This is a working paper that has not yet been peer-reviewed.

What is common is not blameworthy:

How statistical norms impact judgments of blame and praise.

D.H. Bostyn[1] & J. Knobe[2]

[1] Ghent University; Department of Developmental, Personality, and Social Psychology; Henri Dunantlaan 2, B-9000, Ghent, Belgium.

[2] Yale University; Program in Cognitive Science; 2 Hillhouse Ave, New Haven, CT 06520-8205, United States of America.

dries.bostyn@ugent.be, ORCid: 0000-0001-9994-4615

joshua.knobe@yale.edu, ORCid: 0000-0003-0733-3775

Keywords: Blame, Praise, Statistical norms, Moral Cognition

WORD COUNT: 9904

Authors' Note:

This research was supported by postdoctoral research grants from the Research Foundation - Flanders (FWO.3E001619) and the Bijzonder Onderzoeksfonds (BOF22/PDO/119), and a scholarship from the Fulbright Commission in Brussels awarded to the first author as well as by set-up funds given to the second author by Yale University.

Correspondence concerning this article should be addressed to Dries H. Bostyn, Department of Developmental, Personality, and Social Psychology, Henri-Dunantlaan 2, B-9000, Ghent, Belgium. E-mail: Dries.Bostyn@Ugent.be.

Abstract

For many types of behaviors, whether a specific instance of that behavior is blame- or praiseworthy depends on how much of the behavior is done or how people go about doing it. For instance, for a behavior such as "replying to an email in *x* days", whether a specific reply is perceived as blame or praiseworthy will depend on how many days elapsed before the reply. Such behaviors lie on a continuum in which part of the continuum is praiseworthy (replying quickly) and another part of the continuum is blameworthy (replying late). In the current manuscript, we investigate how judgments of blame and praise on such gradual behaviors relate to people's perceptions of the statistical norms surrounding that behavior (i.e., how quickly people usually reply). We find that people do not base judgments of blame and praise on a comparison to the statistically average quantity. Instead, they show an asymmetric effect of statistical frequency: frequent behaviors are typically not considered to be blameworthy but can still be considered praiseworthy. Whereas the frequency of a behavior is strongly related to how blameworthy it is perceived to be, the effect of frequency on judgments of praise is much more diminished.

Keywords: Blame, Praise, Statistical norms, Moral Cognition

This is a working paper that has not yet been peer-reviewed.

1. Introduction

Imagine you are a student writing an important, time-sensitive email to a professor. If the professor responds quickly, you might regard that as praiseworthy. However, if they take an entire month to respond, you will likely view it as blameworthy. Somewhere along the continuum of ever-later replies, a behavior once considered to be good slides into a behavior perceived to be bad. This is not unique to responding to emails. Commonplace behaviors such as spending time with your family, tipping, or the extent to which one listens to a friend in need, have a similar gradual nature. Certain aspects of these behaviors can be varied in a continuous way, and the amount of blame or praise associated with them varies continuously with the values of these parameters. How exactly do people determine when to praise and when to blame if they are confronted with behaviors like this?

Research suggests that judgments of praise and blame are influenced by people's representations of what people typically do (Ericksson, Strimling, Coultas, 2015; Lapinski & Rimal, 2005). So in the example of the email, whether a reply is deemed praiseworthy or blameworthy may depend on how that reply compares to what is typically done. Indeed, there is a long-standing literature in psychology investigating how social and descriptive norms relate to moral judgments (Bear, Bensinger, Jara-Ettinger, Knobe, & Cushman, 2020; Cialdini, et al. 1990; Cialdini & Trost, 1998; Goldring, & Heiphetz, 2020; Monroe, Dillon, Guglielmo, Baumeister, 2018; Nolan, et al., 2008; Philips, Morris, & Cushman, 2019; Roberts, et al., 2019; Wysocki, 2020). People use their perception of social norms as a standard to compare behavior to, and they judge the appropriateness of behavior by how much it deviates from the norm (Baer, Stacy, & Larimer, 1991; Clapp & McDonell, 2000; Schmidt, Rakoczy, & Tomasello, 2012). A question now arises about precisely how to understand the impact of statistical information on people's judgments of blame and praise within the context of gradual behaviors.

1.1. Three hypotheses

Perhaps the simplest and most obvious hypothesis would be that people compare an agent's behavior to whatever they believe to be the statistical *average*. If they conclude that the agent's behavior is worse than average, they see the agent as blameworthy, whereas if they conclude that the agent is better than average, they see the agent as praiseworthy. This hypothesis makes some clear testable predictions. For example, it predicts that if people think that what the agent is doing is completely average, they should see the agent as *neutral*, i.e., as neither blameworthy nor praiseworthy.

However, this is not the only way in which statistical information could impact moral valuations. Rather than being shaped just by a representation of the average, moral valuations might be shaped by a representation of the entire distribution. To illustrate, consider a helpful behavior, and suppose people think the average quantity of this behavior is around three hours per week. Now consider two different ways in which people might represent the distribution. One possibility would be that they think most people do the behavior approximately three hours per week, with relatively few doing it a lot more or a lot less. Another possibility would be that they represent the behavior as having a highly right-skewed distribution: the majority of people never do the behavior at all, but a minority do a large quantity, which drives up the average to approximately three hours. Even though these two distributions have the same mean, they might be associated with very different patterns of praise and blame judgment.

If we assume that judgment of praise and blame are influenced by representations of the entire distribution, one natural hypothesis would be that this influence has a very simple form. Specifically, it might be that there is a tendency such that if a particular quantity is seen as highly frequent, that quantity is seen as less deserving of praise or blame. Thus, consider again our example in which people perceive the average quality to be three hours. On this hypothesis, if people think that three hours is itself a very frequent quantity, then they will think that doing the behavior for three hours deserves little, if any, praise or blame. By contrast, suppose they represent the behavior as having a highly right-skewed distribution, such that most people do less than three hours, and three hours is a quantity with a relatively low frequency. In such a case, they might think that doing the average amount is itself highly praiseworthy.

Finally, a third hypothesis would say that people's praise and blame judgements are indeed shaped by their representation of the whole distribution, but that this effect is not a matter of frequency having a symmetric impact on praise and blame (e.g., decreasing praise and blame equally). Instead, it might be thought that there is a special connection between representing something as statistically frequent and judging that it is *not morally bad*. A number of existing studies seem to point in this direction. Perceiving a behavior as statistically frequent seems to decrease the degree to which people regard it as wrong (Goldring & Heiphetz, 2020; Monroe et al., 2018; Roberts et al., 2017), increase the degree to which it is seen as morally right (Lindström, Jangard, Selbing, & Olsson, 2018), and increase the degree to which people are inclined to perform it themselves (Cialdini et al, 1991). Moreover, there appears to be a fundamental connection between the representation of statistical frequency and the representation of prescriptive goodness (Phillips & Cushman, 2017; Phillips & Knobe, 2018).

This third hypothesis would predict an asymmetric impact of frequency on praise-blame judgments. If a particular quantity is seen as highly frequent, people will tend not to regard it as blameworthy, but they might still regard it as praiseworthy. Take the case in which people think that the average quantity is three hours and that this is also the most frequent quantity. On this third hypothesis, this representation of that statistical distribution will make people less inclined to see doing the behavior for three hours as blameworthy, but it will not make them less inclined to see doing a behavior for three hours as praiseworthy. In other words, even if a particular quantity is seen as the average and as a highly frequent quantity, it might still be seen as praiseworthy.

1.2. The present studies

In what follows, we report the results of three studies exploring the relationship between what people consider to be typical behavior and their moral judgment of gradual behavior. We investigate whether the moral judgment of gradual behaviors is based on a comparison to the average or whether it is based on the frequency of each instance.

In Studies 1a and 1b, we confront people with a variety of common, everyday behaviors described through vignettes and estimate *blame-praise curves.* These curves link the amount of praise and blame to the underlying behavioral continuum and can subsequently be used to estimate a *neutral point* for each curve. This will allow us to compare how these neutral points relate to people's perception of average behavior and provide an initial way of testing broader theories about the impact of statistical information on blame-praise curves.

Building on the results of these studies, we explore blaming and praising in an experimental paradigm. In Study 2, we confront participants with a made-up behavior and give them information about how people typically perform that behavior. In that study, we manipulate whether the made-up behavior is described as either positively valenced (a behavior that is judged to be more praiseworthy as the quantity increases) or negatively valenced (a behavior that is judged to be more blameworthy as the quantity increases).

Finally, in Study 3, we return to investigating naturalistic behaviors described through vignettes. In this study, we confronted participants with multiple instances of each behavior and probed them both for how frequent they judged that instance to be, as well as how blame- or praiseworthy they considered it. This approach allowed us to directly assess the relationship between judgments of frequency and those of blame and praise within each individual participant.

2. Open Practice Statement

In light of well-known concerns about "hypothesizing after results are known" (Kerr, 1998), we want to disclose that the theoretical framework described in this introduction was developed only after we conducted some of the studies reported below. In particular, Studies 1 and 2 were not designed with these hypotheses already in mind. Rather, the framework was developed in part because of the results of those studies. When we argue that the results of Studies 1 and 2 are best understood in terms of this theoretical framework, we do not mean to be implicitly suggesting that we conducted those studies as a way of testing that framework.

Along the way to developing the theory presented in this introduction, we conducted a variety of other studies designed to test other possible hypotheses. For transparency, we include a summarized write-up of these studies in the online supplementary materials. We preregistered all studies we ran. At a minimum, we preregistered the materials, methods, and sample size. For some studies, but not all of them, we also preregistered specific hypotheses and the analyses we would use to test those hypotheses. Within the current manuscript, we explicitly note which hypotheses and analyses were preregistered and which ones were not. As not all studies were conducted before we developed the framework presented in this introduction, some of the analyses we preregistered do not bear directly on this framework. The data, materials, and preregistrations for all studies included in this project are available at https://osf.io/jzvhn/?view_only=7bfc6866d2ab4b9e8b8a7d156348cb3d/. The data and materials for the specific studies included in the current manuscript are available at: https://osf.io/x7wp4/?view_only=fab21b861e4d42e0bf4ec06e4dc6dbcc.

3. Study 1a and 1b

In a first set of studies, we aimed to estimate blame-praise curves for a number of different behaviors. For example, we looked at the way the amount amount of praise and blame participants assign for the behavior *comforting a friend for x hours* changes as one goes from a very small of time to a very large amount of time, and similarly for *responding to an email after x amount of time*, *cheating x number of times during a marriage*, and so forth.

By estimating blame-praise curves, we can find out what the neutral points are for these behaviors and investigate how these neutral points relate to participants' perceptions of statistically average behavior. This should enable us to ask whether the neutral points tend to be approximately equal to the average or whether they differ from averages in some systematic way.

3.1. Pre-studies: determining behavioral ranges

Before running these studies, we needed to determine an appropriate behavioral range we wanted to study for each behavior as the shape of each curve depends on the range that is being studied. Rather than determining these ranges a priori, we wanted to determine the range for each behavior empirically. Accordingly, we ran pre-studies for all behaviors included in Study 1a and 1b to determine what people perceived to be "average" behavior and to find an appropriate behavioral range to study. In Study 1a, we estimated blame-praise curves for 14 positively valenced behaviors, i.e., behaviors where the more of the behavior you do, the more positive it becomes. In Study 1b, we estimated blame-praise curves for 11 negatively valenced behaviors, i.e., behaviors where the more of the behavior you do, the more negative it becomes. Each of the behaviors we studied was described to participants using a vignette.

Anticipating some drop-out, we recruited a total of 111 participants for the pre-study of Study 1a and asked what they thought the average amount was that people typically do for all of the 14 vignettes included in that study. For the pre-study of Study 1b, we recruited a total of 112 North-American participants and confronted them with 11 vignettes. In both pre-studies, we recruited North-American participants from Amazon Mechanical Turk who were paid US$0.40 for the completion of these studies. An example of the vignettes used in these prestudies reads as follows:

"*Jack is the father of a 5-year-old girl. What would you guess is the average amount of hours that fathers like Jack spend playing with their daughter on a regular Saturday?*"

First, participants were presented with a short demographic survey in which they were asked to indicate their age and gender, and were subsequently asked to respond to all the vignettes included in that specific pre-study. After responding to the vignettes, participants were asked to guess the hypothesis of the pre-study as a data quality control question. Using such an open-ended question as a data quality control question allows one to filter out participants with a poor grasp of English. The first author coded these responses for nonsensical answers (e.g. "nice study") or egregious grammatical mistakes. Such participants were considered to be unreliable and eliminated from the study. A total of 20 participants were eliminated from the pre-study of Study 1a, and a total of 18 participants were eliminated from the pre-study of Study 1b. Accordingly, the final sample for pre-study 1a consisted of 91 participants (35 female, 56 male) with a mean age of 35.44. The final sample for pre-study 1b consisted of 94 participants with an average age of 38.65 of which 31 self-identified as female and 63 self-identified as male. We did not inquire about participants' ethnicity in these or any other studies. Participants received US$0.40 in compensation for completing the pre-study.

All behaviors described in our vignettes had a natural minimum: zero. Given the potentially important theoretical relevance of natural minima, zero was used as the minimum for the range of all behaviors that we tested. The responses to the vignettes were right-skewed, with some participants providing very extreme or even impossible answers (e.g. >24 hours for the father vignette described above). To minimize the impact of these extreme responses, the range for each behavior was determined by taking the 95th percentile of participants' responses. These ranges were then used in the main study.

3.2. Main studies: participants and procedure

The goal of the main studies was to estimate blame-praise curves for the behaviors described in the vignettes and use these curves to estimate the neutral point up to a certain degree of uncertainty (5% of each range). We did not conduct a traditional power analysis to determine an appropriate sample size for the current study. Instead, we simulated hypothetical response data from a variety of different types of data-generating models, added random residual noise to all simulated responses, and then retro-fitted different types of curves to this data to test how well we would be able to estimate the neutral point for different amounts of residual noise and different types of data-generating models. These simulations suggested that 300 participants would be sufficient for our purposes. The R code to repeat these simulations is available on OSF (https://osf.io/jzvhn/?view_only=7bfc6866d2ab4b9e8b8a7d156348cb3d).

Anticipating some drop-out, we recruited 331 North-American participants for Study 1a and 317 North-American participants for Study 1b from Amazon Mechanical Turk. Participants received US$0.50 for completing either of the studies. Participants completed demographic information (age and gender), responded to all vignettes included in that study and were asked to guess the hypothesis of the studies as a data quality control measure. A total of 44 participants failed this check in Study 1a, and 25 in Study 1b. Eliminating these latter participants left us with 287 participants with a mean age of 36.13, of which 120 self-identified as female and 167 as male in Study 1a. Study 1b had a final sample consisting of 292 participants with an average age of 38.89, of which 134 self-identified as Female, 156 as Male, and two as Other.

3.3. Main study: vignette task

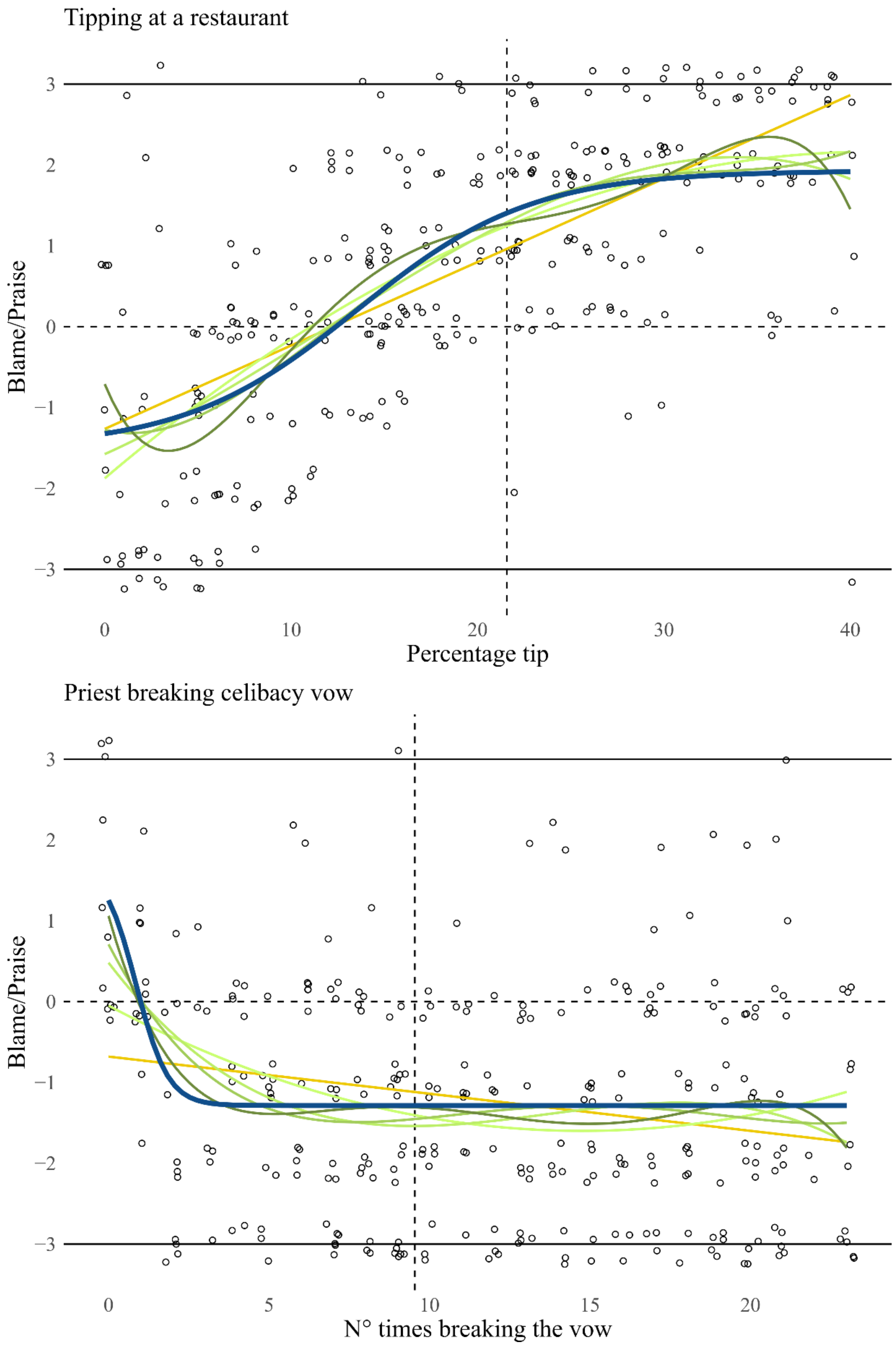
Participants were confronted with a total of 14 vignettes in Study 1a and 11 vignettes in Study 1b, administered in a randomized order. Each vignette described an everyday behavior for which blame- or praiseworthiness is related to how much of the behavior is done. On each vignette, participants were asked to rate the blameworthiness or praiseworthiness of a specific quantity on a behavioral continuum. These quantities were randomly selected integers from the range associated with the specific vignette obtained in the pre-studies. An example for Study 1a reads: "*Jack is the father of a 5-year-old girl. On a regular Saturday, fathers typically spend some time playing with their daughter. Jack spends [0-8] hours playing with his daughter*". An example for Study 1b reads: "*Linda is a university professor and often receives emails from students asking her questions. University professors typically respond to emails of their students, although it sometimes takes a while for them to do so. Linda always takes [0-5] days to respond to the emails of her students.*" All vignettes are available through the OSF page associated with this project (https://osf.io/jzvhn/?view_only=7bfc6866d2ab4b9e8b8a7d156348cb3d). For each behavior participants received, they were asked to give a rating on a 7-point scale spanning from (-3) *Extremely Blameworthy* to (3) *Extremely Praiseworthy*

3.4. Results

The main goal of these studies was to uncover how people's perception of average behavior relates to the neutral point for each vignette, i.e., the specific instance of a behavior at which participants neither blame nor praise. Since we did not want to presume that all curves would have a specific shape, we first fitted non-parametric loess curves to the data of each vignette. Loess curves model the local average of data and allow one to visually inspect the shape of a curve without the need to fit data to a specific type of curve. Subsequently, we fitted multiple types of regression curves to the data of each vignette. For each vignette, we tested a linear fit, a logistic fit, and polynomial fits (of the second up to the fifth order). A more detailed analysis of fit statistics is available through the supplementary materials.

Generally speaking, a number of clear patterns emerged. On nearly all vignettes, linear models had the worst fit. In contrast, polynomial models typically had the best fit with the data, although which order polynomial fit best with the data differed across vignettes. Importantly, these polynomial fits demonstrated clear signs of overfitting. A visual example of all six different fits for two vignettes (one for Study 1a and one for Study 1b) is displayed in Figure 1. The logistic model, while never the best-fitting model, performed adequately and did not lead to unreasonable predictions.

Figure 1



Note. Linear (yellow), Polynomial (green), and Logistic (blue) fits for a vignette of Study 1a (top) and Study 1b (bottom). The vertical dotted line represents the perceived statistical average. Random jitter was added to the plot to eliminate overplotting.

The different models led to (slightly) different estimates for the neutral points. Fortunately, as can be gleaned from Figure 1, the estimates of the neutral point obtained through most types of models were interchangeable. For a number of vignettes, linear fits did deviate from this, but whenever they did, the linear regression line was clearly distinct from all other estimated curves. In those cases, the AIC values of those linear fits also suggested they had a substantially worse fit to the data. As per our preregistration, we decided to use the logistic models to estimate the neutral point for each of the vignettes, except when the estimate obtained through the best-fitting model differed more than 5% of the total range from the estimate obtained through the logistic model. This was not the case for any vignette, and as a result, we used the logistic models to estimate all neutral points. Table 1 shows the neutral point obtained for each of the vignettes along with what the pre-study participants perceived to be the average behavior for those vignettes.[1]

Table 1

Neutral points (as estimated through the logistic model) and perceived averages for all vignettes of Study 1a (top) and Study 1b (bottom).

Vignette	Neutral Point	Perceived Average
Comforting a Friend	0.57 hours	4.22 hours
Playing with daughter	1.24 hours	4.87 hours
Tipping	12.40%	21.57%
Hospital Visits	2.10 visits	8.79 visits
Household work	30.25%	42.40%
Blood Donations	0.79 donations	10.48 donations
Rounds of Drinks	20.61 %	26.57 %
Fireman	6.64 minutes	24.19 minutes
Babyshower Gift	US$ 20.00	US$ 79.56
Orphanage Charity	-0.96%	19.44%
Returning Loan	NA	9.69
Serving Tables	NA	8.97
Teacher Preparation	14.72	56.19

[1] Some of these averages might appear to be on the higher side based on the summary description of the vignettes in this table. This is due to how some of the vignettes were framed. For instance, the “Orphanage Charity” vignette asks people to ponder how much a person who was raised in an orphanage would donate to their former orphanage and further specifies this person has a successful career and attributes a large part of their success to that orphanage.

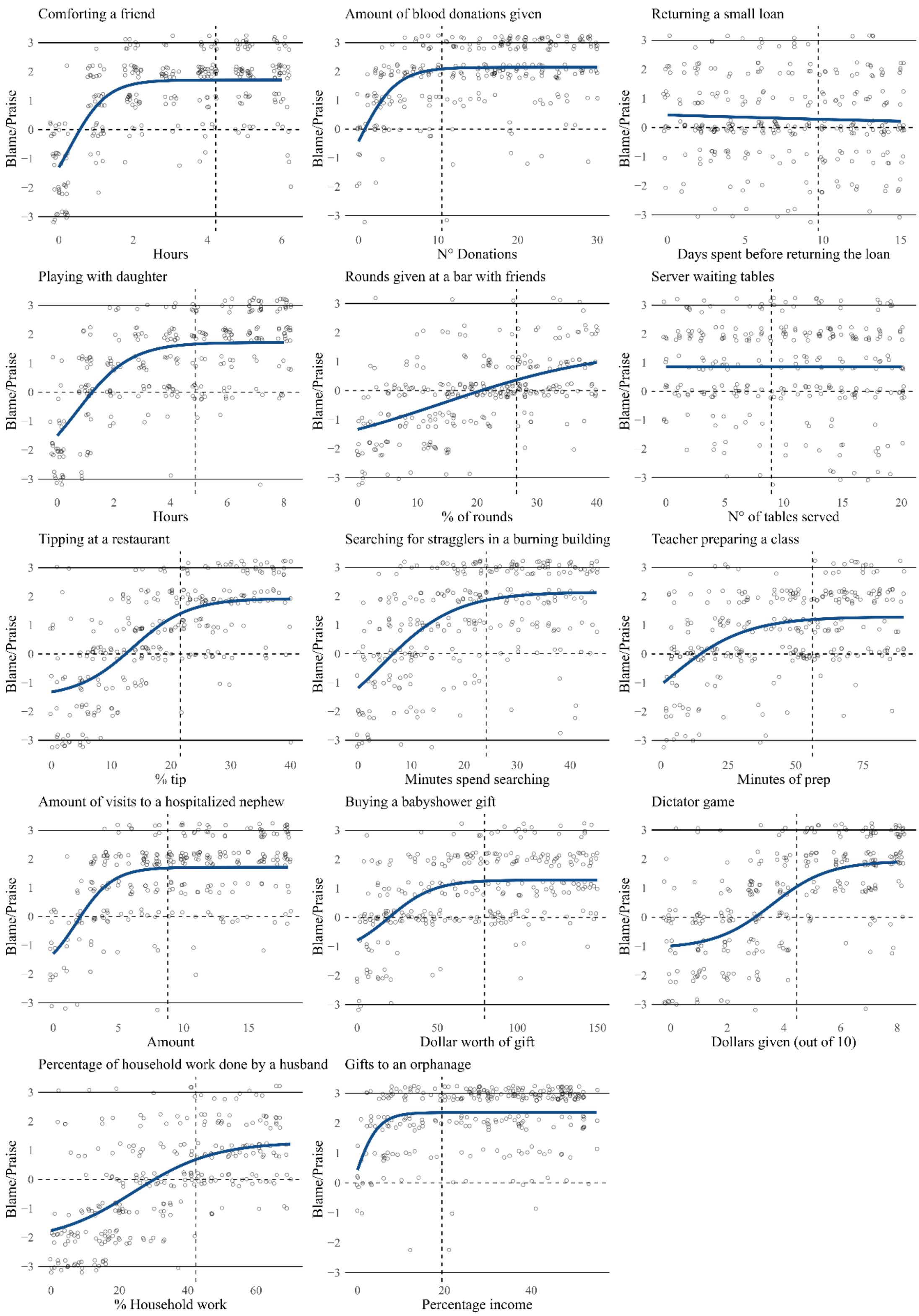
Dictator Game	2.98	4.45
Priest breaking vow	1.00 times	9.54 times
Cheating Husband	0.38 times	5.59 times
Breaking speed limit	6.81 mph	15.93 mph
Responding to emails	2.13 days	3.07 days
Cheating at a game	3.85%	36.59%
Not paying for groceries	0.69 times	5.51 times
Overtime on weekend	5.61 hours	7.84 hours
Arriving late	1.43 minutes	13.63 minutes
Canceling plans	6.63%	29.79%
Flirting	7.69%	40.86%
Missing kids' matches	2.31 matches	6.04 matches

When we initially preregistered these studies, we had hypothesized that behaviors would cluster into two categories: behaviors for which the neutral points and curve-shape would be related to statistical norms, and behaviors for which neutral points and curve-shape would be unrelated to statistical norms. We had no preconceived ideas on what might differentiate behaviors but assumed that categories would emerge as some behaviors we tested have clear moral norms associated with them (e.g. cheating) whereas others are more of a conventional nature (e.g. tipping behavior). Across studies, we failed to find any evidence for differences between categories of behaviors (for more information, see the supplementary materials).

A visual representation of all vignettes, their blame-praise curves (estimated through a logistic model) and the averages associated with these vignettes is available in Figure 2a and Figure 2b. Blame-praise curves have a variety of different shapes: some curves appear linear, others more sigmoidal and yet others are convex or concave shaped. A glance at the neutral points and the perceived averages does show these to be quite distinct (see Figures 2a and 2b), but while they are clearly different, the former do seem to be systematically lower than the latter. Indeed, exploratory paired samples t-tests confirm that the difference between neutral points and averages is statistically significant, both in Study 1a, $t(11) = -3.16$, $p = .009$, $d = -0.91$, as well as in Study 1b, $t(10) = -3.46$, $p = .006$, $d = -1.04$. While these first two studies thus provide clear evidence against the idea that neutral points are anchored at the perceived average, they do suggest that statistical information does have some kind of systematic relation to blame and praise judgments.

Figure 2a

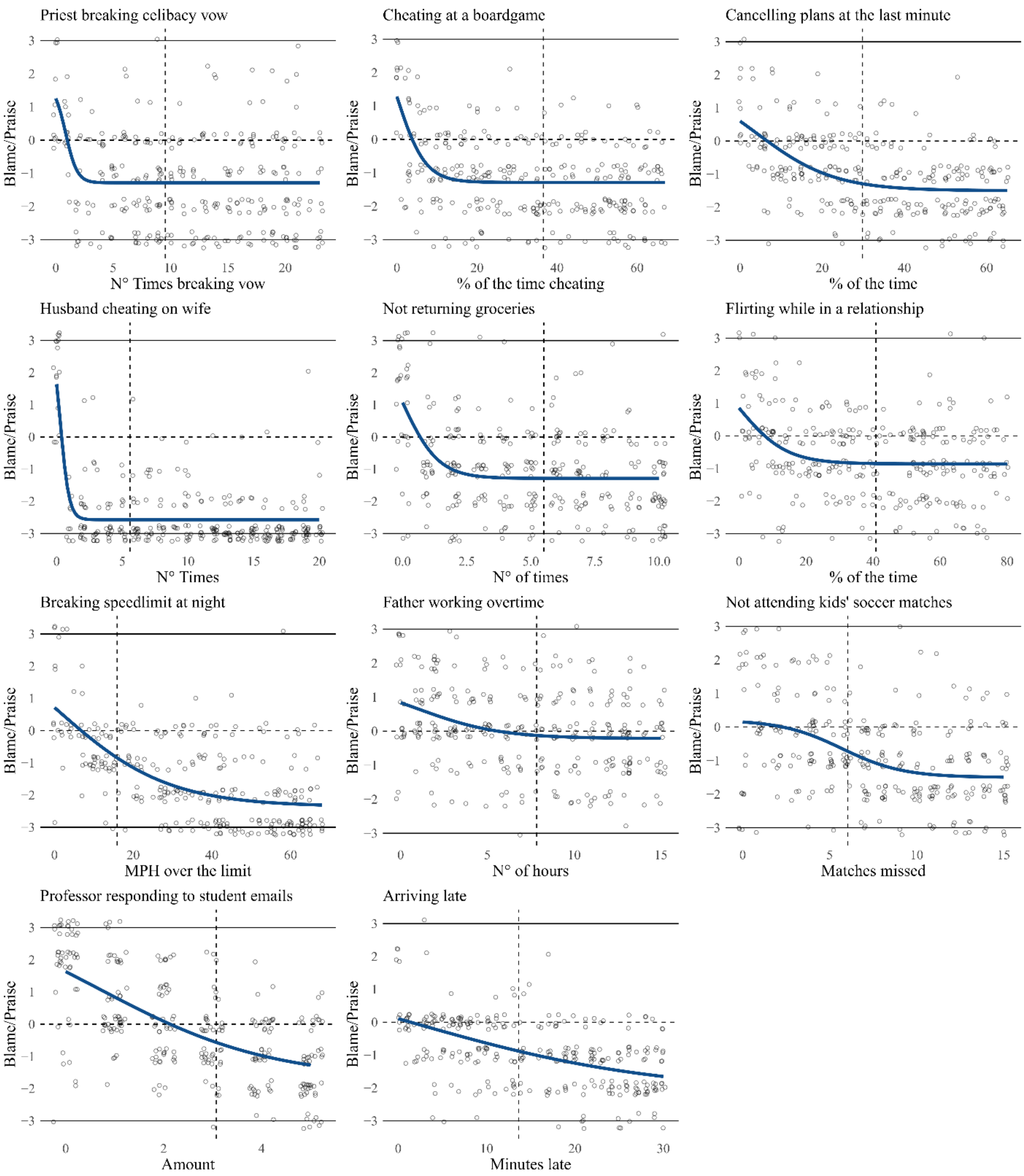
Logistic curves associated with each vignette in Study 1a.



Note. The dotted vertical line represents the perceived average. Random jitter was added to eliminate overplotting.

Figure 2b

Logistic curves associated with each vignette in Study 1b.



Note. The dotted vertical line represents the perceived average. Random jitter was added to eliminate overplotting.

3.5. Discussion

In these first studies, we looked at the perceived averages and the blame-praise curves for a variety of different behaviors. By estimating blame-praise curves, we could investigate the relationship for each behavior between the perceived average and the neutral point. This approach yielded two key results.

First, this analysis confirmed that the neutral point of a blame-praise curve is not equal to people's perception of the statistical average of that behavior. For example, in Study 1b, people thought that the average number of times people cheat on their spouse was six, but cheating on your spouse a single time was still considered to be blameworthy. Conversely, people think that the average number of times to give blood is ten, but they think that giving blood ten times is still praiseworthy. In short, across multiple different behaviors, we find that people are willing to attribute blame or praise even when they believe that what the agent has done is statistically average.

Second, even though the neutral point is not simply equal to the average, it is clear that statistical information is still related to how people think about blame and praise. In particular, we find a very systematic tendency for the neutral point to be lower than the average. This suggests that neutral points, and by extension when people blame or praise, must be closely related to at least some aspects of the statistical distribution of behavior. In other words, the amount of time a person has to spend comforting a friend before they are considered praiseworthy is not simply the statistical average, but it does seem to be related in some important way to statistical facts about how much time people tend to spend comforting their friends.

One explanation for this pattern could be that it arises because people's praise and blame judgments are impacted not by their representation of the statistical mean but by their representation of the full statistical distribution. To illustrate, participants judged that the average number of times for husbands to cheat on their wives was approximately 6, but presumably they do not think that the distribution is one in which the majority of husbands cheat on their wives approximately 6 times and only a minority cheat either substantially less or substantially more. Instead, they might think that the distribution is highly right-skewed: a large proportion of husbands never cheat at all, but a small portion cheat a very large amount, which drives up the average. Thus, even though participants think that the average is approximately six, they might think that it is much more frequent for husbands to cheat zero times than to cheat approximately six times.

One way to further explore this question would be to look at how people respond in cases where they are given more information about the shape of the statistical distribution. Consider what might happen if they learn that the distribution is not skewed but is instead a roughly Gaussian distribution in which points close to the average have very high frequency. If people's praise and blame judgments are indeed shaped by a representation of the entire distribution, one might expect that when the distribution has this sort of shape, the neutral point is sometimes *higher* than the average.

Consider first a negatively valenced behavior, which is seen as more blameworthy the more one does it. If people tend to assign less blame when they see a behavior as frequent, then if points near the average are themselves highly frequent, people should not be inclined to see points near the average as blameworthy. So the neutral point – the point where judgments switch over from praise to blame – should be higher than the average.

Similarly, if it turns out that seeing a behavior as frequent leads to decreased attributions of praise, the neutral point should be higher than the average for positively valenced behaviors. In cases where a behavior is more praiseworthy the more one does it, we should find that participants do not regard to behavior as praiseworthy when someone is simply doing the quantity of it that is most frequent. People should only regard it as praiseworthy when a person goes beyond that quantity.

4. Study 2

In this second study, we used artificial stimuli that allowed us to more precisely control participants' beliefs about the statistical distribution of a behavior. Participants were told about a behavior that was either positively or negatively valenced, but in both cases, they were given reason to think that the frequency with which people performed this behavior had a symmetric distribution. Thus, the point with the highest probability was the statistical mean, and the frequency at points lower than the mean was exactly equal to the frequency at points higher than the mean. The key question then was whether we would find that the neutral point was below the mean in both conditions which would suggest that frequency impacts judgments of blame alone.

4.1. Participants and procedure

Participants were asked to respond to a short demographics questionnaire (age and gender) and then completed an experimental blame-praise task involving a made-up behavior called "blarging". Finally, they were asked to guess the hypothesis of the study as a data quality control question. We recruited a total of 438 North-American participants through Amazon Mechanical Turk, 77 of which failed the data quality control question. All participants were paid US$0.50. The final sample (mean age = 36.31) consisted of 160 participants who self-identified as female, 200 who self-identified as male, and 1 participant who self-identified as "other".

4.2. Measures: experimental blame-praise task

Participants were asked to imagine traveling to a fictional country where people perform a behavior called "blarging". Half of the participants were told that it is good to blarg a lot and bad if people do not blarg enough; the other half were told that it is good not to blarg a lot, and bad if you blarg too often. After learning about blarging in general, we confronted participants with statistical information on the frequency of blarging. More specifically, we displayed how often 60 people within this fictional country blarg. Each amount of blarging was displayed for a single second, and the same statistical information was provided to all participants. The statistical information provided was randomly sampled from a symmetrical beta distribution scaled to produce values within the range of 0 to 100 with a mean of 50.

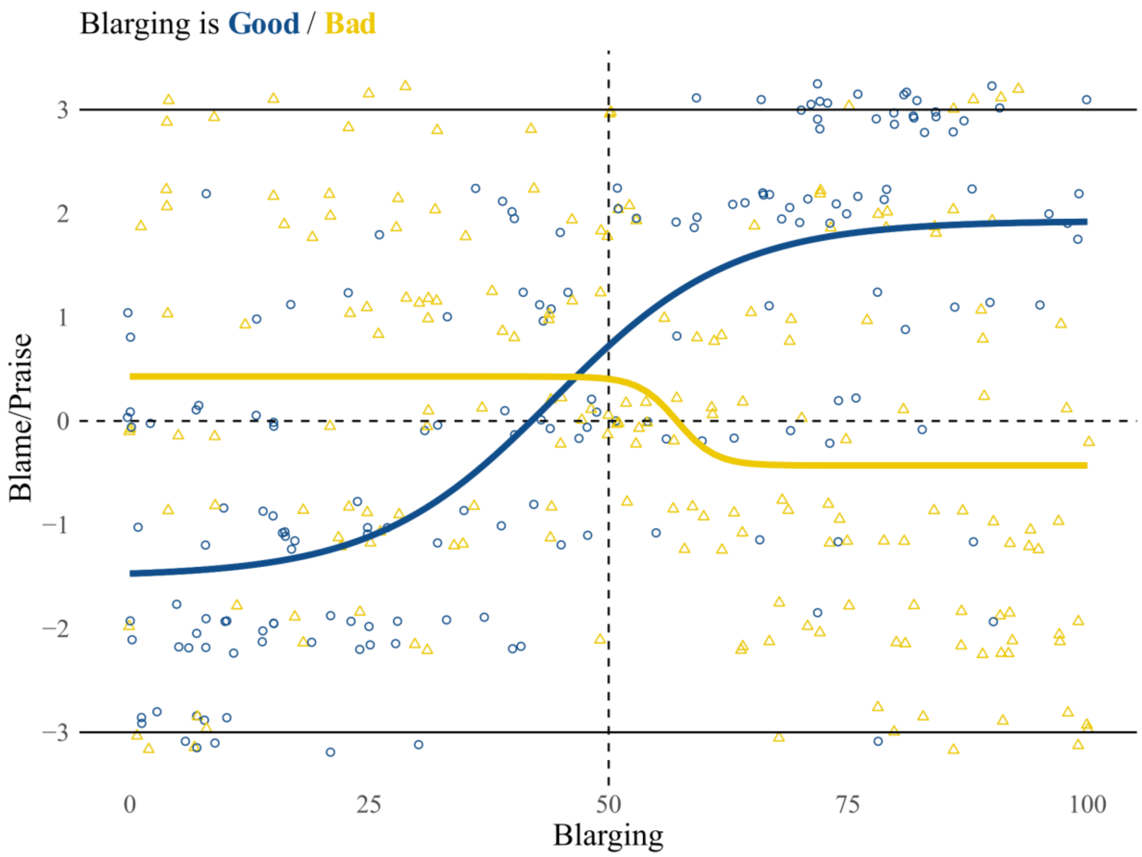
4.3. Results

We preregistered two analyses. First, we wanted to test whether an overall positive-negative asymmetry was present in the data. To do so, we first reverse-scored the blame-praise ratings for the negatively valenced condition and combined this reverse-coded data with the data from the positively valenced condition. Through this recode, we essentially flip the data of the negatively valenced condition vertically around the axis of neutrality. If no positive-negative asymmetry is present, such a procedure would lead to two overlapping blame-praise curves and no difference in the average amount of blame. We ran a linear regression with the condition as a categorical predictor and "amount of blarging" as a control variable. This analysis revealed a significant difference, $t(358) = -2.63$, $\hat{\beta} = -0.12$, $p = .009$. While this result is not straightforwardly interpretable, it does imply that some kind of asymmetry is present in the data.

Subsequently, we fitted logistic curves to the (non-recoded) blame-praise data of both conditions. As Figure 3 demonstrates, we found that the neutral point associated with the positively valenced blarging condition was lower than the average of the statistical information provided to participants, replicating the results of Study 1a. However, in contrast to the results of Study 1b, the neutral point associated with the negatively valenced blarging condition was *higher* than the average.

Figure 3

Logistic fits for the positively valenced (blue) and negatively valenced (yellow) conditions.



Note. The vertical dotted line represents the statistical average.

4.4. Discussion

In this study, participants were asked to judge a made-up behavior they were unfamiliar with. Participants were given information about the statistical distribution of this behavior that specifically indicated that it had a distribution in which quantities close to the average were extremely frequent. The results indicated an asymmetry between blame judgments and praise judgments. Participants tended to judge a person does not deserve blame for doing something that is highly frequent (which meant that the neutral point for negatively valenced behaviors was higher than the average). But praise judgments did not show the same pattern. Participants tended to say that even when a person does something that is highly frequent, that person can be deserving of praise (meaning that the neutral point for positively valenced behaviors was lower than average).

In light of these results from studies with artificial behaviors, we wanted to return to a methodology that involved naturally occurring behaviors. Consequently, we ran a third study using an adapted version of the vignette paradigm. In our first two vignette studies, we estimated a single blame-praise curve per behavior across many participants. We then related a perceived average to the neutral points of blame-praise curves estimated across the responses of many participants. Such a methodology does not allow us to assess the impact of frequency itself. Looking at the average is simply insufficient.

In this final study, we opted to probe participants about five instances of the same behavior. We used a two-wave approach, asking them to judge the frequency of each instance on the first wave and how blame or praiseworthy they considered that instance to be on the second wave. This procedure allowed us to estimate blame-praise curves while also giving us insight into the perceived frequency distributions of behavior. As a result, we can directly test the effect of frequency on judgments of blame and praise.

5. Study 3

5.1. Participants and Procedure

Like Study 1, Study 3 used vignettes describing everyday behaviors. Data for Study 3 were gathered in two waves. A total of 406 Western-European first-year psychology students completed the first wave for course credit and 333 students completed the second wave of data collection. While students were incentivized to complete both waves of the experiment, they were free to participate in other experiments as well. As a result, some students completed only one of the two waves. 308 students completed both waves of the experiment and only these were included in our analyses. The second wave of data collection contained an attention check that merely instructed participants to respond with “not often”. Ten participants failed the attention check and were eliminated from the analysis. Analyses were conducted on the remaining 298 participants (with a mean age of 18.79, of which 255 identified as female, 40 as male, and 3 as non-binary).

In the first wave of data collection, participants completed demographic information (age and gender) and were subsequently asked to respond to a vignette task aimed at measuring how frequent they thought the behavior described in the vignette was. The second wave of data collection took place two weeks after the first wave and used the same procedure as the first wave. Participants were confronted with the same vignettes and the same instances but were asked to rate how praise- or blameworthy the behavior was rather than rate the frequency.

5.2. Measures: vignette task.

Participants were confronted with a total of 16 positively valenced vignettes presented in a randomized order. Each vignette described an everyday behavior for which blame- or praiseworthiness is related to how much of the behavior is done. In contrast to prior studies, participants were not asked to judge only a single instance of each behavior. Instead, participants were asked to respond to five instances of the behavior that were sampled at equidistant points within the behavioral range. The lowest of these instances was determined in a random fashion. To illustrate, imagine that participants were confronted with a behavior with a range of [1 to 15]. Participants were then asked to judge on of three possible sets of instances: [1,4,7,10,13], [2,5,8,11,14] or [3,6,9,12,15]. The behavioral ranges for each vignette were determined a priori by the authors.

In the first wave of data-collection, participants were asked to rate how common or uncommon it would be to do the randomly selected amount of the behavior using a 6-point scale going from (1) Absolutely uncommon to (6) Absolutely common. Additionally, they were asked to guess what people do on average for that behavior. During the second wave of data-collection participants were asked to rate blameworthiness and praiseworthiness on a 7-point scale spanning from (-3) Extremely Blameworthy to (3) Extremely Praiseworthy. An example vignette reads:

"*John is a high school student playing soccer for his high school team. His team is playing in a regional semi-final. John is an important player for his team but stayed up last night playing video games and slept poorly as a result. John puts in a certain amount of effort.*

Wave 1: How common do you think it is for people in John's situation to put in [50 to 100%] effort.

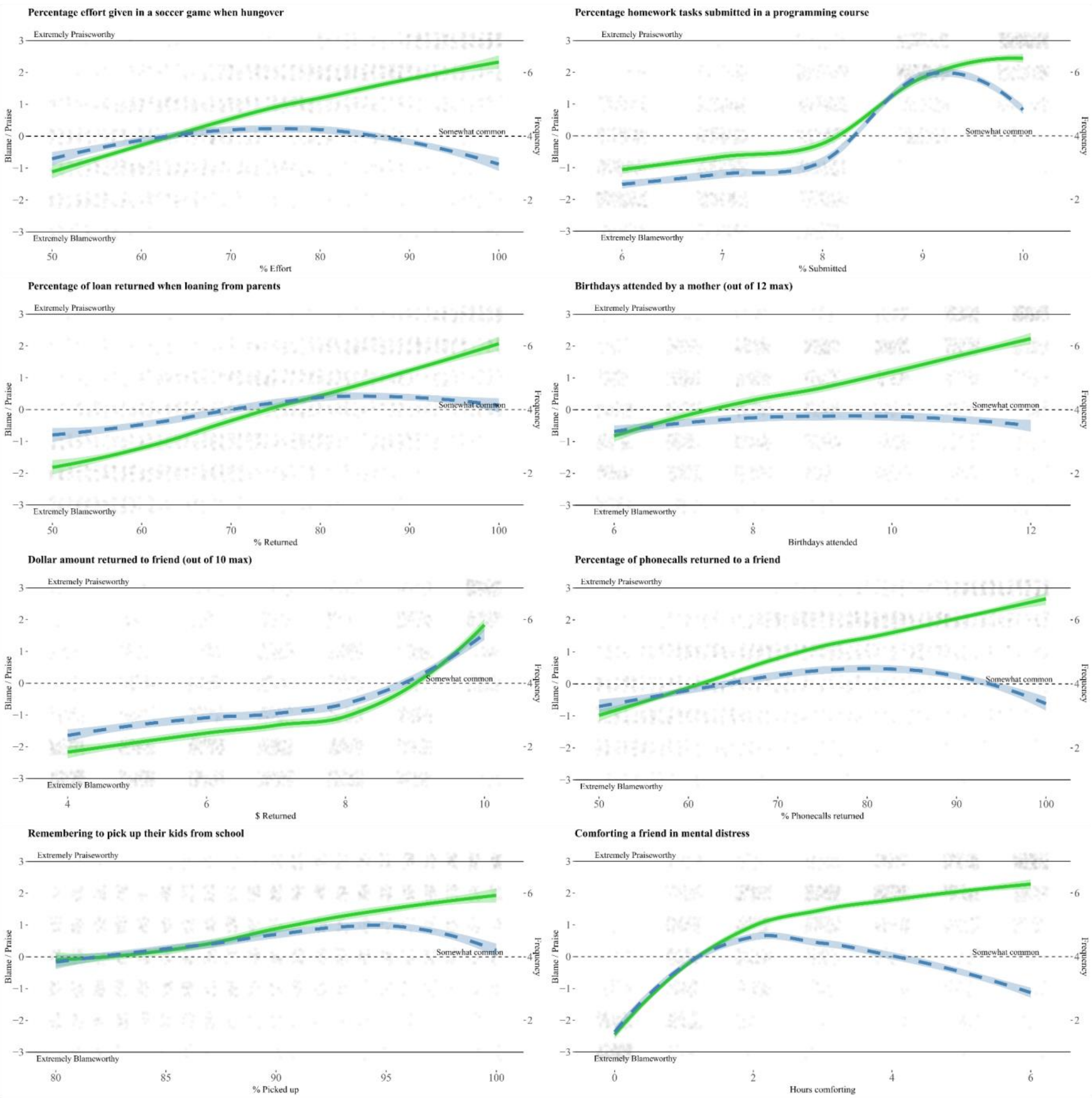
Wave 2: John puts in X% of effort. How praise or blameworthy do you find this behavior?"

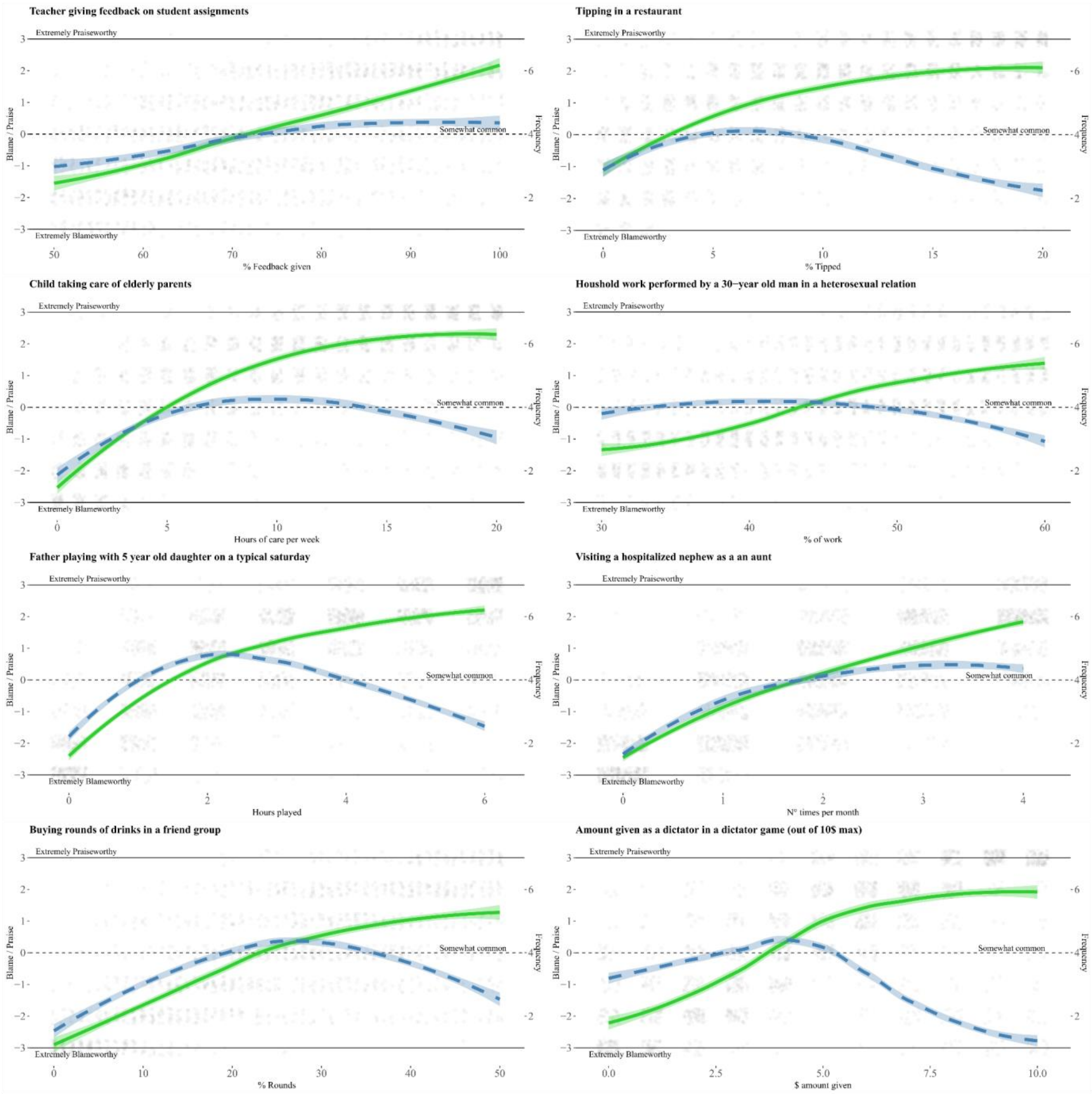
5.3. Results

For each of the vignettes, we started by fitting loess curves both to participants' frequency judgments as well as to their blame-praise judgments. Figure 4 displays the blame-praise curve associated with each behavior as well as the perceived frequency distribution.

Figure 4

Loess curves fitted on blame-praise judgments (in green) and frequency judgments (dotted blue line)





Note. Figure 4 continued.

In Figure 4, one sees clear evidence that the relationship between frequency and blame is different from the relationship between frequency and praise. Consider first the relationship between frequency and blame. Across the different behaviors, we see again and again that when people regard a certain quantity of a behavior as at least somewhat common, they tend not to regard that quantity as blameworthy. Indeed, it often happens that the exact point at which people start regarding behavior as not blameworthy is approximately the same as the point at which they start regarding it as somewhat common. One dramatic example is the programming vignette that probes how blame or praiseworthy it is for students to submit a certain percentage of the homework they are assigned with. The behavior is seen as blameworthy right up until the point it crosses from somewhat uncommon to somewhat common. The one exception to this trend is the blame-praise curve for what percentage of household chores young men should perform. While our participants did think it was somewhat common for a man to do only 40% of the household chores, they did also consider that to be blameworthy. We can only speculate as to why responses to this vignette might be diverging from the others, but perhaps this finding is related to the nature of our sample, i.e. mostly young, university-attending women.

When we turn to the relationship between frequency and praise, we see a completely different picture. Across the different behaviors, there are numerous cases in which people regard a certain quantity of that behavior as highly frequent yet also highly praiseworthy. For example, people think that it is common to spend two hours comforting a friend, but they still think that it is praiseworthy to do so. Similar results emerge in judgments about fathers spending time with their daughters, parents remembering to pick up their children at school, and many of the other behaviors.

For a more formal test of the effect of frequency, we switched to analyzing these data at the level of each individual participant. Thus, instead of aggregating data at the item level, we used each individual participant's frequency judgments to predict that participant's blame and praise judgments. By using a mixed model regression, we can control for the fact that participants were confronted with multiple vignettes and responded multiple times to each vignette.

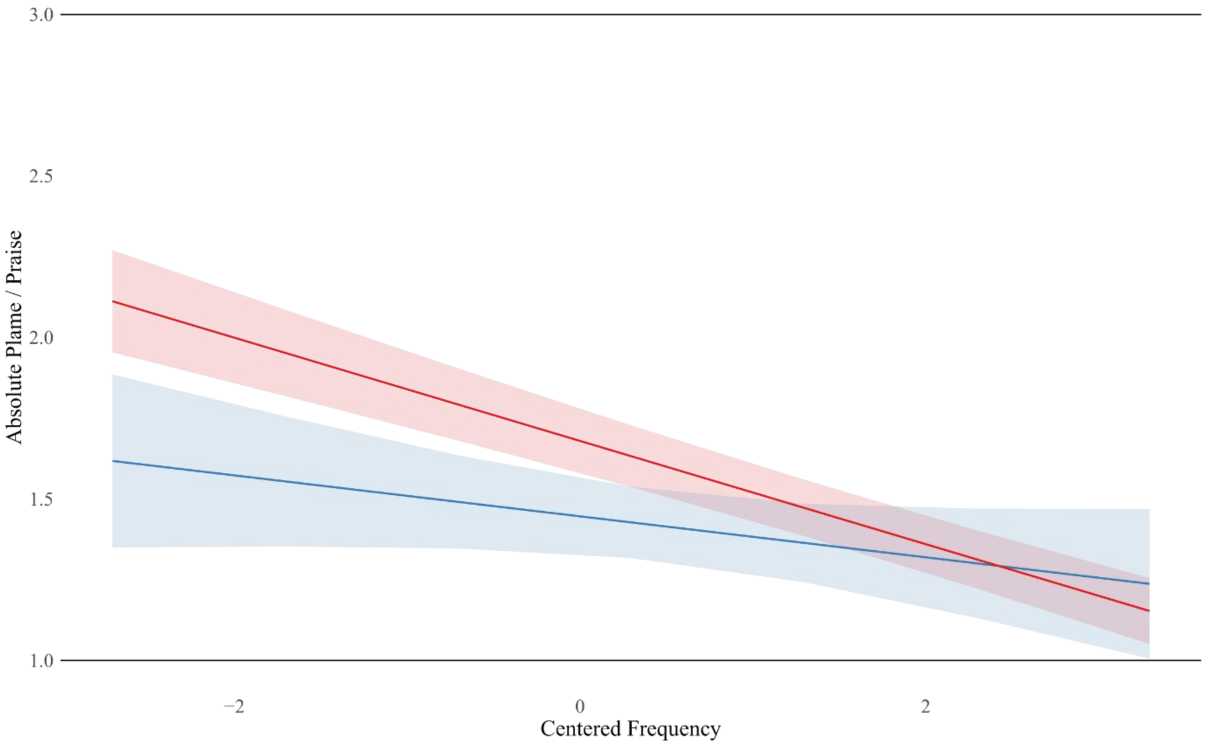
We wanted to test whether frequency has a stronger effect on judgments of blame versus praise or, framed differently, whether the magnitude of the effect of frequency is dependent on the valence of the judgment (blame versus praise). We did not preregister an analysis in this regard, but to test this, we decided to use the absolute values of blame and praise judgments as the dependent variable. We first ran a linear mixed model with random effects for participant ID and Vignette, and a maximal random effects structure. We included a centered version of participants' judgments of frequency, a dummy variable denoting whether a specific judgment is a blame or praise judgment and the interaction of these two variables as predictors.[2] In such a model, the interaction effect between valence and frequency is a direct test of whether valence impacts the magnitude of the effect of frequency. The interaction effect in this model was indeed statistically significant: $\beta = -0.09$, $t(17.14) = -2.40$, $p = .028$. However, the maximal model did not fully converge and collinearity diagnostics suggested that some issues with multicollinearity might be present (i.e., several generalized variance inflation factors were larger than 5). Accordingly, we also fitted a random intercept-only model. This model did converge, demonstrated no problems with multicollinearity, and yielded essentially the results. The interaction effect in this model was statistically significant as well, $\beta = -0.09$, $t(23747) = -10.37$, p < .001. Together, these results demonstrate that the magnitude of the effect of frequency is indeed larger for judgments of blame than for judgments of praise.

[2] Neutral judgments were coded as "praise" judgments. Repeating the analysis with neutral judgments coded as "blame" judgments leads to qualitatively similar results. Detailed results for this analysis are available in the supplementary materials.

Exploring the nature of this interaction effect, we conducted several follow-up tests for simple effects. For judgments of blame, the simple effect of frequency was present both when using a maximal random effects structure, $t(17.45) = -9.84$, $\beta = -0.32$, p < .001, and when using an intercept-only model, $t(7636.99) = -30.97$, $\beta = -0.34$, p < .001. For judgments of praise however, the simple effect was dependent on model specification. While a significant effect was present when using an intercept-only model, $t(16019.40) = -12.38$, $\beta = -0.10$, p < .001, this effect dropped from significance when using a maximal random effects structure, $t(17) = -1.65$, $\beta = -0.09$, p = .117. For a visual depiction of the interaction effect (using the more conservative maximal model), see Figure 5.

Figure 5

The effect of Frequency on Blame (red) and Praise (blue)



6. General Discussion

Many behaviors are not just good or bad. Instead, whether they are blame- or praiseworthy depends on how much of them we do or how we go about doing them. While a quick and timely reply to an email is praiseworthy, a slow and tardy one can be blameworthy. Such behaviors, where blame and praise are a function of an underlying behavioral continuum, are common: from spending time with your family to tipping at a restaurant. Yet it is unclear how people determine which instances of those behaviors are praiseworthy and which ones are blameworthy. Based on existing literature examining the relationship between statistical frequency and moral judgments (Bear & Knobe, 2017; Cialdini et al., 1991; Lindström, et al., 2018; Monroe et al., 2018; Phillips & Knobe, 2018; Schultz et al., 2008), we assumed that people's perceptions of the statistical, descriptive norms surrounding these behaviors would influence when they praise and when they blame.

We explored these phenomena in three studies. In the first set of studies, we confronted participants with vignettes describing everyday, common behaviors. We asked one group of participants to rate how blame- or praiseworthy they found a random instance of those behaviors and used these judgments to estimate blame-praise curves. This allowed us to determine the "neutral point" for each behavior: the point where people neither blame nor praise others for their behavior. We asked a second group of participants how much of these behaviors people typically do on average. The results indicated that the neutral point was not simply equal to the average but also that the neutral point was in some way related to the average. This finding suggests that the neutral point is not simply determined by the statistical average but that it is impacted in some way by statistical information.

In a second study, we confronted participants with a made-up behavior and provided them with descriptive information on how often people perform that made-up behavior. This gave us precise control over the frequency distribution associated with that behavior. We manipulated whether the made-up behavior was described as a positively or negatively valenced behavior to test whether statistical frequency would impact both types of behaviors similarly. We estimated a blame-praise curve within each condition and found that the neutral point was lower than the average for positively valenced behaviors and higher than the average for negatively valenced behaviors. In other words, when a behavior occurred frequently, it was not considered to be blameworthy. This suggested that frequency might impact blame, without impacting praise.

In a final study, we wanted to directly investigate the relationship between frequency and judgments of blame and praise. We again confronted participants with vignettes describing common, everyday behaviors and asked them to rate both the frequency and blame and praiseworthiness of multiple instances of each behavior. We found that as behaviors became more frequent, blame decreased in step with changes in frequency. In contrast, changes in frequency seemed to be much less related to whether behaviors were seen as praiseworthy or not. A formal test of these associative patterns confirmed a statistically significant interaction between frequency and the valence of judgments, confirming that frequency has an asymmetric effect on judgments of blame. These findings line up with a broader literature on moral cognition documenting numerous different asymmetries between blame and praise (Bostyn & Roets, 2016; Rozin & Royzman, 2001; Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Feldman, Wong, & Baumeister, 2016; Knobe & Doris, 2010) and extend prior results by Lindström et al. (2018) suggesting that people use a "common is moral"-heuristic by demonstrating this effect within the context of gradual behaviors.

6.1. Supplementary findings

The studies reported in the main text of this manuscript demonstrate our key finding: i.e., that frequency has an asymmetric impact on judgments of blame for gradual behaviors.

We have conducted several additional studies that have not been included here in the interest of keeping this manuscript concise but some of these studies could help answer some ancillary questions that might exist surrounding this work.

For instance, one might point out that the current work establishes only a cross-sectional association between frequency and judgments of blame but falls short of demonstrating a casual impact. Obviously, prior research using other paradigms has already demonstrated that frequency information can causally impact moral valuations (Cialdini et al., 1990), but additionally, it might be helpful to note that some of these additional studies involved manipulating the frequency information we provided participants with (using the experimental paradigm used in Study 2). These studies confirm that manipulating frequency information impacts the shape of a blame-praise curve.

Additionally, the studies included here do not address how the effect of frequency should be interpreted. Is this an effect of cumulative frequency or one of non-cumulative frequency (i.e., density). While the additional studies are not conclusive in this regard, we did conduct two additional experimental studies aimed at disambiguating this issue and found some evidence that blame-praise curves do not seem to map onto cumulative frequency.

6.2. Need for further computational research

The present manuscript provides evidence for a broad theory that suggests that people tend to think that something is not blameworthy when they perceive it to be frequent. While the studies included in the current manuscript provide a useful starting point, it is worth pointing out that this theory could be developed in much more detail. As of yet, we have merely demonstrated that if a behavior is perceived as occurring frequently, it will typically not be seen as blameworthy. However, it should be possible to say something more precise about how perceived frequency relates to both blame and praise. For instance, how exactly do changes in frequency impact judgments when people think that something *is* blameworthy. The exact mathematical relationship is still unclear. Accordingly, future work could build on these findings by developing a more refined computational account of how frequency and blame relate to one another and attempting to formalize the current theory in a way that makes possible more precise quantitative predictions.

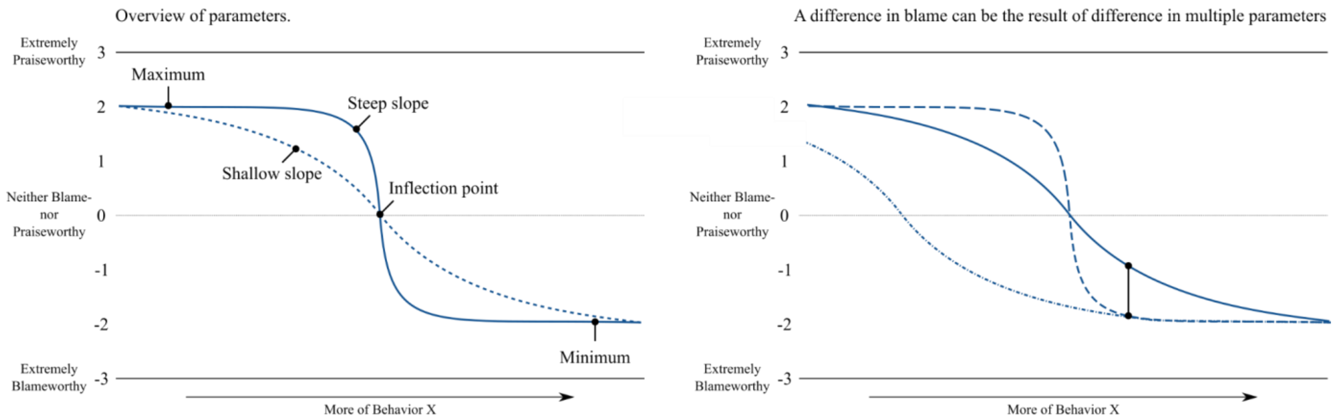
Of course, the current manuscript does not intend to imply that frequency is the sole driver of blame and praise judgments. It is safe to assume that numerous different factors will impact people's attributions. Realistically, there is little chance that we will soon be able to develop a formal theory that accurately models the impact of all of these different factors. Indeed, large amounts of residual noise were present in our data and while some of this residual noise could be explained as measurement error, much of it will reflect the impact of other theoretically meaningful factors. Some of these factors might have a direct impact on judgments of blame and praise, whereas others might moderate the impact of frequency. For instance, Gollwitzer, Marshall & Bargh (2019) found that people with a high aversion for pattern deviancy (e.g., disliking when one triangle is out of line in a row of triangles) tend to dislike members of groups that make up a statistical minority because they consider statistical minorities to be a sort of irregularity. It seems likely that people with a high aversion for pattern deviancy will also consider infrequent blameworthy behavior as meriting more blame than those with a low aversion for pattern deviancy for the same reason.

Still, even without an exploration of other factors that influence blame and praise attribution, it should be feasible to begin developing theories that model the impact of frequency on blame-praise curves. Such theories would not necessarily make direct predictions about the overall level of blame or praise assigned for any given behavior, but they would answer a simpler sort of question: If we manipulate people’s beliefs about the frequencies of certain behaviors while leaving everything else constant, what will be the effect of that manipulation? Developing a more refined computational theory will not only help to clarify the mechanisms that underlie the relationship between frequency and blameworthiness, but could ultimately inform interventions aimed at promoting ethical behavior.

6.3. On Blame-Praise Curves

In the current work, we have primarily estimated blame-praise curves by using a logistic model. Though we cannot preclude the possibility that other curve shapes might emerge in future research, at least in the present work, we were able to describe the full spectrum of judgments of blame and praise associated with a behavior through only four parameters. While the current work has focused on studying the relationship between frequency and neutral points, the methodology we developed should allow for an investigation into these other parameters as well. The left panel of Figure 6 displays an example curve as well as a visual depiction of the four parameters determining the shape of each curve.

Figure 6



Note. Left panel: overview of the different parameters in the logistic model. Right panel: A difference in blame can be the result of differences in multiple parameters.

Minima and maxima relate to the maximum levels of blame and praise that people give for a behavior. While our studies have not focused on analyzing what might determine these parameters, they do demonstrate that each behavior seems to settle at a specific minimum or maximum. Perhaps this could be considered an obvious finding, but the most egregious instance of responding late to an email is never perceived to be as blameworthy as marital infidelity. This suggests that minima and maxima are not a function of frequency but are related to some other aspect of these behaviors.

The other two parameters are similarly meaningful. The inflection point is the point at which the concavity of each curve changes and, as such, determines the location of each curve along its behavioral continuum. When a blame-praise curve is symmetrical around the axis of neutrality, the inflection points will be the same point as the neutral point. Interestingly, it appears as though many of the curves we studied were not symmetrical in that way. Finally, the rate of change determines how steep the slope of each curve is and thus how quickly blame switches to praise and vice-versa. Judgments of behaviors with a low rate of change shift only gradually, whereas those with a high rate of change can shift in an almost binary way from blame to praise. Investigating what determines these parameters will undoubtedly yield novel insights into the moral judgment of gradual behaviors.

There is another benefit to using a modeling approach such as this. The mathematical precision it offers could facilitate the formulation of more specific hypotheses when examining differences in how people blame or praise. The right panel of Figure 6 displays a possible example. Imagine that we uncover a difference in blame at the point noted in Figure 6: both the dotted and the dashed line diverge to the same extent from the full line. Yet, the root cause for this discrepancy is distinct: in comparison to the full line, the dotted line has a shifted inflection point whereas the dashed line has a shallower slope. Given that the values of these parameters are likely to be determined by different factors, the same difference in judgments of blame could be caused by different underlying mechanisms. By estimating blame-praise curves and describing them through their underlying parameters, we can add a level of mathematical precision to our thinking on blame and praise that will allow for more nuanced theories of blame and praise to arise.

6.4. Conclusion

In this paper, we explored the impact of statistical information on people's judgments about the amount of blame or praise one deserves for performing different quantities of a behavior. The results provided initial support for the hypothesis that frequency impacts blame. This hypothesis says that when people regard a certain quantity of a behavior as frequent, they tend not to regard that quantity as blameworthy. Further research on this hypothesis could develop the hypothesis in more technical detail and test it in additional studies.

In addition, moving beyond this hypothesis in particular, further research could continue to explore blame-praise curves. In the present studies, we looked only at the impact of statistical information, but further research could use similar techniques to explore the ways in which blame-praise curves can be shaped by other, very different factors.

7. References

Baer, J. S., Stacy, A., & Larimer, M. (1991). Biases in the perception of drinking norms among college students. *Journal of Studies on Alcohol*, 52(6), 580-586.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), 323-370.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, 167, 25-37.

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind? *Cognition*, 194, 104057.

Bostyn, D. H., & Roets, A. (2016). The morality of action: The asymmetry between judgments of praise and blame in the action–omission effect. *Journal of Experimental Social Psychology*, 63, 19-25.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 151-192). McGraw-Hill

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6), 1015-1026.

Clapp, J. D., & McDonnell, A. L. (2000). The relationship of perceptions of alcohol promotion and peer drinking norms to alcohol problems reported by college students. *Journal of College Student Development*, 41(1), 19–26.

Eriksson, K., Strimling, P., & Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, 129, 59-69.

Feldman, G., Wong, K. F. E., & Baumeister, R. F. (2016). Bad is freer than good: Positive–negative asymmetry in attributions of free will. *Consciousness and Cognition*, 42, 26-40.

Goldring, M. R., & Heiphetz, L., (2020). Social groups matter for norm-based moral judgments. *Journal of Experimental Social Psychology*, *91*, 104025.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, 35(3), 472–482.

Gollwitzer, A., Marshall, J., & Bargh, J. A. (2020). Pattern deviancy aversion predicts prejudice via a dislike of statistical minorities. *Journal of Experimental Psychology: General,* 149(5), 828.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196-217.

Knobe, J. & Doris, J. (2010). Responsibility. In *The Handbook of Moral Psychology*, eds. John Doris, et al., New York: Oxford University Press.

Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication theory*, 15(2), 127-147.

Lindström, B., Jangard, S., Selbing, I., & Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General*, 147(2), 228.

Miller, D. T., & Prentice, D. A. (2016). Changing norms to change behavior. *Annual Review of Psychology*, 67, 339-361.

Monroe, A. E., Dillon, K. D., Guglielmo, S., & Baumeister, R. F. (2018). It's not what you do, but what everyone else does: On the role of descriptive norms and subjectivism in moral judgment. *Journal of Experimental Social Psychology*, 77, 1-10.

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin*, 34(7), 913-923.

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, 114(18), 4649-4654.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, 33(1), 65-94.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences*, 23(12), 1026-1040.

Roberts, S. O., Ho, A. K., & Gelman, S. A. (2019). The role of group norms in evaluating uncommon and negative behaviors. *Journal of Experimental Psychology: General*, 148(2), 374.

Roberts, S. O., Gelman, S. A., & Ho, A. K. (2017). So it is, so it shall be: Group regularities license children's prescriptive judgments. *Cognitive Science*, 41, 576-600.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5, 269–320.

Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition*, 124(3), 325-333.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5), 429-434.

Wysocki, T. (2020). Normality: A two-faced concept. *Review of Philosophy and Psychology,* 1-28.