).

For example, consider a case in which there are two different factors, x and y, that are both necessary for an outcome to arise. Now suppose we ask whether x caused the outcome. In cases of this type, participants tend to regard x as more of a cause to the extent that it is statistically infrequent (Hilton & Slugoski, 1986), but they also tend to regard x as more of a cause to the extent that it is morally wrong (Knobe & Fraser, 2008). It has therefore been suggested that causal judgments might be sensitive to a more general notion of *normality*, where x will be considered more abnormal to the extent that it violates either statistical norms or evaluative norms (Hitchcock & Knobe, 2009).

Drawing on this existing research on causal cognition, we argue, more generally, that people's normality intuitions reflect a mixture of both statistical and evaluative considerations. Specifically, we propose that people's normality intuitions are a function of both the *average* and the *ideal*. On this hypothesis, intuitions about the normal should not be equal either to representations of the average or to representations of the ideal. Instead, a person's intuition about the normal amount should depend on both of these factors, meaning that it would become higher if either her representation of the average or her representation of the ideal were to increase.

We suggest that this idea applies across a broad array of contexts. First, it should apply to people's judgments in a wide variety of different domains (the normal amount of television to watch in a day, the normal percentage of students who are bullied in a middle school, etc.). More importantly, the theory is intended to make a strong claim about how normality is represented in the mind and how it underlies various other aspects of cognition. Thus, our proposal applies not only to people's explicit use of the word "normal," but also more implicit measures that tap into the same cognitive representation.

Gradable Adjectives

We examine participants' normality judgments both by simply asking them what they think are "normal" amounts of various quantities and by using a more implicit and indirect method. Specifically, we look at participants' use of *gradable adjectives*. Consider the distinction people draw between quantities that are "small" and those that are "large." Existing theoretical work on gradable adjectives explains this distinction by positing two key theoretical constructs: first, a *scale* along which any quantity can be assigned a size; second, a *standard* on the scale above which a quantity counts as large (Kennedy, 1999; Kennedy & McNally, 2005). A question then arises, however, as to how the threshold itself is determined.

Existing work has shown that purely statistical considerations do play an important role here (e.g., Barner & Snedeker, 2008; Lassiter & Goodman, 2014). For example, suppose one were trying to determine the standard amount beyond which a number of hours of TV per day counts as large. In doing this, it would clearly be relevant to know the average amount of TV that people watch. However, some intriguing new work by Egré and Cova (in press) suggests that there may be more to the story. For example, when told that 5 out of 10 children died in a fire and 5 survived, people are more inclined to agree that "many children died" than to agree that "many children survived." That is, even though equal numbers of children survived and died in this scenario—in other words, descriptive facts about death and survival were equated—people's evaluative considerations about the desirability of survival versus death seemed to affect their intuitions about "many" (Egré & Cova, in press).

It might be the case that this effect of evaluative content is specific to people's use of gradable expressions in language. But we suggest that this linguistic phenomenon is actually a symptom of a far more general fact about people's representations of normality. Specifically, we argue that judgments of gradable adjectives are based not merely on assessments of the average but on assessments of the *normal*, which is itself affected by representations both of the average and of the ideal. Thus, we hypothesize that participants' use of gradable adjectives will show the same pattern as their explicit use of the word "normal," reflecting a hybrid of statistical and evaluative considerations.

The Present Studies

We begin by examining explicit judgments of normality (Experiment 1a). We then further explore the phenomenon using the more indirect measure of judgments about gradable adjectives, in both a correlational study (Experiment 1b) and a manipulation study (Experiment 2).

Across all these studies, we ask participants about what is statistically average and what is evaluatively ideal and then observe how these variables predict views about normality. In each case, we predict that judgments of normality will be impacted both by statistical judgments and by evaluative judgments.

For each study, we analyze the results using normal linear regression, but we also apply a convex combination model that has just one free parameter. Specifically, we fit participants' responses (N) to the following equation:

$$N = bI + (1 - b)A,$$

where I is the representation of what is normatively ideal and A is the representation of what is statistically average. Intuitively, the free parameter b represents the weight that people attach to the ideal vs. the average. When b is 0, the normal is determined entirely by the average. When b is 1, it is determined entirely by the ideal. Our model predicts that b will have an intermediate value (i.e., the confidence interval for b will not overlap with either 0 or 1), suggesting that the normal is determined in part by the average, in part by the ideal.

To estimate the value of b, we used linear regression. Note that the above equation is equivalent to $(N - A) = b(I - A)$. This equation can then be modeled using simple linear regression without a constant term.

Experiment 1a

In this experiment, we examine how people's intuitions about average and ideal amounts of various behaviors or events relate to what they think are normal amounts of these behaviors and events. Recent work by Wysocki (unpublished data) suggests that people's explicit use of the word "normal" in cases like these may be affected by evaluative considerations. Building on this work, we developed a list of behaviors and events and then used exactly the same method to examine both explicit judgments of normality (Experiment 1a) and gradable adjectives (Experiment 1b). The hypothesis is that people's judgments in both of these cases will be influenced not only by what they consider average, but also by what they consider ideal.

Method

Ninety participants from Amazon's Mechanical Turk were randomly assigned to judge average, ideal, or normal amounts for a set of 20 behaviors or activities, which were presented randomly on a single page. (We picked behaviors and activities that we predicted would have judged averages that were significantly different from their judged ideals.) Thus, for all domains, approximately 30 participants were

asked questions like "What would you guess is the average number of hours of TV that a person watches in a day?"; another 30 participants were asked questions like "What do you think is the ideal number of hours of TV for a person to watch in a day?"; and the remaining participants were asked questions like "What is a normal amount of hours of TV for a person to watch in a day?"

Results

Participants' responses in each condition were averaged for each of our 20 domains (Table 1). Responses from participants who failed an attention check or that were 3 standard deviations away from the mean answer for a given question were excluded.

Since our questions asked about very different kinds of quantities (hours, calories, etc.), assumptions of normality were violated. To address this problem, mean responses for each measure were converted to log scale.

To examine how judgments of averages and ideals affect normality judgments, we compared a regression model in which only average judgments predict normal judgments to a model in which both average and ideal judgments predict these judgments. The latter model reveals that both judged averages, $\beta = .70$, $SE = .09$, $p < .001$, and judged ideals, $\beta = .33$, $SE = .07$, $p = .001$, significantly predict normality judgments. Moreover, the Bayesian Information Criterion (BIC) for this model (19.48) is lower than that for a model in which only judged averages predict normality judgments (30.46), suggesting that it is a more appropriate model of the observed data.

We also performed the convex combination analysis discussed above (see "The Present Studies"). This analysis yields a 95% confidence interval for b of [.18, .39], demonstrating that participant's normality judgments are intermediate between their judgments about what is average ($b > 0$) and their judgments about what is ideal ($b < 1$).

Discussion

In this experiment, people's explicit judgments about normal amounts of various quantities, like hours of television watching, were best explained by considering both statistical reasoning (what is considered average) and evaluative judgments (what is considered ideal).

Of course, this result may reflect something idiosyncratic about people's use of the word "normal," rather than a deeper truth about people's representations of normality. Thus, we use a more implicit measure in the experiments that follow.

Experiment 1b

In this study, we examine whether participants' judged averages and ideals from Experiment 1a predict a more implicit measure of normality: the notion of a standard amount (see Introduction). We assessed this by asking people the degree to which they thought various quantities relating to the domains of Experiment 1a were large or small amounts. Based on these ratings, we could then estimate what amounts would be considered neither large nor small (the standards).

We again hypothesized that these standard amounts would not be predicted just by participants' estimates of averages, but also by what they judged to be ideal.

Method

One hundred new participants were presented with a single question about each of our 20 domains from Experiment 1a, presented in random order on a single page. The questions had the following format (again taking the TV domain as our example): "Imagine that a person watches *y* hours of TV in a day. Please rate the extent to which you think this is a large or small number of hours of TV for a person to watch in a day." Participants responded on a 7-point scale, ranging from "very small" to "very large."

The number *y* was a randomly selected integer that was 50% likely to fall between the average and ideal values from Experiment 1a and 50% likely to fall outside this range. In most cases, there was a 25% chance that a value would be selected from below the range between the average and ideal and a 25% chance that a value would be selected from above this range. However, if this procedure would have resulted in sampling impossible values (e.g., negative numbers or percentages greater than 100), we did not sample these values, but instead increased the number of possible values on the other end of the distribution. (For example, if, for some domain, the average was 1 and the ideal was 4, we would uniformly sample from the integers between 0 and 7, rather than the integers between -1 and 6.)

Results

To calculate standards for each domain, participants' 7-point ratings were mapped into a range from -3 to 3, such that "very small" corresponded to -3 and "very large" corresponded to 3. Consequently, the zero point on this scale corresponded to the standard point *S*, at which some value is judged to be a neither small nor large amount. We estimated this standard point using linear regression according to the following equation:

$$y = b(x - S)$$

where *y* corresponded to participants ratings on the -3 to 3 scale, and *x* corresponded to the randomly queried values that participants were asked about.

As before, responses from participants who failed our attention check or who gave responses more than 3 SDs from the mean responses on a question were excluded from this analysis.

The estimated values for the standard in each domain are shown in Table 1.

Table 1: Mean Average (A), Ideal (I), Normal (N), and Standard (S) Judgments across Domains from Experiment 1

Domain	A	I	N	S

hrs TV watched/day	4.00	2.34	3.03	3.15
sugary drinks/wk	9.67	3.52	7.30	6.50
hrs exercising/wk	5.37	7.31	6.77	5.98
calories consumed/day	2159.26	1757.84	2063.33	1984.12
servings of vegetables/mnth	34.81	67.67	51.97	54.38
lies told/wk	24.25	2.75	8.43	16.40
mins doctor is late/appointment	17.78	3.97	18.47	13.48
books read/yr	10.07	26.15	9.90	15.24
romantic partners/lifetime	8.04	4.25	8.47	6.27
international conflicts/decade	19.30	1.59	4.82	12.59
money cheated on taxes	604.56	136.45	636.60	605.53
percent students cheat on exam	34.64	3.50	15.97	27.39
times checking phone/day	45.33	13.12	37.17	29.67
mins waiting for customer service	15.04	5.78	12.73	8.08
times calling parents/mnth	6.04	6.00	5.23	6.47
times cleaning home/mnth	5.57	6.75	4.72	6.00
computer crashes/mnth	4.78	0.50	1.60	2.77
percent high school dropouts	12.64	3.82	11.13	10.73
percent middle school students bullied	27.59	2.31	27.26	22.14
drinks of frat brother/weekend	16.79	5.91	14.30	10.94

A regression examining the influence of log-converted judged averages and ideals on log-converted standard amounts once again finds that both averages, $\beta = .80$, $SE = .05$, $p < .001$, and ideals, $\beta = .23$, $SE = .04$, $p < .001$, impact these standard values. Moreover, the BIC for this model (-1.90) is lower than that for a simpler model in which only averages predict standard amounts (12.36), indicating that it is a better model of the data.

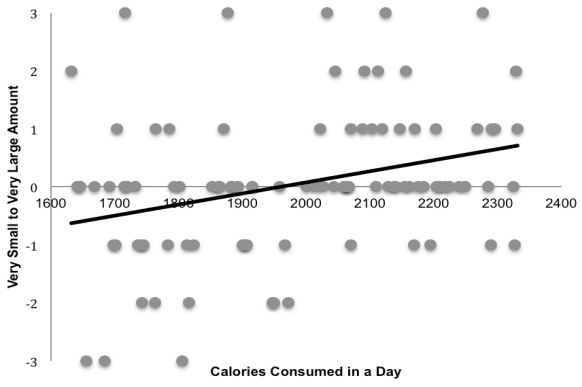

Figure 1: Participants' ratings of the degree to which various daily calorie amounts are small or large. The standard was estimated to be the point at which the regression line crosses the x-axis (1984.12 calories).

Next, as done with normality judgments, we examined whether these standard values could be modeled as a convex combination of average and ideal values. This analysis yields a 95% CI for b of [.13, .26], showing that these standards are once again intermediate between averages and ideals.

Discussion

In conjunction with Experiment 1a, this experiment provides evidence that people's representations of normality are, in fact, influenced by evaluative considerations at a deep cognitive level. Specifically, the effect of judgments about what is ideal extends to both explicit measures of normality and a more implicit measure involving standards from gradable adjectives.

Nevertheless, these results are only correlational. In Experiment 2, we test whether manipulating averages and ideals for a novel category can actually influence how the standard is represented.

Experiment 2

In this study, we experimentally manipulate the average and ideal sizes of a fictional object called a "stagnar." We predicted that, even for this completely novel object, participants' representations of the standard would depend on both the specified average and the specified ideal.

Method

Two hundred participants from Amazon's Mechanical Turk were randomly assigned to receive a certain average and ideal size (in pixels) for a stagnar. Each of these values was independently assigned to participants by randomly sampling from a set of 101 evenly spaced values between 300 and 700 (i.e., the values were all spaced 4 pixels apart).

Participants were first introduced to the concept of an ideal stagnar. They were presented with a description of a fictional hunting tool, along with a green bar presented below this description representing this ideal length for a stagnar to be effective for hunting. (Note that, in order to avoid participants' confusing the concept of an ideal stagnar with the actual stagnars that were supposed to exist in the population, we never presented participants with an image of a stagnar of ideal length. We only presented a green bar to represent the ideal length at all stages of the experiment.) After reading this description and viewing this green bar, participants were told, "If [stagnars] exceed this length too much, they become too difficult to handle; and if they go below this length too much, they become too weak to injure an animal." Participants had to correctly answer two comprehension questions about this ideal stagnar length in order to proceed.

Next, we introduced the average stagnar size by displaying 10 different stagnars (see Figure 2 for an example), which all varied around this average. (These 10

stagnars were always drawn from the same underlying distribution around the average; e.g., the first presented stagnar was always 80 pixels larger than average, the second was always 8 pixels smaller than average, and so on.) Participants were told that "So far, [only] 10 stagnars exist in the world," and that, therefore, the stagnars we would present to them would make up the entire population of stagnars in existence (not just a sample of the population). Participants were further told that we were going to show them each of these stagnars one by one for 5 seconds at a time, along with the green bar below representing the ideal stagnar length. Once participants indicated that they were ready to pay attention, they were presented with each of these 10 stagnars, with the green bar below each of them, on separate pages, which auto-advanced every 5 seconds.

After viewing all 10 stagnars and answering our attention-check question, participants were told that they would be asked a few questions about stagnars. To reduce possible demand characteristics, we attempted to downplay the importance of our crucial dependent variables, involving intuitions about size, by indicating that we first wanted to collect intuitions about the sizes of stagnars, but then we would ask "one further question." In reality, this further question ("Do you think it's better for a stagnar to be smaller than the ideal or larger than the ideal?") was unrelated to the experiment, but was included after the key measures in order to avoid deception.

On separate pages, participants were asked about 5 different hypothetical stagnars. Once again, the sizes of these displayed stagnars were randomly sampled (without replacement) from the 300- to 700-pixel range, spaced 4 pixels apart. For each stagnar shown, participants were asked, "If there were a stagnar that looked like the following, to what extent do you think this would be a large or small stagnar?" and gave answers on a 1–7 scale ranging from "Very small" (1) to "Very large" (7), with "Neither small nor large" (4) in the center of the scale.

Figure 2: Example of a "stagnar" presented to participants in Experiment 2.

Results

Using the same method from Experiment 1b, each participant's standard stagnar size was calculated based on responses to the five questions described above. (Two participants were excluded from further analysis for failing the attention check. One additional participant was excluded from analysis because their judgments of size were negatively correlated with the sizes of stagnars presented.)

To examine whether both the manipulated average and manipulated ideal affect participants' standards, we regressed standard stagnar sizes on averages and ideals. This analysis revealed that both averages, $\beta = .19$, $SE = .04$, $p < .001$, and ideals, $\beta = .65$, $SE = .04$, $p < .001$, significantly predict standard amounts.

As in Experiment 1b, we compared this model to a simpler model in which only the average predicts the standard. The BIC of the more complex model with the ideal (1680.84) is lower than that for this simpler model (1791.36), suggesting it is a better model of the data.

We also examined whether the standard stagnar sizes could be modeled as a convex combination of the manipulated average and ideal values. This analysis finds that standard values are intermediate between averages and ideals, with a 95% CI for b of [.64, .77].

Discussion

Building on the results of Experiment 1b, this experiment demonstrates that people's representations of normality (as manifested in their use of gradable adjectives) are causally influenced by both statistical and evaluative notions.

This experiment further suggests that these representations can be updated quickly, with limited information. Participants in this study had no prior knowledge about stagnars, and they were introduced to this category with only a few short sentences and a picture. Moreover, they were presented with only 10 examples of the category and were merely shown a green line representing the ideal size. Nevertheless, participants' standard stagnar sizes were highly influenced by averages and ideals.

General Discussion

Three studies suggest that people's representations of normality are guided by both statistical and evaluative considerations. Study 1a shows that people's explicit judgments about what are the most normal quantities of various behaviors or events are driven by what are believed to be the average and ideal amounts of these quantities. Study 1b shows that this is also true for the more implicit representation of a "standard," based on the use of the gradable adjectives "large" and "small." Study 2 demonstrates that these standards can be manipulated by changing the average and ideal sizes of a novel object category.

Of course, we suggest that these studies serve as a mere case study in the context of a broader theory of normality. For example, although the present studies focused on *quantities* of various behaviors and events (e.g., hours of television watching), the theory of normality we propose may apply to more qualitative kinds of stimuli, as well. Indeed, research on concepts suggests that people's judgments about prototypical exemplars of categories are influenced by evaluative (as well as statistical) considerations (Barsalou, 1985; though see also Kim & Murphy, 2011). This result could perhaps be explained by the same theory of normality that explains the data we observe.

More importantly, as discussed in the Introduction, normality has been implicated in a number of ongoing research programs—from epistemology in philosophy to behavioral economics. If, as we suggest, people's normality

judgments are affected by both statistical and evaluative considerations, this theory of normality could help explain how people use language, how they reason causally, how they cooperate with others, and how they ultimately do many other things.

Acknowledgments

We are grateful to Paul Egré, Nick Stagnaro and Zoltán Szabó for valuable help and for comments on an earlier draft of the present paper.

References

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child Development*, *79*(3), 594-608.

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*(4), 629.

Dowty, D. (1979). *Word meaning and Montague grammar*. Dordrecht: Reidel.

Egré, P. & Cova, F. (in press). Moral asymmetries and the semantics of "many." *Semantics and Pragmatics*.

Goldman, A. I. (1986). *Epistemology and cognition*. Harvard University Press.

Halpern, J. Y., & Hitchcock, C. (2014). Graded causation and defaults. *The British Journal for the Philosophy of Science*.

Hart, H. L. A., & Honoré, T. (1985). *Causation in the law*. Oxford University Press.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, *93*(1), 75.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, *106*(11), 587-612.

Kennedy, C. (1999). *Projecting the adjective: The syntax and semantics of gradability and comparison*. Routledge.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, *81*, 345-381.

Kim, S., & Murphy, G. L. (2011). Ideals and category typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1092.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In Moral Psychology, Volume 2 (pp. 441–448). Cambridge, MA: MIT Press.

Lassiter, D., & Goodman, N. D. (2014, April). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of SALT* (Vol. 23, pp. 587-610).

Nickel, B. (2008). Generics and the ways of normality. *Linguistics and Philosophy*, *31*(6), 629-648.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*.

Wysocki, T. (Unpublished data). Explicit judgments of normality. University of Washington in St. Louis.