The European Journal of International Law Vol. 31 no. 2

Unreliable Protection: An Experimental Study of Experts' In Bello *Proportionality Decisions*

Daniel Statman,[*] Raanan Sulitzeanu-Kenan,[**] Micha Mandel,[***] Michael Skerker[****] and Steven De Wijze[*****]

Abstract

The proportionality principle is an international humanitarian law requirement intended to constrain the use of military force in order to protect civilians in armed conflicts. This research experimentally assesses the reliability of its application by legal and moral experts (in 11 countries), by military officers (in two countries) and by laypeople. Reliability was evaluated according to three criteria: inter-expert convergence; sensitivity to relevant factors; and robustness – relative (lack of) susceptibility to biases. Unlike laypeople, experts and military officers performed well on the sensitivity criterion and manifested an appropriate understanding of the principle at the abstract level. However, both groups of experts failed to reach reasonable judgment convergence. These findings cast doubt on the reliability of the protection provided to civilians during warfare, even when warring parties attempt to abide by the proportionality principle.

[*] Department of Philosophy, Haifa University, Israel. Email: dstatman@research.haifa.ac.il.
[**] Federmann School of Public Policy and Department of Political Science, Hebrew University of Jerusalem, Israel. Email: raanan.s-k@mail.huji.ac.il.
[**] Department of Statistics and Data Science, the Hebrew University of Jerusalem, Israel. micha.mandel@mail.huji.ac.il.
[****] Department of Leadership, Ethics, and Law, United States Naval Academy, Annapolis, MD, USA. Email: skerker@usna.edu.
[*****] School of Social Sciences, Manchester University, UK. Email: dewijze@manchester.ac.uk.
We thank Keith Dowding, Mordechai Kremnitzer, Ilana Ritov and participants of seminars at Harvard University, Auckland University, Haifa University, Hebrew University, the Israeli Democracy Institute, London School of Economics, Rutgers University and Southern Denmark University for valuable comments. Excellent research assistance was provided by Marina Motsenok and Victoria Uchitel. Funding for this research was received from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC Grant no. 324182.

1 Introduction

A major goal of international humanitarian law (IHL) is the protection of civilians in armed conflicts. Central to this aim is the proportionality principle that prohibits the use of force when collateral harm to civilians is expected to be disproportionate to the military value of an attack.[1] Application of the proportionality principle thus requires balancing the military value of an attack against the foreseeable harm to civilians. Yet, despite the importance of the proportionality principle, little is known about the capacity of experts to apply it. The present study attempts to offer initial empirical evidence on the reliability of expert proportionality judgments.

Normative judgments lack an objective benchmark to determine their truth value.[2] To overcome this challenge in assessing the validity of proportionality judgments, we rely on three criteria of judgment reliability: inter-expert convergence; sensitivity to relevant factors; and robustness – relative (lack of) susceptibility to biases. Inter-evaluator convergence refers to the distribution of judgments regarding a given situation across a set of evaluators. Perfect, or at least reasonable, convergence among expert judgments comprises a necessary (though insufficient) condition for their ability to identify the true proportional response. The secondary measure of judgment reliability rests on experts' sensitivity to variations in military value. The third reliability criterion refers to the extent to which judgments of proportionality are susceptible to irrelevant conditions (biases), such as the order in which the evaluator is presented with different scenarios, the temporal perspective (judging a future versus past event) or exposure to a numerical anchor.

These measures of judgment reliability were implemented in a novel vignette-based experimental paradigm to assess the reliability of *in bello* proportionality judgments of: (i) academic experts in the ethics and law of war from 11 countries (N = 289); (ii) military officers from the USA and Israel (N = 234); and (iii) a sample of US non-experts (N = 960). Unlike laypeople, academic experts and military officers performed well on the sensitivity criterion and generally manifested an appropriate understanding of the principle at the abstract level. However, they did not reach reasonable judgment convergence. Academic experts were no less susceptible to biases than non-experts, while no significant biases were found in the case of military officers.

Two interesting findings that go beyond the stated aim of this study are also reported and discussed. First, cultural differences in the application of *in bello* proportionality were found in both types of expert groups. The median response of American academic experts and military officers was higher (that is, more permissive), and their level of judgment convergence was lower, compared to their respective non-American counterparts. Second, our findings point to a consistent relationship between the median proportionality judgment of a group and its judgment convergence, in line with the previous findings of 'psychic numbing' in valuations of human lives.[3]

1 Hurka, 'Proportionality in the Morality of War', 33(1) *Philosophy and Public Affairs* (2005) 34; M. Walzer, *Just and Unjust Wars: A Moral Argument with Historical Illustrations* (1977).

2 Gert and Gert, 'The Definition of Morality', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (2017).

3 Dickert *et al.*, 'Scope Insensitivity: The Limits of Intuitive Valuation of Human Lives in Public Policy', 4(3) *Journal of Applied Research in Memory and Cognition* (2015) 248.

The results of this research carry important implications for the ethics and laws of armed conflict. The apparent inability of experts – both academics and military officers – to implement the proportionality principle in a reliable manner, casts doubt on the merit of their contribution in guiding behavior in warfare as well as on their potential role in post-war assessments of the legality of military actions. If proportionality judgments are unreliable, so is the protection of civilians during warfare, even when warring parties attempt to abide by the proportionality principle.

The next part briefly reviews the principle of proportionality in war. The third part discusses the challenge of evaluating normative judgments, offering a simple formal model of proportionality and developing the three measures of judgment reliability. The fourth part describes the experimental design and empirical methods used. This part is then followed by a discussion of the results. In this discussion, we summarize our findings and discuss their implications for our main research question. We conclude with a discussion of the normative and policy implications of the results.

2 Proportionality in War

Just war theory and IHL impose a normative distinction between combatants and civilians in war and oblige military forces to restrict the use of force in order to provide a certain level of protection for civilians. This protection is delineated by two fundamental principles of just war theory: (i) civilians ought never to be intentionally targeted, and (ii) although civilians may be harmed as a side effect of legitimate attacks on military targets, the harm they suffer must not be disproportionate[4] to the military value of the attack.[5] Thus, states may use military force to fight their enemies effectively by attacking military targets even when such attacks place civilians at risk of harm, yet this permission is constrained by the requirement of proportionality.

Contrary to a popular understanding, proportionality is not merely a matter of counting casualties, so to say, assuming that if the number of enemy civilians killed in an attack exceeds the number of enemy soldiers, the attack is disproportionate and therefore illegitimate. Legitimate military goals are not restricted to the killing of enemy soldiers. For example, destroying a central communication centre may be of high military value, even if no enemy soldiers are killed in its course. Given its value, attacking the centre might be justified even if it leads to collateral harm among civilians. What proportionality in warfare requires is to compare the military value

[4] Or 'excessive', as per the Geneva Convention's wording. Additional Protocol I to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts 1977, 1125 UNTS 3, Article 57(2)(a)(iii). Note that this article addresses proportionality in the context of *jus in bello* – the rules that govern the means used to fight a war. The proportionality principle also plays a role in *jus ad bellum* – the rules determining the permissibility of entering into war – where it requires states to weigh the benefits that they expect to achieve against the costs or harms of such an undertaking. If the latter outweigh the former, then the war is morally unjustified and ought not to be launched. See Hurka, *supra* note 1, at 35–36.

[5] Hurka, *supra* note 1; Walzer, *supra* note 1.

of the attack with the foreseeable (yet unintended) harm to civilians (or to civilian infrastructure).[6]

Together with the absolute prohibition on the deliberate targeting of civilians, the proportionality requirement is meant to provide reasonable protection to civilians in times of war. Of course, it would be better if civilians could be spared the savagery of war altogether by granting them immunity, even from collateral harm. However, that would undermine the ability of countries to effectively use military force when they have a just cause for doing so, as in clear cases of national defence. This compromise is reflected by international humanitarian law, which allows states to carry out attacks even when harm to civilians is foreseen, provided that the harm is not disproportionate.[7]

3 Assessing the Application of the Proportionality Principle

In recent decades, the proportionality principle has attracted much attention in discussions about the morality and legality of armed conflicts.[8] The discourse on proportionality by academics, journalists and politicians typically posits that failures to comply with this principle are either intentional or due to negligence. Some recent studies also point to the practical problem of inaccurate accounting of civilian losses[9] as a potential source of unintentional disproportionate military actions. Still, an assumption shared by all previous analyses of *in bello* proportionality is that when warring parties choose to abide by the proportionality principle, they can practically implement it. The present study seeks to examine whether this is indeed the case.

For the sake of clarity in developing our argument, we utilize the following simple formal representation of proportionality in war: let V_{LT} represent the military value of legitimate target *LT*. The proportionality principle implicitly assumes that V_{LT} can be elicited from the characteristics of a military target and explicitly determines that the maximum number of collateral civilian casualties permitted (that is, not 'excessive')

[6] The traditional understanding of *in bello* proportionality (expressed, for instance, in the above citation from the Geneva Convention) suggests that it weighs the military value, in terms of advancement of victory against the harm to civilians and to civilian infrastructure, separately from the issue of *ad bellum* proportionality – that is, the justification of the war. This traditional understanding has been challenged by a few philosophers, notably Jeff McMahan, who posit that combatants of an unjust war can never satisfy the *jus in bello* requirement of proportionality. J. McMahan, *Killing in War* (2009), at 18–32. We avoid this controversy by focusing on military operations undertaken by combatants waging a clearly just war.

[7] For an attempt to ground the permissibility of this collateral harm in an agreement between the international players, see Y. Benbaji and D. Statman, *War by Agreement* (2019).

[8] Barber, 'The Proportionality Equation: Balancing Military Objectives with Civilian Lives in the Armed Conflict in Afghanistan', 15(3) *Journal of Conflict and Security Law* (2010) 467; J. Gardam, *Necessity, Proportionality and the Use of Force by States* (2004); Gross, 'The Second Lebanon War: The Question of Proportionality and the Prospect of Non-Lethal Warfare', 7(1) *Journal of Military Ethics* (2008) 1; Hurka, *supra* note 1.

[9] Jewell, Spagat and Jewell, 'Accounting for Civilian Casualties: From the Past to the Future', 42(3) *Social Science History* (2018) 379; Sloboda, 'Can There Be Any "Just War" If We Do Not Document the Dead and Injured?' (2008), available at http://www.fredsakademiet.dk/tid/2000/2008/april08/org.pdf.

– C_{LT}^{max} – is proportional to V_{LT}. Importantly, C_{LT}^{max} is a normative determination rather than a factual one. It refers to the number of civilians it would be legitimate to collaterally kill, given the target's military value, not to the number of civilians that are expected to be killed. The relationship between C_{LT}^{max} and V_{LT} can be represented by the following linear equation: $C_{LT}^{max} = \alpha V_{LT}$, where α is a positive coefficient for converting military value to a maximum number of collateral civilian casualties ($\alpha > 0$). An additional realistic constraint is that the maximum number of collateral civilian casualties permitted is a natural number: $C_{LT}^{max} \in N^0$.[10] Since α depends on a normative judgment and V_{LT} is determined subjectively, $C_{LT}^{max} = \alpha V_{LT}$ does not provide a unique value for C_{LT}^{max} given a specific legitimate target but, rather, assumes that such cognitive processes can yield decisions that conform to proportionality.

The proportionality principle is therefore not self-applicable; humans – typically, experts – carry out its practical implementation. Such experts include international lawyers and moral philosophers who specialize in this field of knowledge as well as military officers. In order to assess the reliability of applying the proportionality principle, we examine proportionality judgments of academic experts in the legality and ethics of war (N = 289) from 11 countries, military officers (N = 234) from two countries and a representative sample of the US population (N = 960). The selection of academic and military experts from multiple countries enhances the external validity of our sample, as they represent the relevant experts involved in such decisions. The two types of experts differ in the sense that academic experts possess extensive legal and moral knowledge, while military officers typically have less formal knowledge about proportionality in armed conflict, but possess more specialized practical experience in such decision-making. The representative sample of the US population provides a baseline of laypeople's intuitions regarding proportionality in war, a baseline that enables us to identify the role of expertise in forming proportionality judgments. More specifically, the comparison to laypeople may facilitate the interpretation of the results. If experts are found to apply the proportionality principle reliably (our main research question), we can further assess whether expertise is a necessary condition for this capacity.

A Measures of Reliability

The foremost challenge in any attempt to assess the validity of proportionality judgments stems from their normative nature. With regard to normative judgments, there is no empirical test that can determine their truth-value; this characteristic lies at the core of the fundamental distinction between descriptive and normative judgments.[11] To (partially) overcome the inability to directly assess the truth-value of proportionality judgments, we propose three measures of judgment reliability: (i) convergence,

[10] Note that there can be a level of military value that is greater than zero but does not justify the risk of even one civilian casualty. Such low military value targets may justify collateral damage to property and/or an expected number of casualties that is less than one – for example, a probability of 0.30 that one civilian will be killed.

[11] Gert and Gert, *supra* note 2.

(ii) sensitivity and (ii) robustness. Our key measure of reliability is inter-evaluator convergence, which refers to the distribution of judgments regarding a given dilemma across a set of evaluators.[12] Unreasonable divergence of views regarding common tasks points to the dominance of opinion rather than expertise.[13] Reasonable convergence among expert judgments comprises a necessary condition for their ability to identify the true proportional response,[14] yet it is not sufficient – as they may collectively be wrong. Conversely, non-convergence among experts casts serious doubt on their capacity to identify the true proportional response.

A possible objection to this criterion of judgment reliability might be that mere disagreement provides no reason to retreat from anyone's view nor even to moderate it. Surely, 'truth' is not a matter of a majority decision. However, even if one expert expresses the true answer, the normative nature of the judgment bars us from identifying this expert. In line with the philosophical problem of peer disagreement,[15] if two people disagree on a matter regarding which they are epistemically equal in terms of their intelligence, education, training and (relevant) knowledge, then neither has any basis to assume that she herself (rather than the other person) got it right. In such cases, many philosophers believe that both sides must suspend judgment.[16] When sizable samples of epistemic communities fail to converge on the normative estimates for the number of permissible civilian casualties, only extreme vanity would allow an individual evaluator to believe that she enjoys a privileged epistemic position that would enable her to answer the question correctly in contrast to all others. Under such circumstances, the inevitable conclusion is that the epistemic community has no answer to it.[17]

The measure of convergence yields itself to descriptive analyses rather than to a clear hypothesis testing, since any threshold of sufficient convergence may be contested. Our analyses of judgment convergence within each sub-sample are thus descriptive, allowing readers to draw their independent normative conclusions from the raw results. These are complemented by two testable hypotheses. The first takes a comparative approach, by positing that judgment convergence among experts should be higher compared to non-expert groups:

> *H1a: Judgment convergence among academic experts and military officers is higher than among lay respondents.*

The second hypothesis tests judgment convergence against a measure of relative convergence. Previous findings suggest that valuations of human lives are subject to the

[12] Dowding, 'Moral and Political Expertise', working paper (2016).

[13] D. Kahneman, *Thinking Fast and Slow* (2011), ch. 21.

[14] Adcock, 'Measurement Validity: A Shared Standard for Qualitative and Quantitative Research', 95(3) *American Political Science Review* (*APSR*) (2001) 529.

[15] Goldman, Alvin and Blanchard, 'Social Epistemology', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (2001), especially s. 3.4.

[16] Elga, 'Reflection and Disagreement', 41(3) *Noûs* (2007) 478; Feldman and Richard, 'Reasonable Religious Disagreements', in L.M. Antony (ed.), *Philosophers without Gods* (2007) 194.

[17] Cross, 'Moral Philosophy, Moral Expertise, and the Argument from Disagreement', 30(3) *Bioethics* (2016) 188; McGrath, 'Moral Disagreement and Moral Expertise', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics* (2008), vol. 3, 87.

psychophysical principle known as the Weber-Fechner law.[18] This cognitive principle suggests that people's sensitivity to quantities diminishes as they evaluate increasingly larger values;[19] thus, the level of convergence is expected to decrease as larger numbers are considered. To accommodate the diminishing sensitivity to human lives when evaluating increasingly larger numbers, we estimate the ratio between judgment convergence (dispersion), measured by the interquartile range (25th–75th percentile) and the median judgment of a set of evaluators. Based on this relative convergence measure, we propose a practical benchmark of one – that is, the interquartile range must not exceed the value of the median judgment:

H1b: The (lack of) judgment convergence (measured by the interquartile range) does not exceed the median judgment among academic experts and military officers.

Inferring judgment reliability from the level of convergence rests on the assumption that the observed distribution of judgments reflects an aggregate result of multiple instances of individual discretion rather than a mechanistic adherence to an arbitrary rule. To address this, we rely on a second measure of judgment reliability, which gauges evaluators' sensitivity to variations in military value, in line with the proportionality principle.[20] By presenting experts with two different military targets with diverging military values, we complement convergence by measuring central tendency shifts in proportionality judgment distributions. Specifically, since α is greater than zero, in determining C_{LT}^{max} for two legitimate targets with clearly different military values – $V_{LT1} > V_{LT2}$ – the value of C_{LT}^{max} should be greater for the target of higher military value, such that $C_{LT1}^{max} > C_{LT2}^{max}$. Our estimate of sensitivity relies on the ratio (at the individual level, which is detailed below). Given that $V_{LT1} > V_{LT2}$, a ratio greater than one conforms to the sensitivity criterion,[21] yielding the following hypothesis:

H2: The average within-subject ratio between the maximum permissible number of civilian casualties (C_{LT}^{max}) in the case of a high-value target and a low-value target is greater than one.

The third judgment reliability criterion refers to the extent to which judgments of proportionality are susceptible to irrelevant considerations (biases), such as the order in which an evaluator is presented with different scenarios, the temporal perspective (judging a future versus past event) or the exposure to a numerical anchor. Robustness is assessed by the following hypotheses:

H3a: The order of target affects the maximum permissible number of civilian casualties (C_{LT}^{max}).
H3b: The temporal perspective affects the maximum permissible number of civilian casualties (C_{LT}^{max}).
H3c: Exposure to a numerical anchor affects the maximum permissible number of civilian casualties (C_{LT}^{max}).

[18] Dickert, *supra* note 3.

[19] Dehaene, 'The Neural Basis of the Weber–Fechner Law: A Logarithmic Mental Number Line', 7(4) *Trends in Cognitive Sciences* (2003) 145.

[20] Sulitzeanu-Kenan, Kremnitzer and Alon, 'Facts, Preferences, and Doctrine: An Empirical Analysis of Proportionality Judgment', 50(2) *Law & Society Review* (2016) 348.

[21] This measure taps within-subject sensitivity, thereby restricting potential inter-subject variance. Second, choosing a ratio conforms to the psychophysical principle of Weber's Law (detailed below).

These three measures of (un)reliability are closely related. When evaluators do not possess clear and feasible means to adjudicate a particular dilemma, we can expect the distribution of their judgments to be more dispersed. Moreover, the absence of a well-defined formula to address a problem tends to increase susceptibility to implicit biases,[22] a result found also for expert judgments.[23] The only empirical study on experts' judgments of proportionality of which we are aware found evidence for sensitivity to variations in the military value of a potential attack when experts were called upon to decide whether an attack was proportional or not, yet this study also found strong correlational evidence for (ideological) bias in these decisions.[24] However, unlike the current study, this research included a general sample of lawyers from one country rather than specialists in international law from multiple countries, and it did not address the key criterion of judgment convergence. The following part describes the experimental design and methods used for assessing the three measures of judgment reliability.

4 Experimental Design

A *Participants*

All three respondent samples (N = 1,483) participated in a vignette-based experiment, in which they were asked to read two descriptions of wartime military operations and answer questions regarding the permissible collateral damage in each case.[25] We categorized as 'experts' academics who had published academic studies on the morality or legality of war. To create a comprehensive list of academic experts, we conducted a set of searches in the legal database HeinOnline (https://home.heinonline.org) and in the Philosopher's Index database (https://philindex.org) for articles that included a set of relevant keywords.[26] These searches were carried out in English, Dutch, French, German, Italian and Spanish. The raw search results were reviewed to omit substantively irrelevant articles. The resulting authors' list yielded 938 experts.

Administration of the survey experiment followed a standard two-stage approach. Email participation requests were sent to all 938 experts during March 2015. Removing dysfunctional (returned) email addresses from this list left us with 825 international experts who plausibly received an invitation to complete the online questionnaire. Two rounds of email reminders followed the initial invitation. This procedure yielded 289 respondents – a 35 per cent response rate. Beyond the inclusion criteria

[22] Bertrand, Chugh and Mullainathan, 'Implicit Discrimination', 95(2) *American Economic Review* (2005) 94.

[23] Gazal-Ayal and Sulitzeanu-Kenan, 'Let My People Go: Ethnic In-Group Bias in Judicial Decisions: Evidence from a Randomized Natural Experiment', 7(3) *Journal of Empirical Legal Studies* (2010) 403.

[24] Sulitzeanu-Kenan, Kremnitzer and Alon, *supra* note 20.

[25] Institutional review board (IRB) approvals for research on human subjects were obtained from the authors' academic institutions (details omitted here for blind review purposes).

[26] <proportionality in war>, <war AND collateral>, <war AND innocent>, <warfare AND moral> and <in bello>. The term <in bello> yielded too many unrelated HeinOnline results; hence, the more restrictive terms <'jus in bello' AND disproportionate> were used for this database.

that identified 'experts' invited to participate in this research, we added a question at the end of the 'experts' version' of the questionnaire, asking whether respondents considered themselves qualified to serve as panel members or as expert witnesses in a national or international investigation into the morality and legality of military operations. Seventy per cent of our respondents in the expert sample answered 'yes' to this question (for further descriptive details of this sample, see Table S1 in the Appendix).[27]

The sample of military officers (N = 234) includes officers from two countries – the USA and Israel. These particular countries have been involved in active warfare in recent decades, and we thus assumed that officers in their military organizations are more likely to possess knowledge and experience in making *in bello* proportionality judgments. The American officers were recruited by email invitations sent to all of the officers in the military faculty at the US Naval Academy (USNA) during December 2015. Of the 253 officers who received an invitation, 123 completed the questionnaire – a 48.6 per cent response rate. The Israeli officers were recruited by administering the questionnaire to officers who attended the National Security College (MABAL) and the Command and Staff College (PUM) in March 2017. All 111 officers present completed the questionnaire in the lecture halls where it was administered. Importantly, 62 per cent of the officers in our sample reported having combat experience. We estimate that officers with combat experience have a 74.8 per cent chance of making a decision whether to open fire during a combat situation in the course of their career (for details, see online Appendix, p. 3). Lastly, a non-probability sample of the US population (N = 960) provides a set of layperson respondents. Given that, by far, the largest share of academic experts was from the USA, as well as the majority of the military officers, a US sample of laypeople was selected to serve as the non-expert reference group. Respondents were recruited by Qualtrics, utilizing an opt-in panel that covers the US population aged 18 years and older during September 2015. The sample approximates the US population in most respects, including gender, age, ethnicity, income and education (for details, see Table S3 in the Appendix).[28]

B Experimental Procedure

After completing a set of background questions, respondents took part in a vignette-based experiment. Such experiments are used widely in social science and public health research and have been found to have strong external validity in predicting the behaviour of both laypeople[29] and professionals,[30] while avoiding some of the ethical, practical and other limitations involved in alternative methods. Respondents were

[27] The Appendix in available online at https://openscholar.huji.ac.il/sites/default/files/raanansulitzeanukenan/files/unreliable_protection_online_appendix_2020.pdf.

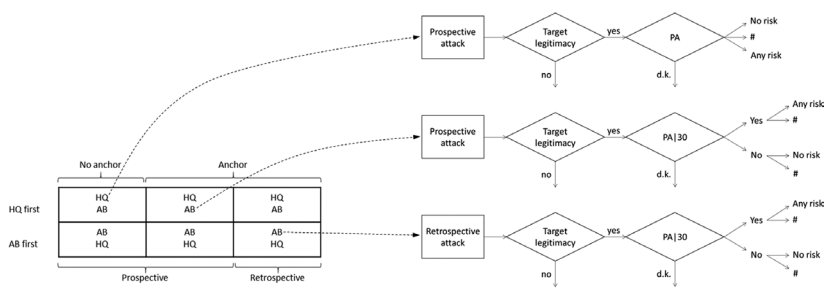
[28] Notable deviations are 4 per cent and 5 per cent over-representation of female and Caucasian respondents, respectively; and 5 per cent and 6 per cent under-representation of Hispanics and low-education respondents ("some high school or less"), respectively.

[29] Hainmueller, Hangartner and Yamamoto, 'Validating Vignette and Conjoint Survey Experiments against Real-World Behavior', 112(8) *Proceedings of the National Academy of Sciences* (2015) 2395.

[30] Evans *et al.*, 'Vignette Methodologies for Studying Clinicians' Decision-Making: Validity, Utility, and Application in ICD-11 Field Studies', 15(2) *International Journal of Clinical and Health Psychology* (2015) 160; Peabody *et al.*, 'Measuring the Quality of Physician Practice by Using Clinical Vignettes: A Prospective Validation Study', 141(10) *Annals of Internal Medicine* (2004) 771.

presented with descriptions of military operations in the context of a war, regarding which they had to determine the maximum permissible number of collateral civilian casualties.[31] As detailed in the Appendix, the preparation of these vignettes included several pilot interviews with a small group of experts. Since the requirement of proportionality is often confused with that of necessity,[32] the vignettes indicated that the necessity requirement was satisfied – that is, that there was no less costly way of attaining the military advantage that was to be obtained by the attack. Respondents were thus confronted with a clear dilemma of *in bello* proportionality, requiring them to strike the proper balance between military value, on the one hand, and harm to (enemy) civilians, on the other.

This design allowed us to assess judgment convergence among each set of respondents. Additionally, by presenting each respondent with descriptions of two military operations (in random order) that vary in their respective military value (within-subject treatment of military value), we were able to estimate sensitivity among respondents. Third, by varying normatively irrelevant attributes of the task, we were able to assess respondents' susceptibility to biases – that is, their judgment robustness. The normatively irrelevant attributes were manipulated by three between-subject treatments: (i) the order in which the two military operations were presented; (ii) exposure (or lack thereof) to a numerical anchor; and (iii) the temporal perspective – that is, whether the operation was presented prospectively or retrospectively. These considerations resulted in a two (military value) by three (temporal perspective and anchor) by two (order) mixed between-within subject design, with military value as the within-subject factor. Figure 1 presents the experimental design graphically. The two military operations, designed to differ in their military value, involved an attack on the 'main



A: Each respondent was randomly assigned to one of the six conditions, in which they read descriptions of two military attacks and answered a set of question regarding collateral damage. The experimental conditions vary the exposure to a numeric anchor, the temporal perspective of the task (prospective/retrospective evaluation), and the task order (Headquarters (HQ), or Airbas (AB) first).

B: Under the 'prospective no-anchor' condition (top flow-chart), respondents were presented with a prospective description of a potential attack, followed by a question regarding the legitimacy of the target and a set of questions regarding permissible collateral damage; Under the 'prospective anchor' condition (middle flow-chart), respondents were asked to consider whether a risk of 30 civilian casualties constitutes permissible collateral damage in the context of the respective attack. After replying, additional questions regarding collateral damage were asked. The 'retrospective anchor' condition had the same structure as 'prospective anchor', but the attacks were presented as if they already occurred, and the questions regarding collateral damage were retrospective.

Figure 1: *Experimental design*

[31] Our empirical analyses were restricted to judgments of attacks carried out by the 'just' side, regarding which the revisionists versus traditionalists controversy is immaterial.

[32] McMahan, *supra* note 6, at 23.

headquarters' (HQ) and an attack on a 'relatively small military airbase' (AB) in a rival country during war (the vignettes are provided in the Appendix).

After reading each of the two scenarios, respondents were asked whether the target in question was a legitimate military target for attack (regardless of collateral damage).[33] This question refers to the principle of distinction. In line with the doctrine, only respondents who considered the target as legitimate were presented with subsequent questions regarding collateral damage.[34] Two questions were used to tap our main dependent variable – respondents' proportionality judgments. (A third question addressing proportionality appeared in the 'anchor' conditions.) The first question was:

> Assuming that an attack on the <airbase> / <headquarters> will cause collateral damage to civilian lives, which of the following options is closest to your judgment:
> i. The attack is permitted only if it poses no risk to civilian lives, otherwise this <airbase> / <headquarters> should not be attacked. (For the sake of simplicity, please ignore other harms such as injuries, damage to infrastructure and so on.)
> ii. The attack is permitted at almost any risk to civilian lives.
> iii. The attack is permitted only if the estimated number of civilian causalities does not exceed [*please write a number*]; otherwise, the <airbase> / <headquarters> should not be attacked.
> iv. I cannot offer a reasoned answer to this question.

This question offered respondents the entire range of possible answers to the question of collateral damage. The first two options suggested extreme responses – either zero or any number of civilian casualties as permissible collateral damage. Respondents who rejected these options could then either give a numerical response (the software accepted only numbers in this option) or choose not to give a substantive response. While the principle of proportionality does not specify the exact proportional response in a given situation, the principle does point to non-extreme responses. At least in the case of legitimate military targets with non-negligible military value, we would expect the maximum permissible collateral damage to be above zero casualties, though bounded by some figure – that is, inconsistent with the option that any number of casualties is permissible. Given that experts are typically more knowledgeable of the problem categories than novices,[35] we expected respondents who were familiar with the proportionality principle to refrain from the first two extreme options and to opt for the third – providing a proportional number. This structure of categorization enabled us to assess respondents' understanding of the proportionality principle, prior to assessing their ability to implement it by specifying exact numerical responses. The

[33] Do you think that, viewed by itself (namely, before taking into consideration possible collateral harm), the <headquarters> / < airbase> of Army B is a legitimate military target for attack by Army A?

[34] This structure of the experiment is intended to identify the normative reasons for the respondents' decisions. For example, whether they object to the attack due to reasons of target legitimacy or due to disproportionality. While this staged structure may create selection effects (as those who deemed the target as illegitimate are not asked to assess the proportionality of attacking it), it conforms to the process of decision-making regarding the permissibility of attacks, and, therefore, it is realistic to have such selection effect in real-life decisions.

[35] Chi, Feltovich and Glaser, 'Categorization and Representation of Physics Problems by Experts and Novices', 5(2) *Cognitive Science* (1981) 121.

figures (maximum number of collateral civilian casualties) provided by respondents who opted for the third option provide a quantitative indicator of each such respondent's application of the proportionality principle for each of the two military targets described in the scenarios. This measure allowed us to estimate the three reliability criteria. Under the 'anchor' conditions, we elicited respondents' proportionality judgments by presenting the following questions:

> Assuming that an attack on the <headquarters> / <airbase> is estimated to claim the lives of 30 civilians, which of the following options is closest to your judgment?[36]
> i. The attack on the <headquarters> / <airbase> of army B is permitted.
> ii. The attack on the <headquarters> / <airbase> of army B is not permitted, hence army A must refrain from carrying it out.
> iii. I cannot offer a reasoned answer to this question.

After responding to the numerical anchor (30 casualties), each respondent received the appropriate follow-up question, based on his or her answer, so as to obtain a proportionality judgment as in the 'no anchor' condition (above). Thus, respondents who considered the operation permissible given collateral damage of 30 civilian casualties were asked whether the attack would be permissible at 'almost any number' or permissible at 'a specific number' and were offered the option of admitting that they 'cannot answer'. A respondent who considered the operation impermissible given 30 casualties could choose between 'zero casualties', 'a specific number' or 'cannot answer'. The questionnaire structure in the different conditions is presented graphically in Figure 1. This structure provides an adaptation of classical anchoring measures to the context of proportionality decisions.[37] However, given the limited sample sizes of experts and officers, the design includes only one anchor size (30) and a calibration group.

5 Results

We begin by conducting a set of balance tests to assess the effectiveness of the random assignment of respondents to experimental conditions. A set of multinomial logistic regressions were conducted, each estimating the experimental condition based on the available individual characteristic for each sub-sample. None of the models are statistically significant, providing no evidence for imbalance in the assignment of respondents to experimental conditions. These results are reported in Table S6 in the Appendix.

[36] This is the wording under the 'prospective anchor' condition. In the 'retrospective anchor' condition, the question was: given that the attack on the <headquarters> / <airbase> had claimed the lives of 30 civilians, which of the following options is closest to your judgment? Answer options were: '(1) The attack on the <headquarters> / <airbase> of army B was permissible; (2) The attack on the <headquarters> / <airbase> of army B was impermissible; hence army A should have refrained from carrying it out; (3) I cannot offer a reasoned answer to this question.'

[37] E.g. Jacowitz and Kahneman, 'Measures of Anchoring in Estimation', 21(11) *Personality and Social Psychology Bulletin* (1995) 1161.

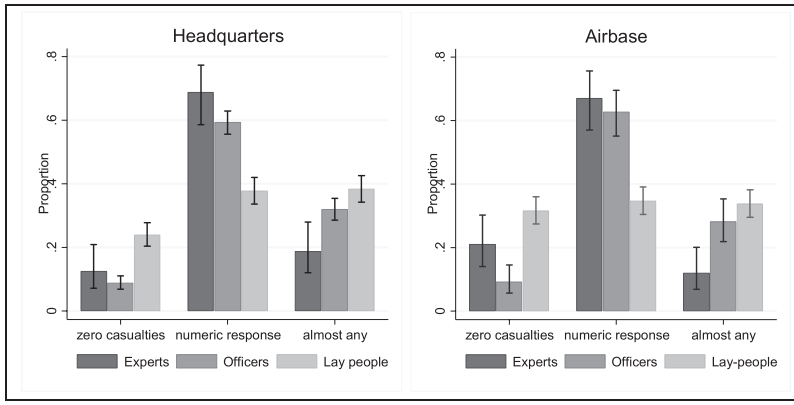


Figure 2: *Proportions of categorical responses. The graph on the left shows responses regarding the 'headquarters' scenario, and that on the right addresses the 'airbase' scenario, with 95% confidence intervals.*

A Target Legitimacy

Immediately after reading each of the two scenarios, respondents were asked whether the target in question was a legitimate military target for attack (regardless of collateral damage). The three groups significantly differed in their responses to this question regarding both targets (HQ: LR$\chi^2 = 107.4$, $p < 0.001$; AB: LR$\chi^2 = 106.10$, $p < 0.001$). Academic experts were more likely to identify the two targets as legitimate (HQ: 91.1%; AB: 89.3%) compared to lay respondents (HQ: 73.9%; AB: 69.9%, $p < 0.001$), and military officers were more likely to identify the two targets as legitimate (HQ: 97.4%; AB: 94.9%) compared to academic experts (HQ: $p = 0.005$; AB: $p = 0.028$).[38]

B Initial Categorical Choices

As explained in the measures of reliability section, the distribution of categorical choices enables us to assess respondents' level of understanding of the proportionality principle. The overall distribution of responses across the three groups differs for both military targets (HQ: LR$\chi^2 = 73.5$, $p < 0.001$; AB: LR$\chi^2 = 79.3$, $p < 0.001$).[39] Experts and military officers were less likely to choose the extreme responses (either 'zero casualties' or 'almost any number') compared to lay respondents, as is evident in Figure 2. These differences are statistically significant based on the set of multinomial logistic regressions (reported in Table S7 in the Appendix).

These results are consistent with our expectation that the distribution of categorical responses reflects respondents' understanding of the proportionality principle. As expected, the modal choice of respondents who were expected to be more knowledgeable about the proportionality principle – experts and officers – was clearly the

[38] The experimental conditions had no significant effect on target legitimacy judgments for any of the groups.

[39] These analyses include respondents who provided a substantive response – that is, they exclude those who answered 'I cannot give a reasoned answer'. Yet the substantive results hold when including the latter respondents in the analysis as well (see Table SI in the Appendix).

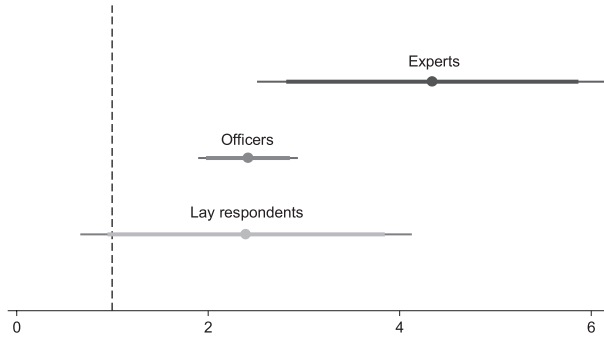


Figure 3: *Sensitivity to target military value in determining the maximum number of collateral civilian casualties permitted. Estimates represent mean (individual level) ratio, with 95% (thin) and 90% (thick) confidence intervals.*

non-extreme one. Among lay respondents, in contrast, the distribution of categorical choices was relatively uniform across the three substantive options. These results support the premise that respondents from the two expert groups – academic experts and military officers – indeed diverge from non-experts in their level of understanding of the proportionality principle.

C Sensitivity

The patterns of categorical choices of experts, officers and lay respondents suggest that the former two groups systematically employ a distinct theory in coping with such dilemmas (in contrast to lay respondents). The choice to provide a numerical response implicitly indicates an intention to apply a proportionate response. The distribution of the numerical responses allows us to assess the extent to which this intention was realized reliably. We begin by assessing respondents' sensitivity to the military value of the targets presented to them. Mean sensitivity levels (ratios) were greater than one and statistically significant – in line with hypothesis 2 – in the samples of academic experts and military officers ($p < 0.001$), but statistically insignificant in the case of the lay respondents ($p = 0.113$), based on one-sample t-tests. These results are presented graphically in Figure 3. Hypothesis 2 is thus supported only among professional respondents. Only these respondents systematically permit relatively larger collateral damage when the military value of the target is greater. Note that the point estimates of the sensitivity levels of lay respondents and officers are not statistically different and that the main difference between these groups is in the variance, which is significantly smaller among officers (K–S = 0.291, $p = 0.001$).[40] This difference reflects officers' ability to consistently identify the different levels of military values, in contrast with lay respondents.

The sample of lay respondents differs from the samples of the experts and officers in gender proportion as well as in the level of education (as shown in Table S4 in the Appendix). To assess whether these characteristics account for the different sensitivity

[40] Based on exact Kolmogorov-Smirnov tests of equality of distributions.

levels across the samples, we regressed the sensitivity of lay respondents on gender and academic education (a dummy for a college degree and above). No significant difference was found between females and males, but sensitivity was higher among the more educated lay respondents at a marginally significant level ($p = 0.061$). These results suggest that education differences may account for the differences between lay respondents and experts and officers. The following subpart provides the main assessment of judgment reliability – that is, inter-judge convergence.

D Convergence

Table 1 presents descriptive statistics for the numerical responses of the three groups. Evidently, the distribution of numerical responses is over-dispersed (standard deviations are larger than the means) in the three groups for the two targets. The large differences between the means and the medians also suggest that the distribution is highly skewed (skewness values of eight and above) and that the samples include extreme outliers.

Given these distributional characteristics, we utilized the median response and percentiles for assessing convergence, as they are robust to distributional assumptions and outliers. Figure 4 graphically presents the distribution of the numerical responses of academic experts, military officers and lay respondents regarding the 'headquarters' (left panel) and the 'airbase' (right panel) targets in two box plots. The y-axes indicate the maximum number of casualties permitted to be at risk in order to carry out the attack. The distribution of each respondent sample is depicted by a box plot, which presents the range between the value (in casualty numbers) of the response at the 25th percentile and the response at the 75th percentile, which is also known as the interquartile range. The numerical value of the interquartile range of each sample is labelled. The horizontal line within each box denotes the median response.

We begin our analysis with the type of respondents most likely to converge in their specific proportionality judgment – academic experts. The median value that experts specified as the maximum number of casualties that may be risked in the case of an attack on an enemy's headquarters is 125, and the box plot shows substantial dispersion in the responses. The interquartile range shows a difference of 575 casualties

Table 1: Numerical responses

	Experts	**Military officers**	**Lay respondents**
Headquarters	2,027 (12,299)	11,448 (100,536)	53,393 (644,907)
Mean (SD)			
Median (P_{25}; P_{75})	125 (50; 625)	50 (30; 200)	47.5 (10; 100)
Observations	66	99	196
Airbase	527 (2,464)	30,888 (301,465)	90,259 (812,081)
Mean (SD)			
Median (P_{25}; P_{75})	50 (12.5; 125)	50 (20; 100)	40 (12; 100)
Observations	69	99	160

Note: SD: standard deviation; P_{25} and P_{75}: 25th and 75th percentile, respectively.

between the 25th and 75th percentile – 4.6 times the value of the median response. In the case of the attack on the tactical target (the military airbase), experts' median response is 50, and the interquartile range presents a difference of 85 casualties, which is 1.7 times the median response for this target. These results, especially regarding the strategic target (HQ), reflect substantial dispersion in the experts' judgments, which fails to support a reasonable level of inter-expert convergence, as even non-extreme responses widely diverge.[41] Quite similar distributions were found for the military officers' sample, yet in the case of the strategic target (HQ), both the median response and the interquartile range were smaller compared to the experts: 50 (versus 125) and 170 (versus 575), respectively. Response distributions of laypeople were similar to those of the experts and military officers in the case of the airbase. However, consistent with the lay respondents' lack of sensitivity for the different military values of the two targets (presented above), their response distribution in the case of the strategic target was nearly identical to their response distribution in the case of the non-strategic target and thus reflected relatively lower median response and dispersion.[42]

Beyond these descriptive results, it is evident that the level of convergence of experts' responses is lower (that is, more dispersed) compared with lay respondents in considering both military targets. These results are statistically significant in the case of strategic target (HQ) and marginally significant in the case of the tactical target

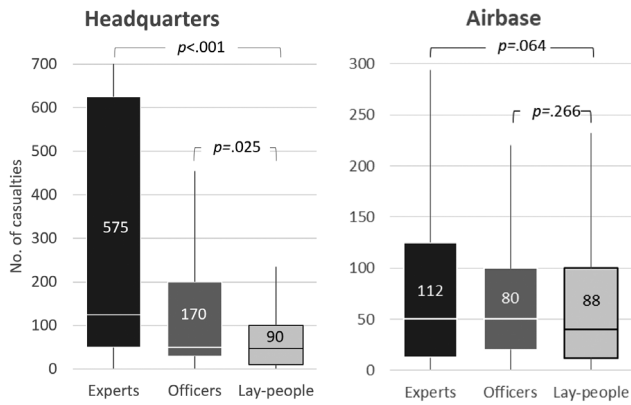


Figure 4: *Convergence measures of experts, officers, and lay respondents' numerical responses. The y-axes present the maximum number of civilian casualties that may be risked for the respective target – 'headquarters' (left) or 'airbase' (right). Box plots represent the interquartile range (25th to 75th percentiles), median values (dark horizontal line) and whiskers for each of the three samples.*[43] *Significance levels of differences in convergence are based on exact Kolmogorov-Smirnov tests.*

41 Excluding experts who did not consider themselves qualified to be members of an official inquiry from the analysis yielded nearly identical distribution results.

42 Notably, only a minority of the lay respondents opted for the option to give a numerical response (and thus are included here), while most of them chose the extreme options ('zero casualties' or 'almost any number').

43 Denoting 1.5 interquartile ranges. See J.W. Tukey, *Exploratory Data Analysis* (1977). In the headquarters plot, the top whisker for experts is not fully presented to facilitate a convenient y-scale.

(AB), based on Kolmogorov-Smirnov tests of equality of distributions (K – S = 0.208, $p < 0.001$; K – S = 0.185, $p = 0.064$, respectively). The convergence levels of officers' responses were lower and statistically significant compared to lay respondents in considering the strategic target (K – S = 0.179, $p = 0.025$) and slightly higher than lay responders, but statistically insignificant in the case of the tactical target (K – S = 0.125, $p = 0.256$). These results provide no support for the hypothesis that judgment convergence among experts and officers is higher compared to lay respondents (H1a). Furthermore, the results provide no support for the hypothesis that interquartile ranges among academic experts and military officers do not exceed the size of their respective median responses (H1b).

To assess whether the different convergence levels across the samples may be accounted for by gender or education level, we utilized exact Kolmogorov-Smirnov tests to compare the distribution of responses among laypeople across gender and education level. No significant differences were found for both targets (HQ and AB), suggesting that gender and education level do not account for the differences in judgment convergence across the samples.

To summarize the results so far, experts and officers were much less likely than lay respondents to choose extreme options ('zero casualties' or 'almost any number') when seeking the proportionate response, suggesting a higher level of understanding of proportionality. Experts and officers also exhibited sensitivity to variation in military value when applying the proportionality principle. However, none of the three groups demonstrated a high level of judgment convergence in applying the proportionality principle. Academic experts and military officers did not present a higher level of judgment convergence compared with lay respondents, and, in all of the analyses, the interquartile ranges were larger than the respective median response, even among those expected to possess the highest level of expertise.

Both academic experts and military officers were expected to share a mutual understanding of the proportionality principle within their respective epistemic communities.[44] Yet, in view of the apparent judgment heterogeneity within the two groups, we further sought to assess whether intra-group differences account for this lack of convergence. For this purpose, we used both professional and cultural criteria. As noted above, the sample of experts encompasses several academic disciplines such as law, moral philosophy, political science and history, with most of the experts (79.6 per cent) from the first two fields. We found no significant difference between the distribution of responses of experts from these academic disciplines (law versus non-law, moral philosophy versus non-philosophy). For the military officers' group, we found a significant difference in the distribution of responses of officers with and without combat experience, but only in considering the strategic target (HQ) (these results are reported in Table S8 in the Appendix). Our cultural criteria utilized the national identity of both expert groups. The largest national group of academic experts was American (50.6 per cent); similarly, the officers were divided between American (52.6

[44] Haas, 'Introduction: Epistemic Communities and International Policy Coordination', 46(1) *International Organization* (1992) 1.

per cent) and Israeli (47.4 per cent). The findings, shown in Figure 5, depict consistent and sizable differences in judgment convergence of American and non-American academic experts and military officers (non-US experts include academic experts from over 10 countries; all non-US officers are Israeli [IL]).

Judgment convergence among American experts and officers is clearly lower (than among their non-US counterparts), leading to the apparent conclusion that the judgments of US experts and military officers are less reliable. However, a closer look at the differences between US and non-US experts and officers also suggests that the median permissible number of civilian casualties of American academic experts and military officers is consistently larger compared to their non-American counterparts. This pattern points to the possibility that the level of convergence decreases as larger

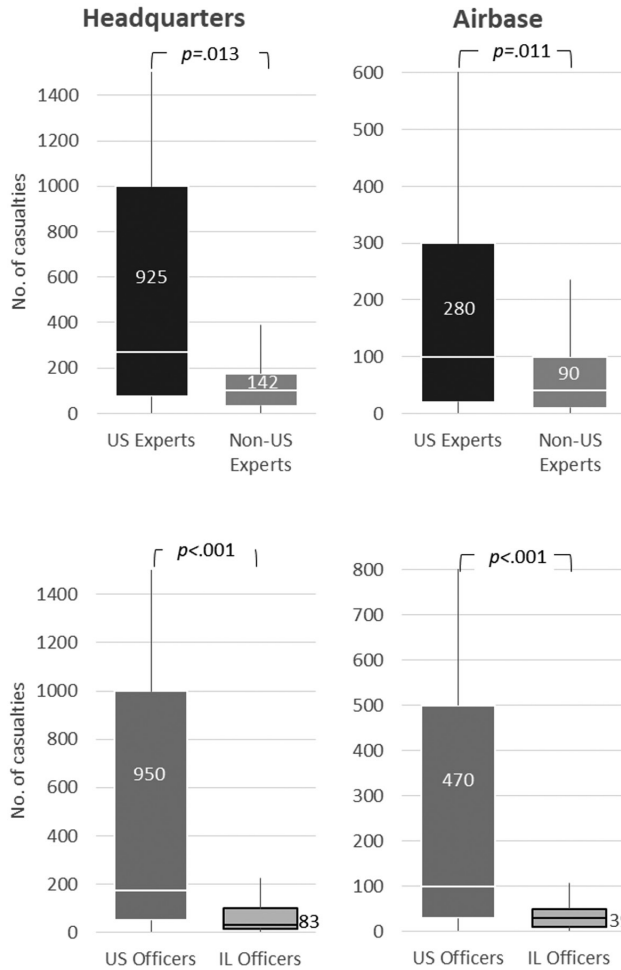


Figure 5: *Convergence measures (interquartile ranges) of experts' and officers' numerical responses for the two military targets, across US and non-US experts/officers. Significance levels of differences in convergence are based on exact Kolmogorov-Smirnov tests.*

numbers are considered, which is consistent with previous findings that valuations of human lives are subject to the Weber-Fechner law.[45] This cognitive principle suggests that people's sensitivity to quantities diminishes as they evaluate increasingly larger values.[46] Therefore, we posit that the cultural difference found between American and non-American experts and officers accounts for the differences in their median permissible number of civilian casualties; given the Weber-Fechner law, the difference in median response resulted in different convergence levels.

Figure 6 demonstrates the application of the Weber-Fechner law to our results, by presenting the relationship between the (logged) median and (logged) interquartile range for the two targets in each group. The dots follow a roughly linear increasing trend, suggesting that the level of convergence decreases (increasing interquartile ranges) as the median response increases. The dashed line specifies the points at which the interquartile ranges are equal to the median response. As clearly shown, all of the estimates are above this line, indicating that all four groups of experts and military officers fail to satisfy this modest requirement regarding both the evaluated targets, providing no support for Hypothesis H1b. We consider the implications of these results in the discussion later in this article. In the following subpart, we analyse the various respondents' susceptibility to biases.

E Judgment Robustness

In this subpart, we report the effects of the three potential biases tested in the experiment – anchor, temporal perspective and task order – for the three respondent groups. Given the non-linear distribution of the numerical responses and the existence of outliers, estimating treatment effects relies on rank regressions, which are robust to a wide range of distributional assumptions and outliers (the results are presented graphically in Figure 7, and tabulated results are provided in Table S9 in the Appendix).[47] All models control for the target size (a dummy variable indicating HQ). Coefficients represent hazard ratios – that is, the ratio between the hazard rate under a marginal increase in the covariate and the base hazard rate. Thus, coefficients smaller than one indicate a decreasing effect on the hazard rate, which, in our data, indicates a larger number of permitted civilian casualties (C_{LT}^{max}); coefficients larger than one designate an increasing effect on the hazard rate and, thus, a smaller number of permissible civilian casualties. The analyses account for the multilevel structure of the data (as each respondent provided two observations) by applying a rank model (Cox proportional hazard) with a shared frailty factor.[48] The statistical assumption of proportionality is supported for the four independent variables across the four respondent groups, based on Grambsch and Therneau tests.[49]

[45] Dickert, *supra* note 3.

[46] Dehaene, *supra* note 19.

[47] Cuzick and Jack, 'Rank Regression', in *Encyclopedia of Biostatistics* (2005).

[48] A frailty factor is a subject-specific random effect, which accounts for the fact that observations are nested within respondents.

[49] Grambsch and Therneau, 'Proportional Hazards Tests and Diagnostics Based on Weighted Residuals', 81(3) *Biometrika* (1994) 515.

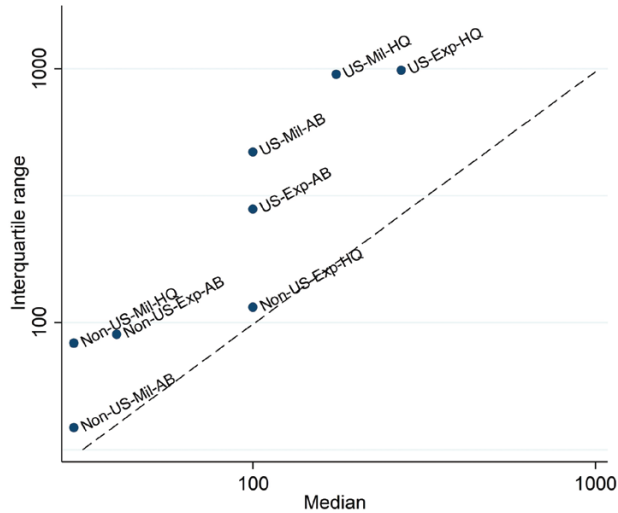


Figure 6: *The relationship between interquartile ranges and median judgments (with logarithmic scales). The dashed line indicates a ratio of one between the two values, with collective judgments above the line indicating interquartile ranges that are larger than the median.*

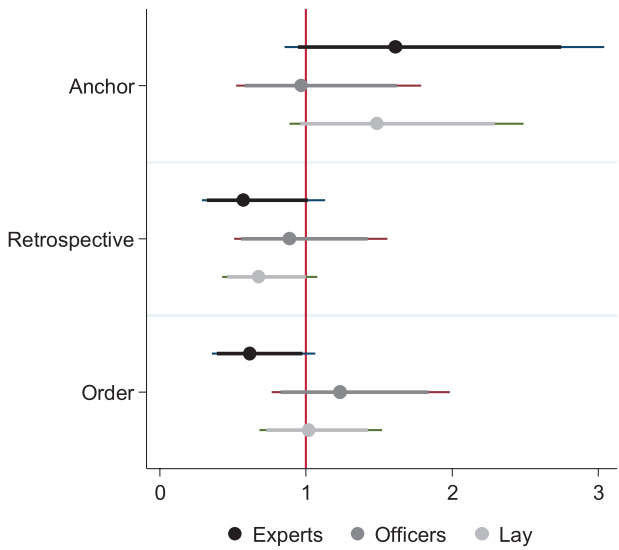


Figure 7: *The effects of anchor, temporal perspective and target order on numerical responses. Point estimates are presented with 95% (thin) and 90% (thick) confidence intervals.*

As evident from Figure 7, exposure to an anchor (30 casualties) seems to have decreased the permissible number of casualties suggested by experts and lay respondents, but these effects are statistically insignificant ($p = 0.141$ and $p = 0.133$, respectively). Retrospective evaluations of academic experts and lay respondents

tended to permit more civilian casualties at marginal levels of significance ($p = 0.108$ and $p = 0.100$, respectively). Experts were affected by the task order. Their permissible maximum number of collateral civilian casualties was higher when the HQ operation was presented first, suggesting that the target to which the experts were exposed first significantly influenced their decision regarding the subsequent target ($p = 0.082$). Notably, military officers were not significantly affected by any of the three treatments. Very similar results were obtained by estimating the effects with random effect linear regressions with logged response as the dependent variable (see Table S10 in the Appendix). Notable differences between the two regression types are that the negative effect of the anchor on lay respondents reaches statistical significance in this analysis and that the order effect on experts is not statistically significant ($p = 0.160$).

To summarize, we found academic experts and lay respondents to be susceptible to at least one of the three irrelevant factors, but military officers were not. An order effect was found for the experts and the temporal perspective affected the decisions of lay respondents. Military officers demonstrated robustness to irrelevant factors. Lastly, American identity was associated with higher numbers of permissible civilian casualties, in line with our convergence analyses above, and the type of target (HQ/AB) affected the decisions of experts and officers but not of lay respondents, which was in line with our sensitivity analyses.

To assess whether the different treatment effects across the samples may be accounted for by gender or education level, we ran two additional regression analyses using the lay respondents sample (Models 4 and 5 in Tables S9 and S10 in the Appendix). These models include interaction terms for each treatment and gender (Model 4) and academic degree (college degree or higher) (Model 5). No significant interaction was found, suggesting that the treatment effects are not different across gender and education levels. In the random effect linear models (Table S10), a significant interaction was found for the order effect and gender, suggesting that this effect differs between women and men, but the effect for each gender is not statistically significant. Furthermore, it appears that the effect of temporal perspective is mostly driven by educated respondents, although no such effect was found for experts and officers.

6 Discussion

The aim of this study was to offer an empirical assessment of the reliability of *in bello* proportionality judgments. Our point of departure was that, in the absence of an empirical criterion for the truth-value of such judgments, their reliability provides the next best criterion as it is a necessary condition for judgment validity. Our analysis utilizes three measures of reliability to assess the capacity of experts to reliably implement the proportionality principle.

Table 2 succinctly summarizes our findings. The most important measure of reliability is inter-expert judgment convergence. Our results show that academic experts

and military officers fail to reach reasonable convergence regarding the maximum number of civilian lives that may be risked in the scenarios presented to them. The levels of judgment convergence of all three respondent groups consistently fell short of what reasonably can be expected, even when accommodating for a known psychophysical limitation – the Weber-Fechner law and consequent 'psychic numbing'.[50]

The results are less clear-cut when considering sensitivity and robustness. Both types of experts demonstrated reasonable sensitivity to variations in military value, unlike lay respondents. This successful performance can be attributed to the experts' understanding of the proportionality principle and conforms to previous findings,[51] yet we cannot discard the possibility that education level may account for the differences in sensitivity between lay respondents and experts and officers. Lastly, in line with the studies showing that normative experts are susceptible to cognitive biases,[52] our study also finds that academic experts are susceptible to biases to a similar degree as lay respondents. However, we find no indication for the susceptibility of military officers to these biases.

It should be noted that the professional knowledge of academic experts and military officers does give them some advantage in judgments of *in bello* proportionality as it reduces their likelihood of choosing extreme options (see Figure 2). These findings suggest that experts share a theory of proportionality (to the extent that the 'correct answer will not be found in extreme responses'); however, analysis of their judgment convergence indicates that they lack the ability to reliably apply this theory. Real world dilemmas concerning *in bello* proportionality usually reside in between the very extreme (and clearly disproportionate) options, and it is in this crucial range of alternatives that the response pattern of experts appears highly dispersed, indicating their unreliability.

Two unexpected findings beg further discussion. First, despite the wide acceptance of the proportionality principle, we found consistent cultural differences in its application by both academic experts and military officers. The median response of American experts and military officers was higher, and their level of judgment convergence was lower (that is, more dispersed) compared to their respective non-American counterparts. Note, however, that even within-culture analyses of judgment convergence failed to yield sufficiently reliable levels of inter-expert/officer convergence. Second, our findings point to a consistent negative relationship between the median proportionality judgment of a group and its level of judgment convergence. If experts or

Table 2: Summary results

Reliability criteria	Academic experts	Military officers	Lay respondents
Convergence	×	×	×
Sensitivity	√	√	×
Robustness	×	√	×

[50] Dickert, *supra* note 3.

[51] Sulitzeanu-Kenan, Kremnitzer and Alon, *supra* note 20.

[52] Schwitzgebel and Cushman, 'Philosophers' Biased Judgments Persist Despite Training, Expertise and Reflection', 141 *Cognition* (2015) 127.

military officers had a reliable method of applying the proportionality principle, one would expect to find similar levels of (high) convergence, regardless of the median judgment. However, relatively lower dispersion was observed only for group judgments featuring a relatively low median response. These results suggest that 'psychic numbing' in valuations of human lives also applies to experts' decision-making.[53]

The main conclusion from these findings is not simply that concepts often have vague boundary zones in which their application is unclear but also, more disturbingly, that some notions may prove to be of no avail precisely where they are most needed – as in the present context – such as in attacks on military targets that are expected to lead to a non-negligible number of civilian casualties. What then do these results entail regarding the idea of proportionality in warfare? One possible inference is that, in cases where no reasonable convergence can be achieved among evaluators, there is simply no truth of the matter. However, we see no reason to commit ourselves to this conclusion, and our empirical evidence does not, by itself, demonstrate that there is no right answer to be discovered.

Assuming then, that there is a true answer regarding the maximum number of collateral civilian casualties that is permissible in cases like the headquarters example, our study shows that given their epistemic imperfections, humans fail to reveal it. While they have strong views about the permissibility of attacks involving very few or no civilian casualties, and about the impermissibility of attacks involving extremely large numbers of enemy casualties, when it comes to cases that lie between these extremes – which constitute most of the cases in real world wars – they appear to be simply guessing. Since reasonable convergence of opinion in a field of knowledge is part of what defines expertise, the endemic disagreement between experts about the right proportion between the value of attacking military targets and the harm that will ensue to enemy civilians, we do not find support for the existence of expertise in this field. This disconcerting conclusion should be slightly moderated, as our findings show that expertise has some informative contribution by reducing the likelihood of choosing extreme options.[54] However, we expect experts to provide us with more informative advice than that.

The model of proportionality points to two mutually non-exclusive sources of disagreement among experts, which may account for the resulting judgment convergence. The first relates to the military value of the target (V_{LT}) and the other pertains to the coefficient α for converting military value to a normative maximal limit of collateral harm (C_{LT}^{max}). The former reflects inconsistent approaches for assessing the military value of targets, whereas the latter manifests a normative disagreement on the application of the proportionality principle itself.[55] This study provides empirical evidence for the low level of convergence of experts' proportionality judgments, but our results cannot disentangle them from among those sources of disagreement. Identifying the role of each of these sources of disagreement in determining the

[53] Dickert, *supra* note 3.

[54] Recall that among laypeople almost 30 per cent opted for the 'almost any number' response in the headquarters scenario.

[55] We are grateful to one of the anonymous reviewers for making this distinction.

consequent judgment carries normative and practical implications and presents potentially important goals and challenges for future research.

One might argue that the weak convergence among experts does not testify to their lack of expertise but, rather, to the result of the limited information provided in the vignettes. If only the experts had been given more information – about how exactly the proposed attack would shorten the war, about how many civilian lives (in country A) would exactly be saved by it and so on – they would have fared much better in terms of judgment convergence. However, in the typical fog of war, armies rarely have access to such detailed information. The results of this research carry important implications for the ethics and laws of armed conflict. First, the apparent inability of experts – both academics and military officers – to consistently determine the application of *in bello* proportionality implies that the protection of civilians during warfare is unreliable, even when warring parties attempt to abide by the proportionality principle. The protection of civilians may be insufficient in some cases, and, in others, the risk to civilians may overly restrict legitimate military plans. Second, the apparent inability of experts to reliably determine the correct application of *in bello* proportionality casts serious doubt over such *ex-post* judgments. Given that the distribution of experts' proportionality judgments appears as an aggregation of mere guesses, *ex-post* approvals as well as the condemnations of military actions are often unwarranted. The fact that judgments in this field are such easy prey for biases, especially political ones,[56] provides further support for scepticism about the ability to get the proportionality calculation right in any specific case. Lastly, in this respect, while many studies in recent years have been devoted to the difficult question of carefully accounting for civilian losses,[57] the present study points to a more fundamental problem of applying the proportionality principle.

The low reliability of *in bello* proportionality judgments may also have adverse implications for the level of states' compliance with the laws of war. Such compliance is determined by publicly accepted and legally binding agreements that create incentives for the parties to enforce those agreements through reciprocity.[58] Reciprocity is premised on the ability of states to mutually assess compliance levels. However, such assessments are often made under levels of noise. For example, it is often unclear whether violations by individual combatants are a product of state policy or not.[59] The level of noise hinders the effectiveness of reciprocity and consequently dampens compliance level. Yet James Morrow finds lower compliance in aerial bombing – a surprising finding given that such actions result from relatively centralized decisions. We suggest that the unreliability of proportionality judgments – among attackers as well as observers – likely contributes to this level of noise in estimating the compliance of countries to the laws of war and thus hinders reciprocity as a mechanism of upholding compliance.

[56] Sulitzeanu-Kenan, Kremnitzer and Alon, *supra* note 20.

[57] For a recent review, see Jewell, Spagat and Jewell, *supra* note 9.

[58] Morrow, 'When Do States Follow the Laws of War?', 101(3) *APSR* (2007) 559.

[59] Morrow, 'The Institutional Features of the Prisoners of War Treaties', 55(4) *International Organization* (2001) 971.

One limitation of this study stems from the fact that we compare groups – academic experts, officers and laypeople – to which respondents were not randomly assigned. One cannot infer that the differences observed were caused by group membership rather than, for example by self-selection. However, our interest is in the comparative descriptive reliability of different groups of respondents that are particularly relevant to *in bello* proportionality decisions. The causal reasons for their level of judgment reliability would require additional research. The reader might worry that, if the proportionality principle plays such a limited role in restricting the use of force during war, then civilians will be stripped of their immunity and the door will be opened to total war. We wish to say three things in response. First, our study deals only with collateral harm to civilians, and our results do not undermine the validity or the practicality of the blanket prohibition of intentional attacks on civilians. Second, due proportion is not the only restriction on the unintentional, yet foreseeable, harming of civilians. International humanitarian law includes another restriction – namely, that the attackers select the least harmful measure and make a sincere effort to minimize harm to civilians:

> Simply not to intend the death of civilians is too easy. … What we look for in such cases is some sign of a positive commitment to save civilian lives. … War necessarily places civilians in danger; that is another aspect of its hellishness. We can only ask soldiers to minimize the dangers they impose.[60]

Third, given the limited ability of the proportionality criterion to restrict collateral attacks on civilians, we may reconsider the proposal developed by W. Hays Parks to the effect that the protection of civilians in wartime should be the concern of their own countries, not – or not only – of their enemies.[61] In practice, this means that countries ought not to locate military headquarters or facilities in close proximity to residential areas and definitely ought not to launch military attacks from within such areas. For the sake of the present discussion, one need not accept Parks' entire proposal, but just the moderate idea that the protection of civilians from collateral harm should not be the sole responsibility of the side fighting against them. This would be enough to soften the worry that the unreliability of the proportionality principle leaves civilians without a reasonable protection from collateral harm.

Finally, although this study was about proportionality in warfare, the method it utilized may be applicable to other normative notions in an attempt to assess the extent to which they can be reliably applied. We hope to implement it in the future in other domains, and we encourage others to do so as well. It is our contention that the use of norms, regarding which reasonable convergence cannot be achieved, should be minimized. If this suggestion is endorsed, it may lead to a normative discourse that is both fairer and more useful as a guide to individual behaviour and public policy.

[60] Walzer, *supra* note 1, at 155–156; see also Additional Protocol I, *supra* note 4: '[W]hen a choice is possible between several military objectives for obtaining a similar military advantage, the objective to be selected shall be that the attack on which may be expected to cause the least danger to civilian lives and to civilian objects.'

[61] Parks, 'Air War and the Law of War', 32(1) *Air Force Law Review* (1990) 146.