Lay evaluations of police and civilian use of force: Action severity scales

Bradley D. Celestin & John K. Kruschke

Indiana University

In press: Law & Human Behavior

Author Note

Bradley D. Celestin and John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University

Portions of this work have been previously presented at the Society for Personality and Social Psychology annual conferences (2017, 2018, 2019) and the associated pre-conferences on Justice and Morality and Law and Psychology. Additionally, separate analyses of these data are reported in a companion manuscript (Celestin & Kruschke, 2019). Although the two manuscripts are based on the same data, the analyses are (a) mathematically different and (b) focus on very different aspects of the models. The latent severity scales reported here remain essentially unchanged when using more elaborate models, so we present the most parsimonious model here in order to maximize the clarity of the latent scale derivation and the implications of those results.

Correspondence can be addressed to Bradley D. Celestin, Department of Psychological and Brain Sciences, Indiana University, 1101 E 10th St., Bloomington IN 47405-7007, or via electronic mail to bradcelestin@gmail.com.

Abstract

In modern societies, citizens cede the legitimate use of violence to law enforcement agents who act on their behalf. However, little is known about the extent to which lay evaluations of forceful actions align with or diverge from official use-of-force policies and heuristics that officers use to choose appropriate levels of responsive force. Moreover, it is impossible to accurately compare official policies and lay intuitions without first measuring the perceived severity of a set of representative actions. To map these psychometric scale values precisely, we presented participants (N=411 MTurkers, N=395 undergraduates) with minimal vignettes describing officer and civilian actions that span the entire range of force options (from polite dialogue to lethal force), and asked them to rate physical magnitude and moral appropriateness. We used Bayesian methods to model the ratings as functions of simultaneously estimated scale values of the actions. Results indicated that the perceived severity of actions across all physical but non-lethal categories clustered tightly together, while actions at the extreme levels were relatively spread out. Moreover, less normative officer actions were perceived as especially morally severe. Broadly, our findings reveal divergence between lay perceptions of force severity and official law enforcement policies, and they imply that the groundwork for disagreement about the legitimacy of police and civilian actions may be partially rooted in the differential way that action severity is perceived by law enforcement relative to civilian observers.

Keywords: Use of force, police, morality, violence, Bayesian, latent scale estimation

Public Significance Statement: We report public perceptions of moral and physical severity for a set of police and civilian forceful actions that are representative of the complete range of force options, from polite dialogue to lethal force. Our results reveal that non-expert perceptions of police and civilian violence are quite different from official police use-of-force policies. Our results also provide researchers with a unique tool and foundation for future investigations into discrepancies between non-expert evaluations and legal models of legitimate use of force.

Lay evaluations of police and civilian use of force: Action severity scales

From 1998 to 2017, there were 2.54 violent crimes committed per minute (on average) in the United States (Federal Bureau of Investigation, 2017). Responding to this criminal violence often requires some form of socially-sanctioned counter-violence. But average citizens lack the ability and equipment to apply violence in a measured and precise manner, and even if citizens possessed the necessary skills and tools, doing so requires that they expose themselves to elevated risk and eliminates the neutrality of third-party intervention. Consequently, the bargain that citizens make is to willingly cede the use of physical force to a ruling government (with limited exceptions for immediate defense of self or others). In exchange, citizens gain the ability to call on specialized government agents to use force on their behalf for two primary purposes: to maintain law and social order (e.g., to arrest resisting criminal suspects); and to provide skilled protection from other violent threats (e.g., active shooters). Ultimately, the result of this social agreement is a more peaceful and just society than one controlled by cyclical revenge and domination by the largest, strongest, and most brutal (Pinker, 2011). Nevertheless, the tragic nature of any use of violence demands that those who employ it professionally do so with maximum solemnity, care, and public accountability. In modern society, the legitimate use of violence is entrusted to commissioned law enforcement officers (Friedman, 2017).

Ordinal Levels of Force and Heuristics For Decision-Making

Across the United States (as of 2008), there are approximately 18,000 state and municipal police agencies, employing in excess of 750,000 full-time law enforcement officers (Reaves, 2011). In addition, there are at least 24 federal agencies that employ roughly 120,000 full-time law enforcement officers (Reaves, 2012). At least 80% of those agencies utilize a detailed list of rank-ordered force options (commonly referred to as a "force continuum") as part of their official use-of-force policy (Terrill, Paoline, III, & Ingram, 2011). These formal classifications provide policy-level guidance for how the officers employed at a given government agency are allowed to use force (Terrill et al., 2011).

Rank-ordered classifications of force are also utilized within the social scientific literatures, where police and civilian actions are sometimes separated into distinct lists of role-specific categories. Researchers have used these lists to calculate the proportionality of police use of force relative to civilian resistance by quantifying a mathematical difference score, dubbed a "force factor" (Terrill, Alpert, Dunham, & Smith, 2003; Alpert & Dunham, 2004). An advantage of this method is that it explicitly

quantifies the moral evaluation of police and civilian violence (moral in the sense that it indicates whether force is insufficient, appropriate, or excessive), and there is evidence that such an algorithmic approach can yield reasonable inter-rater reliability between two groups of evaluators (Hickman, Atherley, Lowery, & Alpert, 2015). However, the core of the force factor approach requires a clear and uncontested assignment of actions to categorical severity levels, which is a condition that does not reflect the variety of actual law enforcement classification policies. Moreover, the courts have explicitly ruled that evaluations of specific instances of police use of force are so idiosyncratic that they must be evaluated individually with all relevant facts and circumstances taken into account (*Bell v. Wolfish*, 441 U.S. 520, 1979). Thus, while strictly algorithmic approaches to the moral evaluation of force may be useful for researchers, they are insufficient tools for retrospective courtroom decision making. Moreover, the legal precedent the courts have set requiring legitimate police use of force to be "objectively reasonable" (*Graham v. Connor*, 490 U.S. 3, 1989) is too ambiguous to be useful for officers who must make prospective decisions in real time under time pressure and threat.

Acknowledging that law enforcement agencies and researchers have approached force classification and moral evaluation similarly due to the nature of use-of-force decision-making, it is, nevertheless, unclear whether lay people making evaluations of force are framing force in terms of a sequence of categories (like police) or doing anything resembling a force-factor calculation (like researchers). Even among professionals, there is variation in the precise structure of use-of-force policies. There is no absolute consensus between law enforcement agencies regarding a uniformly accepted list of rank-ordered force options, and policies can differ in terms of the exact placement of some actions (Terrill et al., 2011). For example, conducted energy devices (e.g., tazers) are placed below hard empty hand techniques (e.g., punches or kicks) by some agencies, but the order is reversed by other agencies. Additionally, the number of rank-ordered categories contained within a given use-of-force policy varies between 3 and 9 levels, depending on the law enforcement agency (Terrill et al., 2011). Still, nearly three quarters of all law enforcement agencies use a 5 or 6 level classification (Terrill et al., 2011), and many researchers (Terrill et al., 2003) recommend a six-category list for evaluation of police use of force and civilian resistance. Thus, despite the existence of minor variations, most use-of-force policies exhibit a very similar structure and categorization scheme that is mirrored in research paradigms. Therefore, the current study utilized a six-category list which is depicted in Table 1 and described more completely in the method section. As

noted above, it is common parlance among law enforcement professionals and in the criminal justice literature to refer to these force option lists as a "force continuum," but in actuality they are ordered categories that do not contain information about the distance between levels of force nor about differences within categories. For the purposes of the current research, we refer to this list as the *six ordinal levels of force* (SOLOF) scale.

Law enforcement agencies routinely use a SOLOF-style scale in combination with a heuristic rule to train their officers to quickly identify appropriate levels of responsive force (Garner, Schade, Hepburn, & Buchanan, 1995; Terrill & Paoline, 2013), though it is worth noting that there is considerably less inter-agency consensus about these rules than about the SOLOF-scale itself (Terrill et al., 2011). For example, a policy might dictate that officers are allowed to be one level of force above a resisting civilian in order to maintain control and to minimize the duration of the encounter, as well as to reduce the potential for officer and civilian injury that could result from an extended interaction (e.g., if a civilian punches an officer, an agency might authorize the officer to respond by striking the civilian with a collapsible baton rather than engaging in a protracted punch-for-punch exchange). By memorizing their agency's SOLOF scale and heuristic rule, officers can presumably make consistent choices about the most appropriate level of force to use in real-time encounters with resisting civilians. This strategy is especially useful for professional law enforcement officers who can converse with colleagues about hypothetical situations (a.k.a. "if-then dialogues") and thereby enhance intra-agency consistency.

One problem with the formal SOLOF measure endorsed by Terrill et al. (2003), however, is that it contains ambiguous categories. For example, it is unclear what distinguishes "active resistance" from "aggravated active resistance." Consequently, there is room for disagreement regarding which SOLOF category best accommodates a specific forceful action. It is possible that the widespread conflict surrounding police use of force that is observable in the current national discourse is rooted in this sort of disagreement about action binning, and it is also possible that utilizing a SOLOF list and heuristic rule are largely unrepresentative of what lay people are actually doing when they make moral judgments about use of force. Because there is so little systematic research investigating lay evaluations of specific forceful actions, the extent to which civilians utilize something like a SOLOF-style list and heuristic rule is currently unknown.

Recent Criticism and Disparate Evaluations of Force

Because of the delegation of legitimate use of violence to law enforcement agencies, police use of force in the United States is perennially under scrutiny. In recent years, police use of force has been heavily criticized (Friedersdorf, 2015, 2017). Videos of deadly force involving police and minorities have been prominent in the national news (Handelman, 2014), and harsh criticism of police has consequently proliferated (Coates, 2015). Some of this criticism is clearly warranted and is valuable as a catalyst for reforming unjust or insufficiently skilled practices; law enforcement agents are rightly held to a high standard of excellence, especially when employing violence (Selby, 2016). However, some argue that a disproportionate amount of police violence in the media has resulted in public perceptions that the use of force by police is a common occurrence (Sullivan, 2017) when it is actually quite rare relative to the total number of police-citizen encounters (Hyland, Langton, & Davis, 2015; Lemoine, 2017). Moreover, a careful examination of the rhetoric (both anti-law enforcement and pro-law enforcement) in dialogues about police use of force reveals a regular disconnection from the facts of particular cases, and an unwillingness to wade into the full nuance of the details (Loury, 2015; McWhorter, 2016). Because police use of force is such a highly moralized topic, it is especially vulnerable to ideological distortions, extreme levels of conviction, and confirmation biases that influence the way it is evaluated. In general, moralized attitudes have unique characteristics that include increased intolerance of others with dissimilar views, reduced cooperation, and an inhibited ability to resolve disagreements (Skitka, Bauman, & Sargis, 2005). So, to the extent that moral polarization surrounding police and civilian violence is increasing, third-party observers on opposite sides are prone to oversimplify in favor of their prior moral mandates despite the complexity and nuanced details that are frequently present at the level of individual use-of-force cases.

There are many ways to categorize instances of police use of force. For example, observational and experimental researchers have routinely used dichotomous (e.g., "reasonable" vs. "excessive") or trichotomous (e.g., "no force," "necessary force," "improper force") categories as measurement tools (Worden, 2015; Terrill & Mastrofski, 2002). In other domains, it has been argued that police use of force can be rationally divided into categories that include (1) excessive: cases of egregious, unlawful abuse of authority, (2) negligent: cases of unintentional/non-malicious officer overreaction, often coupled with a lack of sufficient training, and (3) necessary: cases in which law enforcement actions yield tragic outcomes but are, nevertheless, appropriate (Harris & Loury, 2016). Notice that each of these various categorization

strategies decomposes use of force in moral terms (i.e., in terms of perceived right and wrong) rather than in more descriptive/objective terms like raw physical severity, potential to cause injury, etc. In other words, although at face value it seems like questions such as "how many times did police in the U.S. use excessive force in the past year?" are fact-based questions demanding a precise numeric answer, in reality they are value-driven questions that implicitly require binning complex sets of facts into sharply divided and mutually exclusive moral categories (Balko, 2016). This sort of categorization process presupposes a particular set of shared values and experiences (Graham, Haidt, & Nosek, 2009; Haidt, 2012; Friesdorf, Conway, & Gawronski, 2015). In ambiguous cases, whether or not an individual judges an officer to have acted, for example, "too quickly" or "at the appropriate moment," likely depends, in part, on many individual differences including the evaluator's political leaning, gender, personal and vicarious experiences with officers, race, ethnicity, etc. As a result, for many of the cases of police violence recently highlighted by the media, there is widespread social disagreement about whether the details of that particular event belong in the "necessary" or "excessive" bin; people watching the same video or reading the same set of facts seem to be coming to starkly different moral conclusions (Kahan, Hoffman, & Braman, 2009; Pew, 2014).

Parsing the extent to which these disparate evaluations are driven by individual differences or by the actions themselves is a challenging but important goal. To the extent that a society values improving trust between citizens and police, and increasing police efficacy and perceived legitimacy (goals that are worth pursuing, Tyler, 2004; Tyler, Goff, & MacCoun, 2015), it is crucial to develop a clear and detailed empirical understanding of how third-party observers make moral evaluations of use of force, and the ways in which their perceptions align or differ from one another. Despite this pressing need, experimental research into third-party evaluations of use of force is limited in both quantity and scope.

Previous Research on Moral Evaluations of Use Of Force

A variety of previous research has investigated different aspects of police use of force. For example, one body of work has focused on describing the actual rates and correlates of use of force that exist in U.S. policing (Klinger, 1995; Fryer Jr., 2016). These studies analyze real-world data sets (i.e., reports of actual use-of-force incidents) that are usually sampled from one or a small number of police agencies, and they often model relationships between a selection of independent variables (e.g., policing style, agency culture, officer/suspect race, etc.) and rates of "excessive" force. Other research has investigated general public

attitudes toward police use of force using a variety of large-scale methods. Some studies make inquiries about first-person experiences (e.g., the Police-Public Contact Survey, Eith & Durose, 2011), while others rely on a small subset of questions extracted from routinely collected, non-specific datasets such as the general social survey (D. Johnson & Kuhns, 2009; Silver & Pickett, 2015). While all of these studies are scientifically valuable, they do not speak directly to how observers make moral evaluations of use of force, or how observers perceive the severity of specific forceful actions.

In another corpus of research using experimental paradigms, participants have been presented with forceful actions that are divided, *a priori*, into justified and unjustified categories (Bradford, Milani, & Jackson, 2016; Gerber & Jackson, 2017). These studies tend to model associations between ratings of acceptability and individual-difference measures. One variant of these studies restricts the focus to evaluations of deadly force exclusively (Fridell & Binder, 1992; Skinner & Haas, 2016). Such an approach is useful for understanding deadly force but is unlikely to generalize to all levels of force, in part because deadly force is extreme and rare (as rare as a person being struck by lightning, Eith & Durose, 2011; Lemoine, 2017)). Other experimental research in psychology has used time-restricted decision-making tasks (e.g., the "shooter task"), but these studies have almost exclusively focused on biases by police and civilians in the application of force (Correll et al., 2007; Hall, Hall, Perry, & Hall, 2016) rather than on moral evaluations of force from the perspective of lay observers.

All of these previous approaches were driven by top-down strategies that assumed stimuli reflect a moral consensus about the reasonableness or excessiveness of a particular forceful action, for example, that if a civilian is armed then lethal force is acceptable but if the civilian is unarmed then lethal force is excessive (Nix, Campbell, Byers, & Alpert, 2017). However, this method of dividing up force evaluations is likely to be misaligned with the nuance and complexities of real assessment (Klinger & Slocum, 2017). As far as we are aware, none of the studies in the relevant literature has systematically measured perceptions of representative actions across the entire spectrum of possibilities. Because official classifications of forceful actions have a rank-ordered structure, and heuristic rules place legitimate actions relative to other actions within that structure, it is important to assess how lay perceptions of specific actions across the entire range of force severity do or do not align with law enforcement classifications.

Thus, our primary focus for the current research was estimating the underlying moral and physical severity of a wide variety of representative forceful actions in order to generate an emblematic set of

psychometric scale values. These scales enable subsequent comparisons between lay evaluations and the structure of SOLOF models. They also provide a rigorous foundation that makes it possible for future research to ask questions about the general level of severity lay evaluators expect police to use, and, thus, they generate a framework for evaluating common heuristic rules used by police and comparing them to lay expectations.

The Need For Scale Values For Representative Forceful Actions

To measure whether lay force evaluations resemble a SOLOF-style scale, and whether lay evaluations agree with a heuristic such as "officers can use force that is one level above civilian resistance," it is important to know the perceived severity of a wide range of representative forceful actions. Such baseline knowledge is required because these heuristics are, by nature, *relative* references to other possible actions. To know that people generally expect officer force to be above civilian force (or below or equivalent to), one must know where a wide range of actions are located in terms of their perceived severity. It cannot merely be presumed that actions officially categorized at a given SOLOF level are, in fact, perceived by lay evaluators the same way. Thus, if we want to know whether the underlying general tendency of lay evaluators judging a specific action sequence matches or diverges from a particular law enforcement heuristic, we must first have detailed knowledge about the precise perceived severity of the actions in question and a sufficient number of surrounding actions with which to make relative comparisons.

The underlying severity inherent in any given action has at least two components: a raw physical component that reflects descriptive severity, and a moral component that is overtly prescriptive. These components roughly correspond to David Hume's famous is-ought distinction (Hume, 1738/2003) in that the descriptive component reflects how much force there is, while the prescriptive component reflects how much force there ought to be. For this reason, in the current study, we estimated psychometric scale values for a wide range of forceful actions using both descriptive and prescriptive dimensions.

As explained in detail below, we presented participants with minimal vignettes describing an officer and civilian encounter that involved one officer action and one civilian action. The actions in these paired vignettes spanned the complete range of force options. Participants rated the officer and civilian actions on moral dimensions (i.e., acceptability, appropriateness, and punishability of the action) and a physical dimension (i.e., how much force). These ratings were used to derive underlying moral and physical scale

values for each action.

Method

All of the experiments presented in this article were approved by the Indiana University Institutional Review Board in Bloomington, Indiana.

Participants

To obtain a diverse sample, we collected data from both the Indiana University (IU) undergraduate subject pool and Amazon Mechanical Turk (MTurk). To verify that the two different subject pools did not differ in their ratings, we gathered reasonably large samples from IU and MTurk and conducted the analysis described below separately for each sample. Results showed virtually identical model descriptions, and therefore we combined the samples together in order to increase the precision of the estimates. Separate analyses for IU and MTurk subject pools are available at <https://osf.io/h6jg3>.

A total of 411 participants were recruited through MTurk and 395 participants were recruited through the IU undergraduate subject pool. Each participant was presented with a consent form explaining the procedure for the study and the participation requirements, which included being a current resident of the United States, at least 18 years old, and a fluent speaker of the English language. Thirty eight participants were excluded from the MTurk sample for failing to participate from within the US (as verified by IP address geolocation), and 23 were additionally excluded because they failed more than one of the random manipulation checks or because the mean absolute deviation (MAD) of their ratings for the entire experiment exceeded 3 standard deviations below the MAD for all MTurk participants (i.e., their responses hardly varied across questions or items, suggesting they may have left the response slider at its default level for most trials). Participation within the US was guaranteed with the IU sample, but 78 participants were excluded for failure of more than one manipulation check or if the MAD exceeded 3 standard deviations below the MAD for all IU participants. This yielded a total of 350 MTurk participants and 317 IU participants, for a combined total of 667 participants.

Of the included 350 MTurk participants, 196 (56%) identified as female, 152 (43.4%) identified as male, and 2 (0.6%) identified as other. Two (0.6%) identified as American Indian/Alaska Native, 23 (6.6%) identified as Asian, 27 (7.7%) identified as Black or African American, 20 (5.7%) identified as Hispanic or Latino, 269 (76.9%) identified as White, 8 (2.3%) identified as More than one, and 1 (0.3%) identified as Unknown.

Of the included 317 IU participants, 208 (65.6%) identified as female, 109 (34.4%) identified as male, and 0 (0%) identified as other. One (0.3%) identified as American Indian/Alaska Native, 48 (15.1%) identified as Asian, 27 (8.5%) identified as Black or African American, 12 (3.8%) identified as Hispanic or Latino, 217 (68.5%) identified as White, 11 (3.5%) identified as More than one, and 1 (0.3%) identified as Unknown.

Stimuli

Set of actions. We began by crafting a role-neutral SOLOF scale that combined Terrill et al. (2003)'s *suspect resistance levels* and *officer force levels* (see Figure 1). Using this unified SOLOF scale, we generated a large list of actions (approximately 100) that could plausibly, or at least possibly, be committed by both an officer and a civilian, being sure to include multiple actions that rationally fit into each of the six SOLOF levels. It is worth noting that this initial list was generated loosely in the sense that it functioned as an initial brainstorm of options and as a mechanism to automatically exclude any obviously implausible actions. Importantly, the initial list did not attempt to generate a set of actions that are equally plausible for both an officer and a civilian because generating such a list is likely to be impossible; in the context of a police-civilian interaction, some actions are intrinsically more associated with police (e.g., handcuffing), and the same is true for other actions that are probably more associated with civilians (e.g., calling the other person a "pig" or pulling away).

After composing this larger list, we set out to narrow it down to a more manageable set of 22 actions per role while retaining at least three actions in each of the six SOLOF categories. The first thing we observed was that, although we were able to generate a large number of actions, there seemed to be a reasonably small core set of violent actions that can plausibly occur within a dyadic pair. Beyond the core set, additional actions are most readily generated by specifying severity modifiers such as the target of the action or the number of times an action occurred. For example, an officer could kick a civilian in the face, stomach, leg, etc., and could do so once or more than once. We immediately eliminated actions that occurred more than once in order to avoid the confound of comparing multiple actions (e.g., punching repeatedly) with one action. To deal with action target, we initially tried to specify actions in their most generic form (e.g., *the civilian punched the officer*), but realized that target clearly matters for judgments of action severity (e.g., kicking someone in the head versus kicking them in the hip), and that leaving the target ambiguous would likely increase participant confusion and decrease measurement precision (e.g.,

some participants might infer a default target if left unspecified). For that reason, we chose to specify action target (e.g., *the civilian punched the officer in the face*) in all cases except those in which the target was obvious (e.g., handcuffing) or seemed largely irrelevant. This target specification is congruent with law enforcement training procedures that incorporate target into judgments of force severity for intermediate severity levels (Jenkinson, Neeson, & Bleetman, 2006).

To further hone the set of experimental actions, we considered selecting officer actions according to their actual frequency of use, but there are several problems with that approach. First, police in the United States only use force in about 1-2% of all police-citizen encounters, so it is a relatively rare occurrence overall. Secondly, the distribution of force severity among those 1-2% of encounters is heavily skewed because most police use of force is comprised of low severity actions (Hyland et al., 2015; Davis, Whyde, & Langton, 2018). As a result, selecting actions according to how commonly they are used by police in the field does not yield a set of actions that include representative options across the full spectrum of severity; instead, such an approach would produce many actions at low levels of severity and few or no actions at high levels of severity. Because we are attempting to map actions across the entire range of severity, selecting actions based on their raw or proportional frequency of use was not appropriate.

With all of these considerations in mind, we ultimately selected actions that were broadly distributed across all severity levels, that we believed were widely representative of real-life force options, and that also reflected within-category variety and maximal role-plausibility. Specifically, we first chose officer actions that are in common use among law enforcement agencies and frequently appear on the majority of use-of-force policies and in police training materials in the U.S. (Terrill et al., 2011). For example, it is standard practice at most police agencies to issue collapsible metal batons to officers for use as impact weapons (although some agencies distribute other impact weapons such as nunchucks or PR-24 rigid batons, collapsible batons are issued by the majority of modern law enforcement agencies because of their compact size that facilitates ease of carry without significantly reducing efficacy). It is worth noting that, prior to conducting this research, the first author (BDC) spent more than a decade as a commissioned law enforcement officer at municipal and federal levels. During a majority of that time, he functioned as a use-of-force instructor for a joint set of law enforcement agencies, as a subject matter expert on police use of force, and as a member of internal review boards evaluating the legality of specific instances of force application in the field. Thus, our action selection process was also informed by his prior experience and

expertise.

Seventeen of the officer actions we chose could also be plausibly performed by a civilian. However, five of the 22 selected actions differed between the officer and the civilian. For example, one of the officer actions at level 5 of the SOLOF indicates that "the officer shot the civilian in the chest with a beanbag shotgun round." Because it is implausible that any civilian would have access to such a specialized weapon, we matched this officer action with "the civilian hit the officer in the chest with a baseball bat," which makes the action more plausible for a civilian but is reasonably balanced with the general level of force caused by a beanbag shotgun round. We retained these five actions because we wanted to maximize diversity of action within each SOLOF severity level. For a full list of all actions, including the five actions that differ between officer and civilian, please see Figure 1.

Because there were five unique civilian actions matched with five other officer actions, the total number of *unique* actions was 27. Nevertheless, to avoid confusion and to reinforce that these five actions were matched across roles, we refer to the full set of "22 actions" throughout this article. Importantly, although we tried to match the five differing actions a priori in terms of severity, it is not crucial to the experimental design, statistical analyses, or results that these actions ultimately turn out to be well-matched between officer and civilian because each of the 27 actions has its own moral and physical severity estimates that are relative to all of the other actions. As we explain later, results show that the estimates for these five differing actions are remarkably similar.

Vignette. A minimal vignette was constructed to state each action pair without introducing superfluous contextual details:

While working a patrol shift, a police officer was dispatched to a reported crime, and in the course of the investigation, he came into contact with a civilian. During their subsequent interaction, the following actions took place *in the order listed below* (other actions may have occurred before, between, or after):

Two specific actions were then listed, in bold font. There was one action by the civilian and one action by the officer. Figure 2 shows an example of the display seen by participants.

Measures. For each vignette with its two actions, participants were asked to make four ratings of the second action, one rating per screen. Each rating was made using a slider with anchors marked as indicated in parentheses:

- Please rate how much **force** is involved in the officer's/civilian's action: (0 = No Force, 100 = Maximum Force)
- Please rate how **morally acceptable** the officer's/civilian's action is: (0 = Completely Unacceptable, 100 = Completely Acceptable)
- Please rate the extent to which the officer/civilian used an **appropriate** level of force (0 = Insufficient, 50 = Appropriate, 100 = Excessive)
- Please rate the extent to which the officer/civilian should be **punished** for his action: (0 = Not at all, 100 = Severely)

The rating of force was designed to assess evaluations of the severity of the raw physical force of the actions. The other three rating questions were intended to assess moral evaluations. Thus, these dependent measures allowed for a comparison between the physical and moral dimensions of force severity. Only the appropriateness slider had its center point labeled, because we wanted to provide participants with an opportunity to rate officer and civilian actions as insufficient as well as excessive. It is worth noting that we consciously opted to probe physical severity with only one question in order to keep the total number of questions brief enough that the experiment duration remained tolerable for participants. Intuitively, physical severity is a less complex concept than moral severity, and we collected enough data to make precise estimates of scale values. Figure 2 shows an example of the display with the response slider. Although the structure of SOLOF-style policies is generally ordinal, we purposely used a continuous measurement scale in order to maximize precision and to allow participants to respond in an unconstrained way. Importantly, our continuous response scale does allow for actions to cluster into categories (e.g., six categories) if they are perceived that way, but it is also flexible if the underlying perceived severity does not map well into discrete categories.

Procedure and Design

After informed consent, participants read a set of instructions explaining that they would be presented with a scenario that was representative of a real life interaction involving a police officer and a civilian, accompanied by a pair of actions that occurred during the encounter: one action by the police officer, and one by the civilian. Participants were also told that the order in which any given pair of actions

had occurred was the order in which they were presented, and that other actions may have occurred before, between, or after the two listed actions. This latter information was included in order to allow for some ambiguity in the scenario as well as to increase the coherence of actions pairs that might otherwise be perceived as nonsensical. For example, if the civilian strikes the officer with a motor vehicle and subsequently the officer places handcuffs on the civilian, there must have been other things that occurred between the actions. Participants were informed that they would be asked to make ratings of force, as well as appropriateness, moral acceptability, and punishment, for the second action that occurred during the encounter (i.e., always the action that occurred subsequent to the first action). Participants were also advised that there were no right or wrong answers and that we only wanted their intuitive, thoughtful responses. Following the instructions, participants provided ratings of general attitudes toward police (J. A. Johnson, Hogan, & Zonderman, 1981). Complete details are provided at <https://osf.io/h6jg3>.

Participants then viewed a series of 22 vignettes, each with a pair of actions. Officer-civilian action pairings were manipulated within subjects. Because the 22 actions could be performed by both the officer and the civilian in either order, the total number of possible action pairs was 968. Rather than present participants with all 968 action pairs, which would have caused fatigue and inattentiveness, we instead permuted the officer and civilian actions separately and then paired them together to form uniquely randomized action pairs for each participant. By using this method, we were able to present participants with only 22 action pairs while still ensuring that every participant saw the civilian and officer performing all 22 actions without any action repetition within a given role. Once the 22 action pairs had been selected, 11 random pairs were assigned to the civilian evaluation condition (i.e., the officer action was listed first followed by the civilian action, and participants were asked to rate the civilian action), and the remaining 11 pairs were assigned to the officer evaluation condition (i.e., the civilian action was listed first followed by the officer action, and participants were asked to rate the officer action). Finally, all 22 action pairs were randomly ordered so that ratings of civilian and officer actions would be interspersed as the experiment progressed. Participants were not informed about the set of possible actions or the constraints on their pairings. For each vignette, the four measures (force, moral acceptability, appropriateness, punishment) were presented in a random order, one at a time.

Eleven attentiveness checks were also inserted randomly for some of the action pairs. A random number of these catch questions instructed participants to move the slider all the way to the left, and the

remainder all the way to the right. These catch trials were crucial for identifying participants who were not providing thoughtful answers or not following instructions, a concern that is especially important when using crowd-sourced participant pools (Paolacci & Chandler, 2014).

Following the scenario ratings, participants answered a series of questions designed to assess their beliefs about police legitimacy (Gerber & Jackson, 2017), as well as questions measuring the number of personal and vicarious negative experiences with police (Weitzer, 2005). The focus of the current analysis is on developing precise psychometric estimates of the actions, so these individual-difference measures will not be discussed at length here. Further discussion of the individual difference measures can be found in Celestin and Kruschke (2019). The experiment concluded with a series of demographic questions, including a single item measure of political affiliation (Iyer, Koleva, Graham, Ditto, & Haidt, 2012), and a debriefing. Complete details are available at <https://osf.io/h6jg3>.

Software

The experimental sequence was programmed using the JavaScript, html, and CSS programming languages, with the open-source libraries available in the jsPsych package (de Leeuw, 2015). The experiment was made available to participants online using an open-source psiTurk (Gureckis et al., 2015) or Apache server running on a desktop computer in combination with a SQL database or server-side PHP scripts. In addition, data cleaning, pre-processing, and analysis were done using the R programming language, RStudio (Team, 2016), JAGS (Plummer, 2003), runjags (Denwood, 2016), and Stan (Stan Development Team, 2016) software, as well as variants of the Bayesian analysis programs provided by Kruschke (2015).

Results

Data Analysis

In our analyses, the key dependent variables are the moral ratings of acceptability, punishment, excessiveness, and the rating of physical force. Upon visual inspection of the raw data, we observed that the distributions of responses on the 1-100 sliding scale formed a (sometimes skewed) unimodal hill in the middle of the scale with spikes on one or both extreme ends of the scale. To model the end spikes, we treated responses less than 6 or greater than 94 as censored, meaning that they indicated some uncertain value at least that extreme. To make the distributions of uncensored values nearly normal, we applied a

logit transformation to all responses (after dividing by 100). Complete details are in the R code available at <https://osf.io/h6jg3>.

We modeled the (censored and logit transformed) ratings using linear regression in which each trial's rating was described as an additive influence of the latent scale values of civilian and officer actions. In other words, the ratings of the actions are a weighted combination of the underlying scale values of the actions. As an intuitive example, consider rating the acceptability of an officer action. As the severity of the officer action increases, the action's acceptability should (intuitively) decrease, implying that the regression coefficient on officer-action severity would have negative sign. But as the preceding civilian action increases in severity, the acceptability of the officer action should increase, implying that the regression coefficient on civilian-action severity would have positive sign. The analysis estimates the regression coefficients and simultaneously estimates scale values for the actions.

Formally, the regression model states that the predicted rating, $\hat{y}$, is a weighted combination of the civilian action severity, γ , and the officer action severity, χ , as follows:

$$\hat{y}_{q,c,f} = \underbrace{\beta_{0,q}}_{\text{baseline}} + \underbrace{\beta_{1,q}}_{\text{influence of civilian action}} \cdot \underbrace{\gamma_c}_{\text{civilian action severity}} + \underbrace{\beta_{2,q}}_{\text{influence of officer action}} \cdot \underbrace{\chi_f}_{\text{officer action severity}} \quad (1)$$

where $\hat{y}$ represents the predicted value of a rating (logit transformed), q indexes the rating question (1–8, i.e., the four types of ratings of both officer and civilian actions), c indexes the civilian action (1–22), f indexes the officer action (1–22), γ denotes the estimated civilian-action scale value, and χ denotes the estimated officer-action scale value. Each of the β variables is an estimated regression coefficient. The regression coefficients and action scale values were estimated simultaneously. We analyzed the three social/moral questions (i.e., moral acceptability, appropriateness, and punishment) separately from the force question, and derived separate latent scale values for perceived moral and physical severities of the actions. Neither the moral nor the force questions were aggregated; rather, they were included in their respective analyses simultaneously. For example, the moral severity scale estimates are informed by all six social/moral questions (three officer, three civilian).

The estimation of scale values is analogous to structural equation modeling in which observed ratings covary with an underlying latent factor. Because of algebraic indeterminacies inherent in linear

regression, the latent scale must be pinned down at two arbitrary points. We chose to fix the "waved" and "handgunHead" actions (see Figure 1 for full action descriptions) at 1.0 and 6.0, respectively, because they were likely to be the most extreme. The model did not constrain the other action estimates which were allowed to exceed the bounds of the two pinned action values if warranted.

We also analyzed a variety of other regression models that involved individual-difference variables (e.g., beliefs about police legitimacy, negative police experience, political affiliation, participant race, participant gender, etc.) and various interactions of predictors. These alternative analyses can be found at <https://osf.io/h6jg3>. The resulting estimated scale values of actions were remarkably consistent across all of the various models. Because the focus of this article is the scale values of the actions, we report results from the simplest model that yields robustly stable scale estimates in Equation 1. A more elaborate model is reported in a companion manuscript (Celestin & Kruschke, 2019) along with extensive discussion of the regression coefficients and trends in the ratings. Notably, although this companion manuscript and the results we report here are based on the same data, the analyses are mathematically different, focus on unique aspects of the models, and have substantively distinct implications. The current results focus on the latent severity scales implied by the data and how those scales relate to the SOLOF categories used by law enforcement. Whereas, in Celestin and Kruschke (2019), the focus is on the ratings of appropriateness, punishability, etc., and the relation of those ratings to the predictors. For our purposes here, it is of primary importance to note that, despite the use of more elaborate models, the latent severity scales remain essentially the same, and so the results we report currently are robust to the inclusion of additional relevant predictors.

Bayesian method. We used Bayesian estimation of the parameters (Kruschke, 2015; Kruschke & Liddell, 2018a, 2018b). Bayesian methods are particularly useful in this application because they yield a complete posterior probability distribution over all of the parameter values in our models, including the regression coefficients and the scale values of the actions. We summarize the posterior distribution on a parameter by its mode (i.e., most probable value) and its 95% *highest density interval* (HDI), which spans the 95% most probable parameter values. A narrower HDI indicates a more precisely estimated value.

Bayesian analysis requires specification of a prior probability distribution on the parameters. We used a very broad, vague prior relative to the scale of the data. Consequently, the prior distribution had negligible influence on the results.

Posterior distributions were computed using Markov chain Monte Carlo (MCMC) methods, in the R programming language, along with JAGS (Plummer, 2003) and runjags (Denwood, 2016) software, using variants of the Bayesian analysis programs provided by Kruschke (2015). All MCMC chains converged well, and the effective sample size for all parameter chains approaches 10,000.

Comparing Latent Scale Estimates and SOLOF Categories

The latent scale values derived from the ratings of *physical force* are plotted in Figure 3 along the horizontal axis, while the vertical axis represents the SOLOF levels. Detailed numerical values of the latent scale values are shown in Table 2. From Figure 3 it can be seen that the latent severity of physical force tends to increase with the SOLOF levels, but there are also clear discrepancies between the latent severity of physical force and the SOLOF levels. Importantly, actions in SOLOF levels '3', '4', and '5' have latent force severities that overlap substantially, indicating that participants are not strongly distinguishing these actions from one another in terms of physical force. This lack of differentiation among the SOLOF levels '3', '4', and '5' is evident for both civilian and officer actions.

The latent scale values derived from the ratings of *moral* attributes are plotted in Figure 4. Detailed numerical values of the latent scale values are shown in Table 2. From Figure 4 it can be seen that the latent moral severity tends to increase with the SOLOF levels, but there are also clear discrepancies between the latent moral severity and the SOLOF levels. Importantly, for civilians, actions in SOLOF levels '3', '4', and '5' have latent moral severities that overlap substantially, indicating that participants are not distinguishing these actions from one another in terms of moral severity. For officers, SOLOF levels '4' and '5' overlap, with most actions of level '5' actually falling below the moral severity of actions in level '4'. It is notable that the perceived severities of SOLOF level '4' and '5' actions completely overlap. This suggests that once non-lethal actions reach a certain severity threshold, they are not morally very distinguishable from one another for officers or civilians.

Another prominent feature of the scale values is that one action that is in SOLOF level '3' has scale values among the SOLOF level '2' actions. The action in question is handcuffing for officers and its matched action, pulling away, for civilians (see Figure 1 for full action descriptions). The fact that handcuffing by a police officer and pulling away by a civilian are scaled similarly provides some confirmation that those actions were indeed well-matched as comparably severe actions. Handcuffing and pulling away are physical actions, not verbal actions, and therefore would be situated in SOLOF level '3'

(see Table 2). Nevertheless, ratings of physical forcefulness and moral appropriateness place those actions less severely, in the range of the verbal actions of SOLOF level '2'.

Comparing Physical and Moral Scale Estimates

The moral scale values are plotted against the force scale values in Figures 5 and 6. Figure 5 shows the estimates for the officer's actions, and Figure 6 for the civilian's actions. Recall that the least severe action was arbitrarily set at a scale value of 1.0 and the most severe action was arbitrarily set at a scale value of 6.0, and that all other actions were estimated relative to these two fixed points, so it is the relative not the absolute relationships that are informative in these figures. In particular, the diagonal line is displayed merely as a visual guide to proportionality and ordinality. The diagonal line should not be treated as a line of equality across scales because the scale values could be shifted or stretched on either axis. The segments intersecting the plotted points in Figure 5 and Figure 6 represent 95% HDIs. As indicated by the relatively short length of the HDI segments, all of the scale values were estimated with considerable precision on both force and moral dimensions.

One notable pattern in Figures 5 and 6 is non-monotonicities in moral severity as force severity increases. If moral severity were proportional to force severity, then an increase in force severity would always produce an increase in moral severity. But this does not always occur. For example, in Figure 5 the moral severity of handCuff is *less* than the moral severity of nameCall, even though the force severity of handCuff is *greater* than the force severity of nameCall (see Figure 1 for full action descriptions). These non-monotonicities reveal actions that are either perceived as more morally than physically severe or vice-versa.

The lack of differentiation for non-lethal categories observed in Figures 3 and 4 is also prominent in the relatively tight clustering of all three physical but non-lethal actions in Figures 5 and 6. In contrast, the lethal and verbal actions (SOLOF levels 6 and 2, respectively) are relatively spread out – especially the lethal actions. That is, the severity of verbal and lethal actions exhibits considerable variability relative to non-lethal actions.

Because of the potential concern associated with implausible disparate action pairs to skew the results, we applied the model to a subset of data that excluded any pairs of SOLOF level 1 with SOLOF level 6, which comprised about 4.9% of all trials. We observed that responses to these extreme pairings were consistent and reasonable. Additionally, we re-ran our analyses on a subset of data that excluded

these extreme pairs and observed that the latent severity estimates were virtually unchanged. As mentioned in the Method, we also ran a variety of analyses including all of the individual difference measures we collected, and the latent severity estimates are virtually identical to those reported currently. That is, the latent severity estimates we report are robust, and the patterns we describe in this article do not change when extreme action pairings are excluded or when additional relevant individual difference predictors are included in our models.

Discussion

Our results provide a detailed initial map of the latent moral and physical severity of a set of realistic actions that represent the complete range of force options. Because these actions are broadly representative and include many of the most common forceful actions actually employed by and against police, they yield a comprehensive view of action severity perceptions, and offer insight into discrepancies between lay and formal evaluations of force. Moreover, 17 of our 22 actions were identical for the officer and civilian, and the remaining 5 were matched a priori based on our intuitions which were confirmed empirically by the remarkably similar severity estimates when comparing the matched actions in Figures 3 and 4 and Figures 5 and 6. Finally, the combination of a large number of participants and Bayesian methods provided precisely estimated scale values.

This new set of scale values reveals that lay perceptions of police and civilian use of force are substantially disconnected from the way force is traditionally categorized in most law enforcement agencies (i.e., using a SOLOF-style classification). Psychologically, non-lethal physical (beyond verbal) actions cluster together and effectively merge half of the canonical SOLOF categories (i.e., all three levels of non-lethal physical actions). But within the categories of verbal and lethal force, lay participants distinguish much more variable action severities.

Implications

Of the discrepancies revealed in Figures 3 to 6, handcuffing stands out as a unique example. While it is clearly on the physical side of the verbal-physical boundary, it is also a routine police procedure in nearly all arrests, regardless of whether a civilian resists, and therefore might be perceived as especially normal and acceptable relative to other physical actions. Likewise, pulling away (the matched civilian action for handcuffing) may be viewed as maintaining personal autonomy and as more defensive than offensive relative to other physical actions. Moreover, while some handcuffing and pulling away may be highly

forceful, it is possible to handcuff or pull away in a relatively gentle manner, which is something that does not seem to be true for any of the other physical actions.

More broadly, we believe the general pattern of non-monotonicities can largely be explained in terms of the intuitive normativities of the actions for police relative to civilian roles. For police officers, intuitively normative actions involve official tools such as handcuffs, pepper spray, and tazers, while intuitively *non*-normative actions (i.e., actions that are not primarily or uniquely associated with police officers) include name calling and hitting someone with a motor vehicle. In Figure 5 it can be seen that the moral severity of normative actions tends to fall well below the diagonal reference line, while the moral severity of non-normative actions tends to fall above the diagonal reference line.

Thus, for police officers, actions that are not normative (e.g., striking someone with a motor vehicle) may be evaluated as less acceptable than other actions that are more strongly associated with police (e.g., shooting someone in the chest), even if both of those actions are categorized as lethal force according to the SOLOF system. This distinction by lay people is particularly notable because it diverges from the way that courts evaluate actions and the way that police officers train to identify when lethal force is appropriate under realistic time and resource constraints (see the discussion of lethal force below for more on this point).

There are relatively few non-monotonicities for the civilian actions in Figure 6, but the few that occur may also be explained by differential normativity. In some sense, any action by a civilian more severe than cooperation is not prescriptively normative, and might be expected to exhibit greater moral than physical severity as a result. A large majority of the actions in Figure 6 do show this pattern to a slight degree. However, we are primarily concerned here with moral severity (of civilian actions) that is clearly disproportional to physical severity (of those same civilian actions) in response to forceful actions by police officers, and so the actions that stand out most clearly are those in the verbal and proximity categories. Specifically, for civilians it appears to be especially non-normative to call a police officer names or to threaten a police officer, and these actions have moral severities that fall well above the diagonal reference line. In other words, verbally confrontational actions seem to be especially non-normative for civilians in the context of a police-civilian interaction, while more severe actions are more diffuse and less strongly associated specifically with the civilian role.

Normativity in Lethal Force

In general, police rely on legal precedent that treats police use of force as a seizure under the 4th amendment, and only affirms lethal force as lawful if a "citizen poses a threat of serious physical harm, either to the officer or others" (*Tennessee v. Garner*, 471 U.S. 1, 1985); thus, lawful lethal force must be "objectively reasonable" (*Graham v. Connor*, 490 U.S. 3, 1989). Many law enforcement agencies use this criterion to establish a general threshold, and when that threshold is exceeded, any form of lethal force is permitted in order to accomplish the goal of stopping the threat, regardless of how role-normative it is. Thus, lethal force actions are formally evaluated in a binary manner as either justified or not. In contrast, the psychological separation of different lethal force actions observed in Figure 5 suggests that lay perceptions of moral severity are not categorical, and that lay evaluators have a more nuanced view of lethal actions than formal evaluative criteria.

This disconnect between professional and lay evaluations of non-normative lethal force application was especially evident in the case of a sniper attack in Dallas, Texas on July 7, 2016. During this attack, Micah Xavier Johnson killed five police officers, injured nine other officers and two civilians, and then barricaded himself inside of an El Centro College building where he engaged surrounding police with gunfire (Wikipedia Contributors, 2017). After negotiations failed, Dallas police maneuvered a robot carrying an explosive device near Johnson and killed him. This event was the first time a law enforcement agency in the U.S. had used a robot to deploy lethal force, making it highly non-normative. Consistent with the pattern observed in our results, considerable moral outrage was subsequently voiced by the public, not only about the use of lethal force but about the specific and unusual way it was employed (Fountain & Schmidt, 2016; King, 2016).

Our theory, that normative actions are perceived as less morally severe than their physical severity would imply, is post hoc and we will investigate it further in the future. Meanwhile, we note that it is consistent with other recent research that suggests people think of frequently observed actions as morally acceptable actions. Lindström, Jangard, Selbing, and Olsson (2018) showed that people more strongly endorse free riding on a public good when free riding is common. Those authors suggested that people use a general judgment heuristic: *common is moral*. In work using multiple cognitive tasks, Eriksson, Strimling, and Coultas (2014, p. 60) showed that people have strong associations between what is common and what is moral, and cited earlier work showing that "moral judgments of socially undesirable behavior

tend to be less harsh when the behavior is perceived to be common (McGraw, 1985; Trafimow, Reeder, & Blising, 2001; Welch et al., 2005)". Thus, we believe it is very plausible that actions perceived to be normative and common for police officers should also be judged to be less morally severe than those actions' physical force would suggest. Importantly, our conceptualization of normativity only includes actions that are common and also not routinely punished as we discuss below in the limitations.

Dealing with the Disconnect

The patterns in our data point to questions that may promote understanding among all parties invested in police-civilian interactions. For example, the difference in moral severity between an officer yelling "get back" at the civilian and pepper spraying the civilian is quite large according to lay evaluation, but the difference is only one level of force in SOLOF-style scales. Does this imply that official policies should make finer distinctions at the boundary between verbal and physical actions? As another example, it appears that the perceived normativity of a police action influences the way that lay people judge the action's moral severity. Should this be treated as a cognitive bias in lay evaluation that should be corrected because what counts as normative may be an accident of history? Or should this influence of normativity be incorporated into the official classification of actions to emphasize normative actions and de-emphasize nonnormative actions? Finally, it is clear that lay people do not strongly distinguish among SOLOF levels '3,' '4,' and '5' (i.e., the non-lethal beyond-verbal actions). Does this failure to distinguish levels reflect a lack of expertise, implying that law enforcement agencies need to educate the public about important gradations in physical force? Or does the lay perception imply that official classifications should collapse or reorganize some levels?

Importantly, the disconnect does not necessarily imply that SOLOF-style scales should be abandoned. Many activists who are currently calling for reformative changes to how the police use force are explicitly recommending the ubiquitous adoption of SOLOF-style rules by all law enforcement agencies in the US (Campaign Zero, 2018). The widespread adoption of SOLOF-style categorizations presumably means they are useful, and careful consideration of the challenges surrounding use-of-force decision-making reveals the need for some sort of specific structure to guide officers in the field. But our results may suggest ways to bring law-enforcement and lay perceptions into closer alignment. For example, educational outreach programs might help lay people understand how police and the legal system classify use of force, especially targeting the discrepancies from lay intuition revealed by the present research. As

another example, knowing that more role-normative actions are viewed by the public as less morally severe than non-normative actions may guide officers when choosing a force option within a particular SOLOF level. Additionally, the public might be educated to be wary of becoming morally de-sensitized (or hyper-sensitized) to the forcefulness of actions merely because of their normalization (or lack thereof).

Limitations and Future Directions

As with any empirical research, there are limitations that apply to our findings. One is that, although we selected our 22 actions in order to maximize representativeness and intra-SOLOF-category diversity across the whole spectrum of police and civilian force options, those we selected are merely a sample of all possible actions. Nevertheless, because of our selection procedure, we fully expect that alternative sets of actions will replicate the general patterns we observe even as there are likely to be idiosyncratic differences for some actions.

Our vignettes only used two actions, but real-life conflict between police and civilians is comprised of a set of many interdependent actions that occur over the time-course of an encounter. Future research could explore how moral evaluations are affected by a sequence of interdependent actions between officers and civilians, perhaps paying special attention to how increasing or decreasing the severity of force over time alters perceptions about officer or civilian intentions to escalate or de-escalate conflict. In addition, our use of vignettes allows for enhanced control over stimuli, but may be more limited than video in terms of realism.

In this article we have focused on the scale estimates of actions without describing the pattern of moral and physical ratings for specific pairs of actions. In particular, the scale values alone do not reveal whether lay expectations generally align with common law enforcement heuristics such as the "one level above" heuristic. These ratings are described in a companion paper by Celestin and Kruschke (2019). Moreover, future studies can use the scale estimates we provide to precisely compare the severity of actions chosen by lay evaluators as appropriate police responses to specific civilian violence.

Currently, it is unclear to what extent the disconnect we observe between formal use-of-force policies and lay intuitions emerges as a result of differential expertise rather than moral disagreement. To investigate this issue, future research should attempt to disentangle the influence of descriptive and prescriptive effects on lay evaluations of force. One approach would be to teach lay evaluators about SOLOF-style force categorization and common law enforcement heuristic rules and then observe whether

and how that affects subsequent judgments. Care would need to be taken to avoid experimenter demand effects, but such an approach could provide useful information for law enforcement agencies regarding how to allocate resources to target educational public outreach efforts or to hone their use-of-force policies to maximize alignment with public morality.

Finally, future research should clarify what conditions are needed for common (i.e., normative) actions to be perceived as moral (or less immoral) actions. We suspect that actions become perceived as less immoral especially when those actions are not routinely punished or disparaged. Actions that are perceived to be frequent but are also routinely punished may still be perceived to be immoral.

Conclusion

Because police use of force is a highly moralized topic, evaluating specific instances of applied force is likely to always involve some contention between police and civilian observers. In fact, reasonable people — both lay citizens and officers — can and do disagree about whether specific instances of force are necessary or excessive, and even whether instances that are identified as excessive are isolated incidents or whether they imply a broad problem and the need for systemic change (Morin, Parker, Stepler, & Mercer, 2017). Regardless of how one makes such inferences, it has become clear that police actions are promptly evaluated in the court of public opinion, and it matters little to lay evaluators whether a particular set of actions falls within the law or within agency protocol if the intuitive public reaction to viewing the episode evokes sufficient moral outrage. It is our hope that the results presented here will provide an empirically-grounded starting point for helping to bring together lay and professional moral evaluations of use of force.

References

- Alpert, G. P., & Dunham, R. G. (2004). *Understanding Police Use of Force: Officers, Suspects, and Reciprocity*. Cambridge University Press. doi: 10.1017/CBO9780511499449
- Balko, R. (2016, jul). *Why it's impossible to calculate the percentage of police shootings that are legitimate* - *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-watch/wp/2016/07/14/why-its-impossible-to-calculate-the-percentage-of-police-shootings-that-are-legitimate/?utm{_}term=.2116c367c05f
- Bell v. Wolfish*, 441 U.S. 520. (1979).
- Bradford, B., Milani, J., & Jackson, J. (2016). Identity, Legitimacy and 'Making Sense' of Police Violence. *Social Science Research Network*, 1–14.
- Campaign Zero. (2018). *Campaign Zero*. Retrieved 2018-02-03, from <https://www.joincampaignzero.org/solutions/>
- Celestin, B. D., & Kruschke, J. K. (2019). *Lay evaluations of police and civilian use of force: Moral and physical magnitude ratings of officer and civilian actions*. Retrieved from <https://osf.io/aqfeh/>
- Coates, T.-N. (2015, apr). The myth of police reform. *The Atlantic*. Retrieved from <http://www.theatlantic.com/politics/archive/2015/04/the-myth-of-police-reform/390057/>
- Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: police officers and racial bias in the decision to shoot. *Journal of personality and social psychology*, 92(6), 1006–1023.
- Davis, E., Whyde, A., & Langton, L. (2018). *Contacts Between Police and The Public, 2015* (Tech. Rep.). Washington, DC: U.S. Department of Justice Bureau of Justict Statistics.
- de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47, 1–12. doi: 10.3758/s13428-014-0458-y
- Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, 71(9), 1–25. doi: 10.18637/jss.v071.i09
- Eith, C., & Durose, M. R. (2011). *Contacts Between Police and The Public, 2008* (Vol. BJS 2008; Tech. Rep. No. 234599).
- Eriksson, K., Strimling, P., & Coultas, J. C. (2014). Bidirectional associations between descriptive and

- injunctive norms. *Organizational Behavior and Human Decision Processes*, 129, 59–69. doi: 10.1016/j.obhdp.2014.09.011
- Federal Bureau of Investigation. (2017). *Crime in the United States* (Tech. Rep.). Retrieved from <https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/topic-pages/violent-crime>
- Fountain, H., & Schmidt, M. S. (2016, jul). 'Bomb Robot' Takes Down Dallas Gunman, but Raises Enforcement Questions. New York. Retrieved from https://www.nytimes.com/2016/07/09/science/dallas-bomb-robot.html?smid=tw-nytimes&smtyp=cur&_r=0
- Fridell, L. a., & Binder, A. (1992). Police officer decisionmaking in potentially violent confrontations. *Journal of Criminal Justice*, 20(5), 385–399.
- Friedersdorf, C. (2015, jun). It's Time for Good Cops to Do Something About Bad Cops. *The Atlantic*. Retrieved from http://www.theatlantic.com/politics/archive/2015/06/its-time-for-good-cops-to-do-something-about-bad-cops/396890/{\#}disqus{_}thread
- Friedersdorf, C. (2017). *Footage of the Police Shooting of Daniel Shaver - The Atlantic*. Retrieved 2017-12-10, from <https://www.theatlantic.com/politics/archive/2017/12/footage-of-a-police-killing-that-jurors-didnt-punish/547868/>
- Friedman, B. (2017). *Unwarranted : policing without permission*. New York: Farrar, Straus and Giroux.
- Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender Differences in Responses to Moral Dilemmas: A Process Dissociation Analysis. *Personality and Social Psychology Bulletin*, 41(5), 696–713. Retrieved from <http://psp.sagepub.com/cgi/doi/10.1177/0146167215575731> doi: 10.1177/0146167215575731
- Fryer Jr., R. G. (2016). *An empirical analysis of racial differences in police use of force*.
- Garner, J. H., Schade, T., Hepburn, J., & Buchanan, J. (1995). Measuring the continuum of force used by and against the police. *Criminal Justice Review*, 20(2), 146–168.
- Gerber, M. M., & Jackson, J. (2017). Justifying Violence: Legitimacy, Ideology and Public Support for Police Use of Force. *Psychology, Crime and Law*, 23(1), 79–95.
- Graham, J., Haidt, J., & Nosek, B. a. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5), 1029–1046.
- Graham v. Connor*, 490 U.S. 3. (1989).
- Gureckis, T. M., Martin, J., McDonnell, J., Alexander, R. S., Markant, D. B., Coenen, A., . . . Chan, P.

- (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*, 1–15. doi: 10.3758/s13428-015-0642-8
- Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Knopf Doubleday Publishing Group.
- Hall, A. V., Hall, E. V., Perry, J. L., & Hall, A. V. (2016). Black and Blue : Exploring Racial Bias and Law Enforcement in the Killings of Unarmed Black Male Civilians. , *71*(3), 175–186.
- Handelman, S. (2014). *Ferguson Leads TCR's Top News Stories of 2014*. Retrieved from <http://www.thecrimereport.org/news/inside-criminal-justice/2014-12-ferguson-leads-tcrs-top-news-stories-of-2014>
- Harris, S., & Loury, G. C. (2016). *Racism and Violence in America*. Retrieved from <https://www.samharris.org/podcast/item/racism-and-violence-in-america>
- Hickman, M. J., Atherley, L. T., Lowery, P. G., & Alpert, G. P. (2015). Reliability of the Force Factor Method in Police Use-of-Force Research. *Police Quarterly*, 1098611115586175–. doi: 10.1177/1098611115586175
- Hume, D. (2003). *A treatise of human nature*. Dover Publications.
- Hyland, S., Langton, L., & Davis, E. (2015). *Police Use of Nonfatal Force, 2002–11* (Tech. Rep. No. November). U.S. Department of Justice Bureau of Justice Statistics.
- Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE*, 7(8).
- Jenkinson, E., Neeson, C., & Bleetman, A. (2006). The relative risk of police use-of-force options: Evaluating the potential for deployment of electronic weaponry. *Journal of Clinical Forensic Medicine*, 13(5), 229–241. doi: 10.1016/j.jcfm.2005.11.006
- Johnson, D., & Kuhns, J. B. (2009). Striking Out: Race and Support for Police Use of Force. *Justice Quarterly*, 26(3), 592–623. doi: 10.1080/07418820802427825
- Johnson, J. A., Hogan, R., & Zonderman, A. B. (1981). Moral judgment, personality, and attitudes toward authority. *Journal of Personality and Social Psychology*, 40(2), 370–373. doi: 10.1037/0022-3514.40.2.370
- Kahan, D. M., Hoffman, D. A., & Braman, D. (2009). Whose eyes are you going to believe? Scott V. Harris and the perils of cognitive illiberalism. *Harvard Law Review*, 122(3), 837–906. doi:

10.2307/40379726

- King, S. (2016). *Shaun King Facebook Post and Commentary RE Dallas Shooting*. Retrieved 2017-01-01, from <https://m.facebook.com/shaunking/posts/1084886048216979>
- Klinger, D. A. (1995). The micro-structure of nonlethal force: baseline data from an observational study. *Criminal Justice Review*, 20(2), 169–186.
- Klinger, D. A., & Slocum, L. A. (2017). Critical Assessment of an Analysis of a Journalistic Compendium of Citizens Killed by Police Gunfire. *Criminology & Public Policy*, 16(1), 1–14. doi: 10.1111/1745-9133.12283
- Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and STAN* (2nd ed.). Academic Press.
- Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian analysis for newcomers. *Psychonomic Bulletin & Review*, 25(1), 155–177. doi: 10.3758/s13423-017-1272-1
- Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178–206. doi: 10.3758/s13423-016-1221-4
- Lemoine, P. (2017). *Police Violence against Black Men Is Rare: Here's What the Data Actually Say*. Retrieved from <http://www.nationalreview.com/article/451466/police-violence-against-black-men-rare-heres-what-data-actually-say>
- 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(September), 228–242. doi: 10.1037/xge0000365
- Loury, G. C. (2015). *Ferguson Won't Change Anything. What Will?* Retrieved 2016-11-16, from <http://bostonreview.net/forum/glenn-c-loury-ferguson-wont-change-anything-what-will>
- McWhorter, J. (2016, jul). Police kill too many people—white and black. *Time*. Retrieved from <http://time.com/4404987/police-violence/>
- Morin, R., Parker, K., Stepler, R., & Mercer, A. (2017). Behind the Badge. *Pew Research*, 11.
- Nix, J., Campbell, B. A., Byers, E. H., & Alpert, G. P. (2017). A Bird's Eye View of Civilians Killed by Police in 2015: Further Evidence of Implicit Bias. *Criminology and Public Policy*, 16(1), 309–340. doi: 10.1111/1745-9133.12269

- Paolacci, G., & Chandler, J. (2014). Inside the Turk. *Current Directions in Psychological Science*, 23(3), 184–188. doi: 10.1177/0963721414531598
- Pew. (2014). Pew: Sharp Racial Divisions in Reactions to Brown, Garner Decisions. Retrieved from <http://www.people-press.org/files/2014/12/12-8-14-Police-Race-release.pdf>
- Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Penguin Publishing Group.
- Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*.
- Reaves, B. A. (2011). *Census of State and Local Law Enforcement Agencies, 2008* (Vol. Census of; Tech. Rep. No. July). Bureau of Justice Statistics. doi: 10.3886/ICPSR27681
- Reaves, B. A. (2012). *Federal Law Enforcement Officers, 2008* (Tech. Rep. No. June). Bureau of Justice Statistics.
- Selby, N. (2016). *Guest post: The ‘low-hanging fruit’ of police reform - The Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-watch/wp/2016/06/20/guest-post-the-low-hanging-fruit-of-police-reform/?utm{_}term=.85743965ca24
- Silver, J. R., & Pickett, J. T. (2015). Toward a better understanding of politicized policing attitudes: Conflicted conservatism and support for police use of force. *Criminology*, 53(4), 650–676. doi: 10.1111/1745-9125.12092
- Skinner, A. L., & Haas, I. J. (2016). Perceived threat associated with police officers and black men predicts support for policing policy reform. *Frontiers in Psychology*, 7(JUL), 1–17. doi: 10.3389/fpsyg.2016.01057
- Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: another contributor to attitude strength or something more? *Journal of personality and social psychology*, 88(6), 895–917.
- Stan Development Team. (2016). *RStan: the R interface to Stan*. Retrieved from <http://mc-stan.org>
- Sullivan, A. (2017, sep). NFL Protests Obscure the Facts on Race and Policing. *New York Magazine*. Retrieved from <http://nymag.com/daily/intelligencer/2017/09/nfl-protests-obscure-the-facts-on-race-and-policing.html>
- Team, R. (2016). *RStudio: Integrated Development Environment for R*. Retrieved from <http://www.rstudio.com/>
- Tennessee v. Garner*, 471 U.S. 1. (1985).

- Terrill, W., Alpert, G. P., Dunham, R. G., & Smith, M. R. (2003). A management tool for evaluating police use of force: an application of the force factor. *Police Quarterly*, 6(2), 150–171. doi: 10.1177/1098611102250491
- Terrill, W., & Mastrofski, S. D. (2002). Situational and officer-based determinants of police coercion. *Justice Quarterly*, 19(2), 215–248. doi: 10.1080/07418820200095221
- Terrill, W., & Paoline, E. A. (2013). Examining Less Lethal Force Policy and the Force Continuum: Results From a National Use-of-Force Study. *Police Quarterly*, 16(1), 38–65. doi: 10.1177/1098611112451262
- Terrill, W., Paoline, III, E. A., & Ingram, J. (2011). *Final Technical Report Draft: Assessing Police Use of Force Policy and Outcomes* (Tech. Rep.). National Institute of Justice. Retrieved from <https://www.nij.gov/publications/pages/publication-detail.aspx?ncjnumber=237794> doi: NCJNumber:237794
- Tyler, T. R. (2004). Enhancing Police Legitimacy. *The annals of the American academy of political and social science*, 593(1), 84–99.
- Tyler, T. R., Goff, P. A., & MacCoun, R. J. (2015). *The Impact of Psychological Science on Policing in the United States: Procedural Justice, Legitimacy, and Effective Law Enforcement* (Vol. 16) (No. 3). doi: 10.1177/1529100615617791
- Weitzer, R. (2005). Determinants of Public Satisfaction with the Police. *Police Quarterly*, 8(3), 279–297.
- Wikipedia Contributors. (2017). *2016 shooting of Dallas police officers*. Retrieved 2017-01-01, from <https://en.wikipedia.org/w/index.php?title=2016 shooting of Dallas police officers&oldid=792017195>
- Worden, R. E. (2015). The causes of police brutality: The "causes" of police brutality: Theory and evidence on police use of force. In E. R. Maguire & D. E. Duffee (Eds.), *Criminal justice theory: Explaining the nature and behavior of criminal justice* (2nd ed., pp. 149–200). Routledge.

Table 1

Six ordinal levels of force (SOLOF) scale.

Level	Description
1	Cooperation, physical proximity, polite dialogue.
2	Strong verbal interaction.
3	Open hand control, defensive resistance, non-deadly weapon use with minimal injury potential.
4	Closed hand techniques.
5	Intermediate or aggravated force, non-deadly weapon used with moderate injury potential.
6	Deadly force.

Table 2
Estimated scale values for actions

	Action (officer/civilian)	Officer Force	Officer Moral	Civilian Force	Civilian Moral
1	asked	0.98 (0.80, 1.16)	0.99 (0.79, 1.17)	1.11 (0.91, 1.28)	1.01 (0.88, 1.16)
2	waved	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)
3	stoodNear	1.06 (0.89, 1.24)	1.19 (1.00, 1.38)	1.24 (1.07, 1.41)	1.52 (1.40, 1.65)
4	nameCall	1.77 (1.62, 1.92)	2.56 (2.38, 2.70)	1.97 (1.82, 2.12)	2.60 (2.48, 2.70)
5	getBack	1.85 (1.70, 2.00)	1.51 (1.34, 1.70)	2.14 (2.00, 2.29)	2.47 (2.36, 2.57)
6	hurtThreat	2.18 (2.02, 2.30)	2.25 (2.07, 2.39)	2.40 (2.25, 2.53)	3.08 (2.97, 3.18)
7	handsUp/dontListen	2.58 (2.45, 2.72)	2.38 (2.19, 2.52)	2.45 (2.32, 2.59)	2.99 (2.88, 3.08)
8	handcuff/pulledAway	2.81 (2.69, 2.95)	2.05 (1.88, 2.22)	2.36 (2.22, 2.50)	2.44 (2.34, 2.55)
9	pepperSpray	3.52 (3.39, 3.64)	3.19 (3.03, 3.33)	3.82 (3.70, 3.96)	4.19 (4.09, 4.29)
10	armbar	3.72 (3.58, 3.84)	2.96 (2.80, 3.11)	4.14 (4.02, 4.28)	4.28 (4.18, 4.38)
11	flashlight	3.79 (3.67, 3.92)	3.84 (3.70, 4.00)	4.12 (4.00, 4.26)	4.24 (4.14, 4.34)
12	batonLeg/pipeLeg	3.82 (3.69, 3.96)	3.52 (3.36, 3.66)	4.36 (4.22, 4.48)	4.49 (4.38, 4.58)
13	punchFace	3.91 (3.77, 4.03)	4.02 (3.88, 4.17)	4.14 (4.01, 4.26)	4.27 (4.17, 4.37)
14	tazer	3.94 (3.82, 4.07)	3.27 (3.11, 3.41)	4.29 (4.16, 4.43)	4.53 (4.42, 4.62)
15	headbutt	3.98 (3.85, 4.10)	4.00 (3.86, 4.16)	4.19 (4.06, 4.31)	4.28 (4.18, 4.38)
16	kickStom	3.99 (3.86, 4.12)	3.98 (3.84, 4.13)	4.05 (3.93, 4.18)	4.25 (4.13, 4.34)
17	choke	4.07 (3.94, 4.20)	3.80 (3.64, 3.94)	4.30 (4.18, 4.45)	4.46 (4.36, 4.56)
18	beanbag/ballBat	4.20 (4.08, 4.35)	4.01 (3.86, 4.16)	4.44 (4.31, 4.58)	4.63 (4.53, 4.74)
19	headSlam	4.52 (4.39, 4.66)	4.97 (4.81, 5.12)	4.80 (4.66, 4.94)	4.93 (4.82, 5.03)
20	vehicle	5.07 (4.93, 5.22)	5.83 (5.69, 6.03)	5.25 (5.11, 5.39)	5.28 (5.17, 5.39)
21	handgunChest	5.43 (5.27, 5.58)	5.50 (5.35, 5.66)	5.70 (5.55, 5.87)	5.60 (5.49, 5.73)
22	handgunHead	6 (6, 6)	6 (6, 6)	6 (6, 6)	6 (6, 6)

Numerical entries indicate the modal estimate and 95% HDI. Rows are ordered by ascending estimates of the "Officer Force" column. Please see Figure 1 for expanded descriptions of the action names.

SOLOF Level	Action Statement	Action Name
1	the [officer/civilian] stood near the [civilian/officer]	stoodNear
	the [officer/civilian] asked the [civilian/officer] how he was doing	asked
	the [officer/civilian] waved at the [civilian/officer] in greeting	waved
2	the [officer/civilian] yelled “get back” at the [civilian/officer]	getBack
	the [officer/civilian] shouted “get your fucking hands up!” / “I don't have to fucking listen to you!” at the [civilian/officer]	handsUp/ dontListen
	the [officer/civilian] said “if you touch me, you're gonna get hurt” to the [civilian/officer]	hurtThreat
	the [officer/civilian] called the [civilian/officer] a “thug” / “pig”	nameCall
3	the [officer/civilian] put handcuffs on/pulled away from the [civilian/officer]	handcuff/ pullAway
	the [officer/civilian] sprayed the [civilian/officer] with pepper spray	pepperSpray
	the [officer/civilian] shocked the [civilian/officer] with a tazer	tazer
	the [officer/civilian] twisted the arm of the [civilian/officer], forcing him to the ground	armbar
4	the [officer/civilian] punched the [civilian/officer] in the face	punchFace
	the [officer/civilian] kicked the [civilian/officer] in the stomach	kickStom
	the [officer/civilian] applied a chokehold to the [civilian/officer]	choke
	the [officer/civilian] headbutted the [civilian/officer] in the face	headbutt
5	the [officer/civilian] struck the [civilian/officer] on the leg with a collapsible baton/metal pipe	batonLeg/ metalPipe
	the [officer/civilian] shot/struck the [civilian/officer] in the chest with a beanbag shotgun round/baseball bat	beanbag/ ballBat
	the [officer/civilian] struck the [civilian/officer] in the stomach with a metal flashlight	flashlight
6	the [officer/civilian] shot the [civilian/officer] in the chest with a handgun	handgunChest
	the [officer/civilian] shot the [civilian/officer] in the head with a handgun	handgunHead
	the [officer/civilian] struck the [civilian/officer] with a motor vehicle	vehicle
	the [officer/civilian] slammed the [civilian/officer]'s head into the concrete curb	headSlam

Figure 1. All 22 actions are listed in the middle column, with their SOLOF level in the left column and abbreviated name in the right column. Colors indicate the SOLOF level, and will be used in subsequent graphs.

While working a patrol shift, a police officer was dispatched to a reported crime, and in the course of the investigation, he came into contact with a civilian. During their subsequent interaction, the following actions took place *in the order listed below* (other actions may have occurred before, between, or after).

1. **the civilian headbutted the officer in the face**
2. **the officer put handcuffs on the civilian**

Please rate how *morally acceptable* the *officer's* action is:

0 10 20 30 40 50 60 70 80 90 100

Completely Unacceptable Completely Acceptable

Submit Answer

Figure 2. Example screenshot from the experiment.

Figure 3. Latent scale values from ratings of force. The upper panel plots the latent action estimates for the civilian, and the horizontal axis denotes γ in Equation 1. The lower panel plots the latent action estimates for the officer, and the horizontal axis denotes χ in Equation 1.

Figure 4. Latent scales for moral ratings (i.e., appropriateness, acceptability, and punishment). The upper panel plots the latent action estimates for the civilian, and the horizontal axis denotes γ in Equation 1. The lower panel plots the latent action estimates for the officer, and the horizontal axis denotes χ in Equation 1.

Figure 5. Posterior estimates of latent action scale values for the *officer*. The scale values for the moral questions are plotted on the vertical axis as a function of the estimated scale values for the force questions on the horizontal axis. Segments through each point represent the 95% HDI's. Each point is labeled by the abbreviated action name in Table 1.

Figure 6. Posterior estimates of latent action scale values for the *civilian*. The scale values for the moral questions are plotted on the vertical axis as a function of the estimated scale values for the force questions on the horizontal axis. Segments through each point represent the 95% HDI's. Each point is labeled by the abbreviated action name in Figure 1.