Graded Representations of Norm Strength

Bertram F. Malle (bfmalle@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences,
Brown University, 190 Thayer Street, Providence, RI 20906

Abstract

Previous work across multiple disciplines has shown that norms have a powerful impact on behavior. Little is known, however about how norms are represented in the mind. Here we examine whether people's norm representations come in reliably identifiable grades of strength. Classical models of norms distinguished between the broad deontic categories of prescriptions, permissions, and prohibitions. Four studies demonstrate that people consistently and consensually distinguish between deontic expressions that denote grades of prohibition (e.g., frowned upon < unacceptable < forbidden) and grades of prescription (e.g., called for < expected < required). Selecting terms that have mean ratings with nonoverlapping confidence intervals form a bipolar scale that allows researchers to measure prescriptions and prohibitions at five levels of norm strength each.

Keywords: social norms; moral psychology; deontic logic

No human community can exist without norms (Hechter & Opp, 2001). Norms structure social life in powerful ways, guiding individuals to align their behavior with community interests (Cialdini, Kallgren, & Reno, 1991; Krupka & Weber, 2009). Beyond this social-behavioral impact, however, little is known about how norms are represented in the mind. One point of consensus, in an extensive multi-disciplinary literature, is that norms come in three forms: as prescriptions (one should *A*), prohibitions (one must not *A*), and permissions (one may *A*). In much of this literature, however, no further differentiations are made within each type of norm.

We see this categorical treatment of norms most clearly in deontic logic, where single operators exist for prescription, prohibition, and permission (McNamara, 2006). A few exceptions exist in computational modeling of norms, such as Nickles (2007) who introduced agents' graded beliefs that others act in norm-conforming ways. Andrighetto et al. (2010) equipped their multi-agent systems with norms that can vary in prevalence and salience. But the vast majority of work on norms has treated norms as categorical. Even in the psychological literature on deontic reasoning (Beller, 2010) and in careful treatments of cognitive and social properties of norms (Indurkhya, 2016; Schmidt & Rakoczy, 2019; Sripada & Stich, 2006), graded norm strength is not addressed.

Intuitively, norms come in degrees. Talking on the phone in the train is less strongly prohibited than pushing another passenger to the ground; and paying for one's order in the coffee shop is more strongly prescribed than being polite. And even the classic category distinction between conventional and moral norms (Turiel, 1983) may be reframed as a pair of prototypes on a continuum. Moral psychology research also routinely presupposes that people finely differentiate *violations* of norms, measured on rating scales of badness, wrongness, or blame (Malle, in press), and occasionally even permissibility (Kneer & Machery, 2019; O'Hara, Sinnott-Armstrong, & Sinnott-Armstrong, 2010). It would be difficult to explain such graded judgments if norms came only in categories. However, norm violation judgments take into account outcome severity, mental states, even character; so any impact of potentially graded underlying norms is difficult to isolate.

A more direct test is needed to establish whether people represent norms in a graded way. One of the most powerful indicators of socially shared representations is language. I therefore take a linguistic approach that has been successful in other domains of cognition. Researchers have identified conceptual distinctions of time, space, causality, mental states, and personality by studying grammatical and lexical patterns of the domain of interest (de Villiers, 2007; Saucier & Goldberg, 1996; Talmy, 2000). Similarly, I examine whether people makes such conceptual distinctions for the domain of social and moral norms.

To identify these distinctions I propose to observe the way norms are often taught. Norms are instructions to act a certain way in a certain context, guided by the belief that the community demands one to act that way (Bicchieri, 2006; Brennan, Eriksson, Goodin, & Southwood, 2013; Malle, Scheutz, & Austerweil, 2017). This demand, I hypothesize, may come in degrees. As a result, a norm teacher (e.g., a parent teaching a child, a local teaching a visitor) would indicate not only the relevant context and action (e.g., "When entering a Japanese home, take your shoes off") but also signal the strength of the expectation, its *norm strength*. This could be done by saying, "You should…," "You must…," or "You absolutely have to." But the language available to the teacher is even richer than the small number of modals. This language characterizes prescribed actions as "expected," "recommended," or "mandatory"; likewise, it characterizes prohibited actions as "discouraged," "inappropriate," or "forbidden." A brief glance at a Thesaurus shows that these lists are actually much longer and provide a substantial stimulus set for the present project. The hypothesis of graded norm strength is that these sets of terms are not merely synonyms of each other but are systematically ordered along a continuum. I test this hypothesis for English speakers.

This approach shares some similarity with one that has occupied judgment and decision making researchers for decades: how people represent the probability continuum in

©2020 The Author(s). This work is licensed under a Creative
Commons Attribution 4.0 International License (CC BY).

verbal expressions (e.g., Brun & Teigen, 1988; Budescu & Wallsten, 1985; Hamm, 1991). The underlying phenomena, however, are somewhat different. In the case of probability expressions, there is a formal theory that grounds probability as a continuum of numbers. What researchers want to know is how closely and consistently people match their verbal expressions to these numbers. By contrast, the only formal theory of norms (deontic logic) treats them as categories, so we cannot examine analogous matches between verbal expressions and numbers. In addition, even if we postulate a continuum of norm strength, this continuum has no objective reality, no mathematical grounding. If there is any such a reality, it is a socially shared representation that must be inferred from other indicators—such as the differential use of linguistic expressions of norms.

The hypothesis of a socially shared representation of graded norm strength requires that speakers consistently order certain terms above others, and do so with high inter-judge agreement. I was initially agnostic about whether people's representations of norm strength would be tied to specific words (which could then be used as markers on a measurement scale) or to bundles of words that cluster at certain points of the continuum, without further within-cluster differentiation. However, by sampling a sufficiently large number of linguistic expressions we might be able to distinguish these two possibilities.

In four studies I examined consensus and robustness of graded norm strength among English language users. In Study 1, people considered over 20 terms of norm strength and ordered them from most prescribed to least prescribed, or most prohibited to least prohibited. Such a task may seem to pose high experimenter demands. However, there is no reason to believe that people would agree in their ordering of the terms if those terms are merely synonyms of each other, vaguely different versions of the same norm types. If, on the other hand, people do show substantial agreement, then we may infer that they share a graded interpretation of those terms. In Study 2 people considered a slightly different set of terms and made ratings on a numeric scale, which could then be compared to the rank orderings of Study 1. Study 3 repeated the rating approach with small adjustments in measurement and item sets. Study 4 assessed rankings and ratings in the same sample and combined the results to arrive at a measurement instrument that can be used to assess the strength of norms that people teach, learn, or contest.

Study 1

Methods

Participants. 120 participants were recruited from Amazon Mechanical Turk (no demographics were collected). Participants were randomly assigned to either a prohibition condition or a prescription condition. Eleven participants had empty records. Six participants in the prescription condition and nine in the prohibition condition had six or more missing responses, making their ranks difficult to compare to other participants'; these participants were excluded.

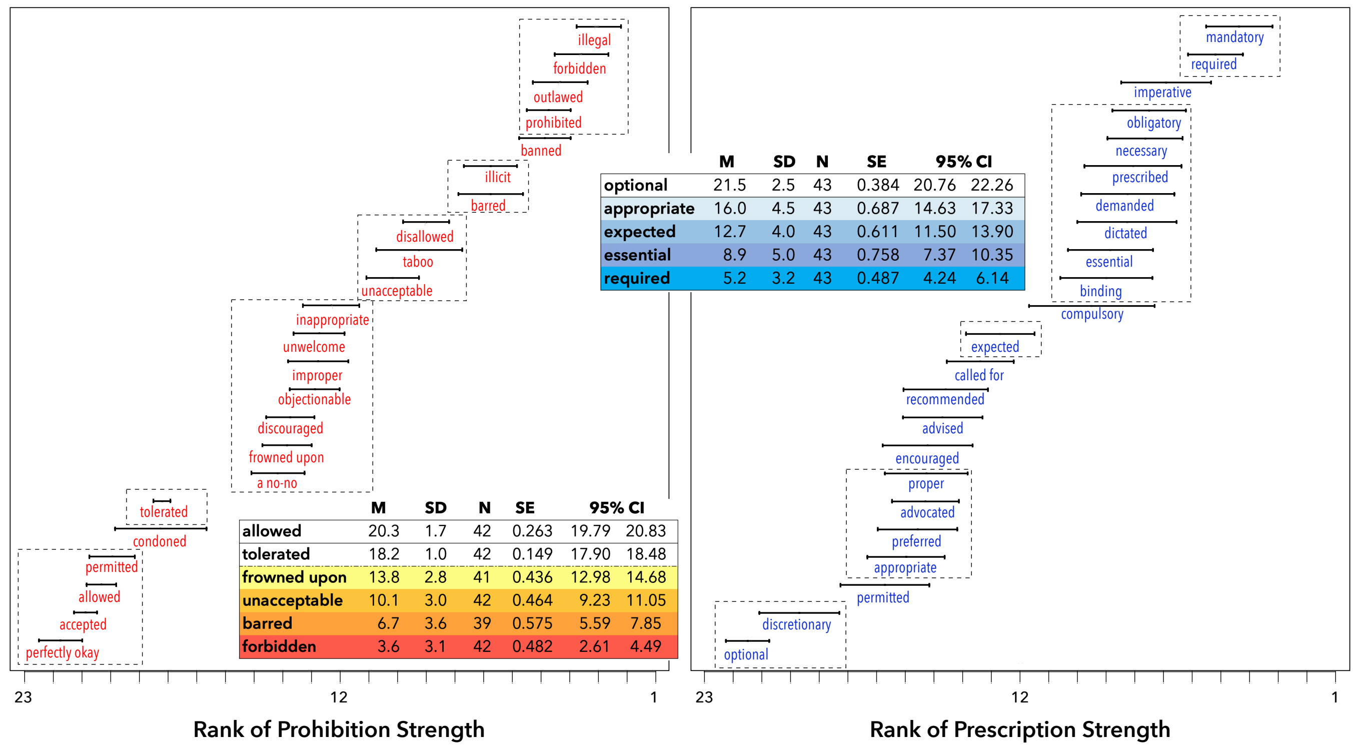
Stimulus design. To create a representative pool of norm strength terms I searched multiple dictionary and Thesaurus sources to identify around 20 terms in each of the categories of prescriptions and prohibitions, with the constraint that the words must function as adjectives to the term *action*. I identified 20 prescription terms and 19 prohibition terms and added three words to the prescription set and four words to the prohibition set. These additional words served as anchors and reached into the permission range (e.g., perfectly okay, permitted, optional). See Figure 1 for all terms used in the study.

Procedure. In the prohibition condition participants read: "Some actions are not prohibited (e.g., sitting in the back of a movie theater); some actions are very strongly prohibited (e.g., killing another person out of hatred); most actions lie somewhere in the middle." In the prescription condition they read: "Some actions are not prescribed (e.g., sitting in the back of a movie theater); some actions are very strongly prescribed (e.g., feeding your infant); most actions lie somewhere in the middle." Participants then ranked the 23 terms (presented alphabetically) for their condition by dragging and dropping them from the left side of the screen into a box on the right and ordering them from "very strongly [prescribed / prohibited]" at the top to "not [prescribed / prohibited]" at the bottom, with "the other ones lying in between." Thus, only abstract anchors were displayed, and the in-between space was left vague. If participants were not sure about the meaning of a word they could place it in a separate box labeled "I don't know the meaning of the word."

Analysis. We computed two indicators of inter-judge agreement. The individual-group agreement, $\bar{r}_{iG}$, measures how well individuals' judgments stand in for the group, or how well the group as a whole represents any individual. Second, the intraclass correlation ICC(2,1) captures the generalizability from one randomly drawn individual to another individual drawn from the same population—a very high bar of agreement. Most important, we recorded the ordering of terms along the hypothesized strength scale and defined boundaries between terms as nonoverlapping 95% confidence intervals (CI = $M \pm 1.96 \cdot s_e$).

Results

Agreement. In the prescription condition, the correlation of any judge with the group as a whole was $\bar{r}_{iG}$= 0.60, with 69% of r_{iG} values above .50. Four judges (9%) had negative r_{iG} values; excluding those improved $\bar{r}_{iG}$ to 0.69, with 76% of r_{iG} values now above .50. The ICC(2,1) was .37 before excluding the four judges with negative values and .48 after. In the prohibition condition, $\bar{r}_{iG}$ was 0.76, with 87% of r_{iG} values above .50. Five judges (11%) had negative values; excluding those improved $\bar{r}_{iG}$ to 0.88, now with 98% of r_{iG} values above .50. ICC(2,1) was .59 before exclusion and .79 after. Subsequent analyses were based on the mean ratings of items across judges that had a positive r_{iG} value, but the results are highly similar with or without this restriction.



	M	SD	N	SE	95% CI	
optional	21.5	2.5	43	0.384	20.76	22.26
appropriate	16.0	4.5	43	0.687	14.63	17.33
expected	12.7	4.0	43	0.611	11.50	13.90
essential	8.9	5.0	43	0.758	7.37	10.35
required	5.2	3.2	43	0.487	4.24	6.14

	M	SD	N	SE	95% CI	
allowed	20.3	1.7	42	0.263	19.79	20.83
tolerated	18.2	1.0	42	0.149	17.90	18.48
frowned upon	13.8	2.8	41	0.436	12.98	14.68
unacceptable	10.1	3.0	42	0.464	9.23	11.05
barred	6.7	3.6	39	0.575	5.59	7.85
forbidden	3.6	3.1	42	0.482	2.61	4.49

Figure 1. Terms of prohibition (left) and prescription (right), ordered by average ranks of rated norm strength (Study 1). Line widths are 95% confidence intervals. Dotted rectangles mark clusters of terms where the top term of one cluster is nonoverlapping with the bottom term of the next cluster. Inserted boxes display candidate scales of nonoverlapping terms.

Clusters of norm strength. Figure 1 displays the 23 terms within each condition, from lowest to highest average rank, where the lowest-ranked items are the permission terms. We can identify clusters of distinct items where at least the highest-ranked term in one cluster and the lowest-ranked term of the next cluster have nonoverlapping CIs. Among prescriptions, we see four clusters above the permission cluster; among prohibitions, we see four clusters plus the term *tolerated* in its own position above the permission cluster. Several preliminary scales of norm strength can be formed by ordering nonoverlapping terms; the inserts in Figure 1 show two such scales, one for prescriptions, the other for prohibitions.

Discussion

The ranking data in Study 1 suggest that normative strength terms elicit substantial inter-judge agreement. As a comparison, I analyzed data from recent studies that contained valence judgments (good-bad) about a variety of morally significant behaviors. Positive behaviors elicited an $\bar{r}_{iG}$ of 0.68 and an ICC(2,1) of 0.46, nearly identical to the agreement for prescription terms in the present study. Negative behaviors elicited an $\bar{r}_{iG}$ of 0.72 and an ICC of 0.51, which are actually lower than the agreement for prohibition terms in the present study.

Four clusters of prescriptions and four clusters of prohibitions emerged. The term *tolerated* is either a weak prohibition (making it five clusters on that side) or a permission with prohibition connotations. Before drawing conclusions about the number or composition of clusters and selecting terms for a candidate scale of graded norm strength, we need to conceptually replicate and sharpen the evidence. With this aim, Study 2 featured three changes. First, because rankings force people to order items, I probed whether similar patterns would emerge when people rate expressions of norm strength on rating scales. Second, I cut the number of items to be judged in half and created two sets of items within each condition. Akin to a construction of parallel forms, this approach examines the comparability of strength clusters across different terms and thus tests whether a given term's perceived strength changes in the context of other terms (Asch, 1946; Hamm, 1991).

Study 2

Methods

Participants. Participants from Amazon Mechanical Turk were randomly assigned to either the prohibition condition or the prescription condition and, within each, to item Set 1 or 2. I had set recruitment to 100 participants per condition to narrow the standard errors somewhat compared to Study 1. Three participants in the prohibition condition and four in the prescription condition gave identical ratings across all items and were excluded. Sample sizes for analysis were therefore 97 and 100 for the two item sets in the prescription condition and 100 and 96 in the prohibition condition.

Stimuli. To select prescription items, Study 1 served as the starting point. I first eliminated two items with the highest standard deviations (e.g., *dictated, compulsory*) and ones that were rare in spoken use, according to the Corpus of Contemporary American English (COCA; Davies, 2008): *advocated, binding, obligatory, imperative*. Then I ordered the remaining items according to the mean ranks in Study 1 and attempted to identify pairs of items within each strength cluster that had highly similar average ranks and could be separated into parallel sets. For the two larger clusters I selected a second pair of items so as to explore possible differentiation within cluster. I included three previous items (*advised, prescribed, mandatory*) in both item sets and added two new items (*insisted on, commanded*) to both sets. These five "twin" items tested whether the context of other items affects perceived norm strength (Hamm, 1991). Finally, I included three permission anchors (optional, discretionary, totally okay). In total, prescription Set 1 contained 11 items, Set 2 contained 10.

To select prohibition items I also eliminated two Study 1 items with high standard deviations (e.g., *condoned, taboo*) and three of rare spoken use (*objectionable, unwelcome, a no-no)*. I also omitted *illicit* because of its predominant legal use. Then I proceeded in the same manner as with prescription items to create parallel item sets. I added one new item (*disapproved*) to appear in both sets and included four items that occurred in both sets (*tolerated, frowned upon, illegal, prohibited*). I also included permission anchors (accepted, permitted, and those that were used in the prescription condition). In total, prohibition Set 1 contained 13 items, Set 2 contained 12 items.

Procedure. Participants received similar instructions as in Study 1. In the prescription condition they read: "In this study we are interested in the kinds of words people use to express how strongly prescribed a given action is. On the next page you will see a list of words that people might use to indicate to which *degree* a given action is prescribed." Participants then rated each term on a scale from 1 to 10. Prohibitions had scale anchors of "this means not at all prohibited" (1) and "this means completely prohibited" (10). Prescriptions had anchors of "this means not at all prescribed" (1) and "this means very strongly prescribed" (10). Items within sets were presented in randomized order.

Results

I first examined items that were included in both Studies 1 and 2 to test how comparable rankings and ratings were overall. The correlations of average rank and average rating were r = .98 for both prescriptions and prohibitions. Thus, differentiation along the strength dimension is overall assessed nearly identically through rankings or ratings.

Agreement. The $\bar{r}_{iG}$ and ICC values across item sets were as high or higher than those in Study 1. For prescription terms, $\bar{r}_{iG}$ = .75 and ICC(2,1) = 0.54. After removing participants with negative r_{iG} values (10 out of 197, 5.1%), $\bar{r}_{iG}$ = .80 and ICC(2,1) = .59. For prohibition terms, $\bar{r}_{iG}$ = .75 and ICC(2,1) = .60. After removing participants with negative r_{iG}values (13 out of 192, 6.8%), $\bar{r}_{iG}$ = .87 and ICC(2,1) = .78. All subsequent analyses were based on the mean ratings of items across judges with a positive r_{iG}value.

Clusters of norm strength. I first examined the full range of distinct norm terms across item sets, averaging the ratings of twin terms that occurred in both sets. The resulting 16 prescription terms and 19 prohibition terms formed patterns highly similar to those in Study 1. Among prescriptions, four distinct clusters emerged above the permission level. As in Study 1, more than one scale of nonoverlapping prescription terms could be formed, but one nearly identical to Study 1 is: (optional) < *appropriate* < *expected* < *insisted on* < *required*. Among prohibitions, four distinct clusters above the permission level emerged as well. One of several possible scales of nonoverlapping prohibition terms is identical to that in Study 1: *(optional)* < *(tolerated)* < *frowned upon* < *unacceptable* < *barred* < *forbidden*, though the last two have a slight overlap of CIs: [8.63; 9.29] and [9.06; 9.58].

Two additional analyses suggest that the data represent cognitively distinct levels of norm strength that are robust across samples and despite varying item contexts. Specifically, average ratings of terms that appeared in both sets (twins) were statistically indistinguishable (M_{diff} = 0.09). Moreover, four out of five prescription twins and three out of five prohibition twins had mean ratings immediately succeeding one another in the overall rating order, even though they were judged by different people and presented in the vicinity of different items. Conversely, three pairs of distinct prescription terms had statistically indistinguishable mean strength ratings within pair (e.g., *called for* = *expected*) but nonoverlapping confidence intervals between pairs; they could therefore substitute for one another in parallel forms of graded norm strength scales. Likewise, three pairs of prohibition terms qualified in the same way for parallel forms (e.g., *unacceptable* = *disallowed*).

Study 3

Study 2's results were highly similar to those of Study 1, even though one used ratings, the other used rankings. The rating scales appeared to slightly reduce the separation of terms at the top end, so in Study 3 I attempted to reduce this upper-range compression by employing a 1-11 scale, which might encourage people to treat the number 11 more distinctly from the remaining scale points. I also made small changes to the item sets. In the prescription condition, I omitted *insisted on* and *commanded* (because I noticed that these terms are more often used as verbs, not adjectives) and brought back *necessary* in their place. I also added *encouraged* to potentially fill a gap at the lower end of prescriptions. In the prohibition condition, I replaced *disapproved* (more often used as an evaluative verb) with *banned,* and I added *encouraged* and *objectionable* to fill a gap at the lower end. None of these changes were expected to have significant impact, but I wanted to further explore category boundaries and capture a natural and wide range of norm strength terms.

The instructions explicitly placed participants in the situation of entering a new community and learning the community's norms. They then evaluated how strongly the community wants its members to (not) perform a behavior when the behavior is, say, "required" or "forbidden." After they worked either on the prescription set (12 items) or the prohibition set (13 items), participants moved to a new screen page and provided judgments of seven permission terms (e.g., *optional, acceptable*). This short rating task examined whether permission terms can be reliably distinguished in the neutral space between prescription and prohibition. Accordingly, participants rated how much each permission term indicates whether the behavior is slightly discouraged (-1), slightly encouraged (+1) or truly neither (0).

Recruited on Amazon Mechanical Turk, 202 participants (57.1% male, Mean age = 36.2, 43% with a Bachelor degree) were randomly assigned to rate 12 prescription terms or 13 prohibition terms (everybody completed ratings of the permission terms). Individuals who had a rating range of 0 or 1 were excluded (one in the prescription condition, five in the prohibition condition, 10 in the permission condition). Inter-judge agreement was somewhat lower than in Studies 1 and 2 (see Table 2). Agreement for permissions was respectable, considering the narrow range of distinctions.

Within prescription and prohibition conditions, I selected terms that formed non-overlapping 95% confidence intervals (see Table 3). Prohibitions separated into five distinct levels (suggesting a beneficial effect of the 1-11 scale), whereas prescriptions maintained the consistent four levels. Most permission terms leaned toward the prescription side of the scale, but three terms formed a mini-scale of permissions, with *permitted* being clearly on the prescription side *(M* = 0.41, [.30, .51]), *discretionary* at the 0 point *(M* = -0.05, [-.14, .04])*,* and *tolerated* being on the prohibition side *(M* = -0.25, [-.37, -.12]).

Study 4

The previous studies showed that distinct levels of norm strength can often be captured by more than one term (thus suggesting that norm strength has a psychological reality, not merely semantic stability), but they also showed that the rank ordering of terms across different levels is highly consistent. In Study 4 I aimed to develop a stable scale of graded norm strength that contained the most sharply differentiated terms, attempted to better fill the lower end of prescriptions, and thus produce five levels of norm strength on both the prohibition side and the prescription side.

Table 2. Inter-judge agreement analysis for Study 3

	Initial sample		After exclusion*		Excluded / Total
	$\bar{r}_{jG}$	*ICC*	$\bar{r}_{jG}$	*ICC*	*N*
Prescriptions	.57	.35	.73	.50	15/101
Prohibitions	.68	.48	.71	.51	3/95
Permissions	.51	.26	.63	.38	22/186

* Exclusion of participants with negative r_{iG} values

Table 3. Nonoverlapping prescription and prohibition terms in Study 3

	95% Confidence Interval	
	low	high
Prescriptions (*n* = 86)		
encouraged	5.40	6.30
called for	6.67	7.53
essential	8.24	9.14
required	9.46	10.12
Prohibitions (*n* = 95)		
frowned upon	3.89	4.85
inappropriate	4.99	6.05
unacceptable	6.10	7.20
barred	7.64	8.52
forbidden	8.57	9.47

Note: Means based on participants with positive r_{iG} values

Methods

For item selection I began with the nonoverlapping terms emerging from Study 3 and added a few terms to better cover the lower range of each side (e.g., weaker than *encouraged*, which starts at 5.73; see Table 3). Each participant completed both a ranking task and then a rating task, and I combined the results to select terms that are distinct in both tasks and thus best represent robust levels of norm strength.

To identify possible low-end prescriptions, I included the terms *approved* and *suggested,* along with the four terms retained from Study 3 (see Table 3). We also added a term at the top (*mandatory,* which had the highest mean in Study 3 but overlapped substantially with *forbidden*) and a middling term (*expected*, which was well differentiated in Studies 1 and 2), yielding eight terms in total. On the prohibition side, I included the five terms retained from Study 3 (see Table 3), one candidate low-end item (*undesirable*), and two in the middle (*objectionable, unacceptable*), also yielding eight.

201 participants were recruited on Amazon Mechanical Turk (no demographics were collected). Three participants had a rating range of 0 or 1 and eight placed four or more terms into the "I don't know" box; they were excluded, leaving 190 for analysis. Randomly assigned to either the prescription or prohibition condition, participants first rank-ordered the eight terms and then, on a separate page, assigned ratings to each term on a 1-11 scale. In the ranking task, modeled after Study 1, participants dragged words into a box and ordered them such that "the top one (1) means *Most strongly [prohibited]/ [prescribed]* and the other ones below it are, in decreasing order, less strongly [prohibited]/ [prescribed]." They also had the option to place a word into the "I don't know the meaning of the word" box. In the rating tasks following the ranking, participants were again asked to imagine that someone was teaching them the norms of a new community they are joining and to guess, from the person's expressions, how strong the norms are. Specifically, each item read, "When somebody says, 'Doing this is [e.g.,

essential; barred], how strongly does the community [want its members to perform] / [try to deter its members from performing] this behavior?" They then selected a number from 1 (A little bit) to 11 (Extremely).

Results

Inter-judge agreement was very high ($\bar{r}_{jG}$ = .74 to .83, *ICC* = .54 to .69) across prohibitions, prescriptions, rankings, ratings, and even better after excluding a small number of participants with negative r_{iG}values (8/190, 4%). Thus, people showed strong consensus on the relative norm strength of the eight presented terms. I then selected the most distinct and precise terms by putting rating and ranking data side by side and identifying terms that had nonoverlapping confidence intervals in *both* assessment methods. The selected terms are shown in Figure 2, forming a bipolar scale of norm strength from the strongest prescription on top to the strongest prohibition on the bottom, with permissions (from Study 3) in between. The attempt to fill the lower end of prescriptions was successful with the addition of *suggested*, while most other terms are familiar from the first three studies as marker variables for degrees of norm strength.

General Discussion

Nearly all formal representations of norms rely on categorical deontic concepts of prescription and prohibition. Psychological research, too, has (often implicitly) treated norms as categorical. Four studies showed that people consensually and robustly distinguish among strength levels of prohibition norms, as denoted by characteristic linguistic phrases such as *frowned upon* or *forbidden*; and they distinguish equally well among strength levels of prescription norms, as denoted by characteristic linguistic phrases such as *suggested* or *required*. Study 4 generated a bipolar scale using specific terms that distinguish (with nonoverlapping confidence interval) five strength levels on the prescription side and five strength levels on the prohibition side. Three permission terms from Study 3 can be included, which may lean slightly toward the prohibition side (*tolerated*) or slightly toward the prescription side (*permitted*).

Even though specific terms were selected for this scale, earlier studies showed that several different combinations of terms can constitute levels of norm strength, and always four or five such levels. This substitutability of terms provides some confidence that the present results are not mere vocabulary distinctions but refer to psychologically real representations of norm strength.

Many open questions remain about these representations and their measurement. For example, we do not yet know whether the representations (and their associated terms) are constant across domains of morality. Is the domain of harm more differentiated than the domains of sanctity or loyalty? Do the specific terms change their meaning when combined with different acts (*cf.* Brun & Teigen, 1988)? That is, is an *unacceptable* interpersonal harm always stronger than an *objectionable* harm, no matter what the specific norm-violating actions are that people consider? Furthermore, we do not know how consistent people are in using the terms across contexts and over time. We can expect, however, that intra-personal inconsistency would limit inter-judge agreement, so given the high levels of agreement we have found across studies, intrapersonal consistency should be no lower (at least in the settings that the present studies created). We also know too little yet about the scale properties of the norm strength scale developed in Study 4. The mean ratings for the strength terms were very similar no matter which other terms they were surrounded by. But do these mean levels reflect an ordinal or interval scale? Is *encouraged* stronger than *suggested* to the same degree (about 1 point; see Figure 2) that *called for* is stronger than *encouraged* (also 1 point)?

In everyday life, when people learn norms of another community or reaffirm the norms of their own community, they have a rich vocabulary available that allows them to interpret or signal how strong a normative expectation is and thus how harsh the sanctions should be in case the norm is violated. Thus, the scale presented here should be a strong predictor of moral judgments: When we know the strength of a norm we also know how bad or wrong it is to violate it. The norm scale may also serve to explain moral disagreement, if it can be shown that one person assigns a higher norm strength to a certain action than another person does. Finally, the scale may set the stage for new formal models of deontic reasoning, which would require either far more operators or a parameter of strength for the main operators (Malle, Bello, & Scheutz, 2019). Such formal models will be needed to equip artificial agents with norm capacities. AI assistants or robot companions, if they co-exist in human communities, will need to know the norms of these communities—not just whether something is prescribed or prohibited, but *how strongly* it is prescribed or prohibited. Reasoning within classic category systems will no longer be sufficient; the present findings suggest that such reasoning must meet people's rich, graded representations of norm strength.

	Mean	95% CI	
required	**10.30**	10.09	10.51
expected	**8.87**	8.56	9.18
called for	**7.31**	6.89	7.73
encouraged	**6.31**	5.88	6.74
suggested	**5.36**	4.93	5.79
permitted			
discretionary			
tolerated			
frowned upon	**4.25**	3.94	5.08
objectionable	**5.89**	5.48	6.30
unacceptable	**7.31**	6.84	7.78
barred	**9.47**	9.08	9.86
forbidden	**10.58**	10.40	10.76

Figure 2: Mean ratings (on a 1-11 scale) of final selected terms of graded norm strength in prescriptions (top 5), prohibitions (bottom 5), along with 3 permission terms.

References

Andrighetto, G., Villatoro, D., & Conte, R. (2010). Norm internalization in artificial societies. *AI Communications*, *23*, 325–339.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*, 258–290. doi:10.1037/h0055756

Beller, S. (2010). Deontic reasoning reviewed: psychological questions, empirical findings, and current theories. *Cognitive Processing*, *11*, 123–132. doi:10.1007/s10339-009-0265-z

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. New York, NY: Cambridge University Press.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining norms*. New York, NY: Oxford University Press.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, *41*, 390–404. doi:10.1016/0749-5978(88)90036-2

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, *36*, 391–405. doi:10.1016/0749-5978(85)90007-X

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 24, pp. 201–234). San Diego, CA: Academic Press.

Davies, M. (2008). The Corpus of Contemporary American English (COCA): One billion million words, 1990-2019. Retrieved from https://www.english-corpora.org/coca/

de Villiers, J. (2007). The interface of language and theory of mind. *Lingua. International Review of General Linguistics. Revue Internationale De Linguistique Generale*, *117*, 1858–1878. doi:10.1016/j.lingua.2006.11.006

Hamm, R. M. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes*, *48*, 193–223. doi:10.1016/0749-5978(91)90012-I

Hechter, M., & Opp, K.-D. (Eds.). (2001). *Social norms*. New York, NY: Russell Sage Foundation.

Indurkhya, B. (2016). A cognitive perspective on norms. In J. Stelmach, B. Brożek, & Ł. Kwiatek (Eds.), *The normative mind* (pp. 35–64). Krakow: Copernicus Center Press.

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348. doi:10.1016/j.cognition.2018.09.003

Krupka, E., & Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, *30*, 307–320. doi:10.1016/j.joep.2008.11.005

Malle, B. F. (in press). Moral judgments. *Annual Review of Psychology*, *72*.

Malle, B. F., Bello, P., & Scheutz, M. (2019). *Requirements for an artificial agent with norm competence*. *Proceedings of 2nd ACM conference on AI and Ethics (AIES'19)* (pp. 21–27). New York, NY: ACM. doi:10.1145/3306618.3314252

Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, & G. S. Virk (Eds.), *A World with Robots: International Conference on Robot Ethics: ICRE 2015* (pp. 3–17). Cham, Switzerland: Springer International Publishing.

McNamara, P. (2006). Deontic logic. In D. M. Gabbay & J. Woods (Eds.), *Handbook of the History of Logic* (Vol. 7, pp. 197–288). North-Holland. doi:10.1016/S1874-5857(06)80029-4

Nickles, M. (2007). *Towards a logic of graded normativity and norm adherence*. In G. Boella, L. van der Torre, & H. Verhagen (Eds.), *Normative Multi-agent Systems: Dagstuhl Seminar Proceedings*. Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

O'Hara, R., E., Sinnott-Armstrong, W., & Sinnott-Armstrong, N. A. (2010). Wording effects in moral judgments. *Judgment and Decision Making*, *5*, 547–554.

Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives.* (pp. 21–50). New York, NY: Guilford Press.

Schmidt, M. F. H., & Rakoczy, H. (2019). On the uniqueness of human normative attitudes. In N. Roughley & K. Bayertz (Eds.), *The normative animal? On the anthropological significance of social, moral, and linguistic norms* (pp. 121–138). New York: Oxford University Press. doi:10.1093/oso/9780190846466.003.0006

Sripada, C. S., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind (Vol. 2: Culture and cognition)* (pp. 280–301). New York, NY: Oxford University Press.

Talmy, L. (2000). *Toward a cognitive semantics. Volume I: Concept structuring systems*. Cambridge, MA: MIT Press.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, England: Cambridge University Press.

Acknowledgments

This research was funded in part by grants from the Office of Naval Research (ONR), N00014-16-1-2278, and from the Defense Advanced Research Projects Agency (DARPA), SIMPLEX 14-46-FP-097. The opinions expressed here do not necessarily reflect the views of ONR or DARPA.