The Impact of Panel Size on the Reliability of Criminal Verdicts in a Military Justice Context

Isaac Kennen [*]

Christopher Stein [†]

Michelle Drouin [‡]

Kenneth Bordens [§]

Dan Coroian [**]

I. Introduction

The American trial by jury has ancient roots—to an English yew tree outside of London overlooking the Runnymede wetlands and the River Thames.[1] About 800 years ago, under the gaze of the Ankerwycke, a group of rebellious barons managed to wrest the right to a jury trial from the grip

[*] *B.A., Louisiana Tech University, Ruston (2003); J.D., University of Colorado at Boulder (2006). Adjunct Professor of Law at Western New England University School of Law, Staff Attorney at Community Legal Aid in Springfield, Massachusetts, and retired United States Air Force judge advocate. As a judge advocate, he served tours as a prosecutor, trial defense counsel, appellate defense counsel, and as legal advisor on the law of armed conflict to commanders conducting combat and strategic deterrence operations worldwide.*

[†] *B.A., University of California, Los Angeles (2005); J.D., William S. Boyd School of Law, University of Nevada, Las Vegas (2008); LL.M., The United States Army Judge Advocate General's Legal Center & School (2018). Active-duty United States Air Force judge advocate. He has served a variety of assignments both stateside and overseas. The views expressed are his own and do not necessarily represent the views of the U.S. Air Force or Department of Defense.*

[‡] *D.Phil., Experimental Psychology, University of Oxford, England (2004); Professor, Purdue University, Fort Wayne.*

[§] *Ph.D, Social Psychology, University of Toledo (1979); Professor, Purdue University, Fort Wayne.*

[**] *Ph.D, Mathematics, University of Iowa (1997); Professor, Purdue University, Fort Wayne.*

[1] *Runnymede and Ankerwycke*, NAT'L TRUST, https://www.nationaltrust.org.uk/visit/surrey/runnymede-and-ankerwycke (last visited May 2, 2024).

of their king.[2] That day, the rebels "gathered with a multitude of most famous knights, armed well at all points."[3] In turn, "[King] John was charming in public [but] behind the scenes he 'gnashed his teeth, rolled his eyes, grabbed sticks and straws and gnawed them like a madman.'"[4] Under the threat of violence, the insurgents forced the Crown into a peace accord, which we now call the *Magna Carta*.[5] As part of that agreement, the King promised to allow for jury trials.[6] Specifically, he swore that "no free man shall be seized or imprisoned, or stripped of his rights or possessions, or outlawed or exiled, or deprived of his standing in any way, nor will we proceed with force against him, or send others to do so, except by the lawful judgment of his equals or by the law of the land."[7] Upon that heritage, American leaders have declared jury trials "the best method of trial that is possible,"[8] "the only anchor, yet imagined by man, by which a government can be held to the principles of its constitution,"[9] "heaven-taught," [10] and "our birth right."[11]

Furthermore, to the American mind, a jury trial means being tried by "a jury of twelve men all concurring in the same judgment."[12] That cultural understanding has fueled the creation of movies like *12 Angry Men*, where Henry Fonda played a lone hold-out juror who stood between the government and the citizen it accused, and who eventually persuaded his fellow jurors to acquit an innocent man.[13] While the U.S. Constitution allows the states to reduce the number of jurors downward from twelve,

[2] DAN JONES, MAGNA CARTA: THE BIRTH OF LIBERTY, 134-35 (2015).

[3] RADULPHI DE COGGESHALL CHRONICON ANGLICANUM 172 (Joseph Stevenson ed., trans. 1875).

[4] MATTHAEI PARISIENSIS, MONACHI SANCTI ALBANI: CHRONICA MAJORA 611 (Henry Richards Luard ed., trans. 1872-1873).

[5] MAGNA CARTA, *supra* note 2, at 138-40.

[6] Magna Carta, 9 Hen. 3 (1215) (Eng.).

[7] *Id.* c. 39.

[8] Thompson v. Utah, 170 U.S. 343, 349-50 (1898) (quoting 1 Hale's P.C. 33).

[9] Letter from Thomas Jefferson to Thomas Paine (July 11, 1789), *in* 15 THE PAPERS OF THOMAS JEFFERSON 266, 270 (Julian P. Boyd ed. 1958).

[10] Claudio v. State, 585 A.2d 1278, 1292 (Del. 1991).

[11] *Id.*

[12] *Thompson*, 170 U.S. at 349-50.

[13] *See* 12 ANGRY MEN (Orion-Nova Production 1957).

few have done so.[14] At any rate, the authority to do so is limited: state juries with fewer than six members are unconstitutional[15] because empirical research in the civilian world has shown that juries so small tend to be inconsistent and unreliable.[16] Similarly, the Federal Constitution requires unanimous verdicts in civilian criminal trials at both the state and Federal level.[17] In explaining the importance of the requirement for unanimity, Associate Justice Brett Kavanaugh opined:

> [N]on-unanimous juries can silence the voices and negate the votes of [B]lack jurors, especially in cases with [B]lack defendants or [B]lack victims, and only one or two [B]lack jurors. . . . That reality—and the resulting perception of unfairness and racial bias—can undermine confidence in and respect for the criminal justice system.[18]

Despite those historical, cultural, and legal imperatives that implore the use of full-size, unanimous juries, not all Americans have received their inheritance. A Federal military conviction carries the same consequences as a civilian one,[19] but rather than being tried by a random selection of their peers, military court-martial panels are made up of a collection of the accused's superiors who are hand-picked by the officer who ordered the trial to proceed.[20] In some cases, military panels may have as few as four members.[21] Further, in most cases, military panels are not required to be unanimous to convict the accused—a mere three-fourths majority vote will suffice.[22] The military's Service courts have resisted arguments that these practices are unconstitutional.[23] The highest military

[14] Nate Raymond, *U.S. Supreme Court's Gorsuch Says Justices Should Require 12-Person Juries*, REUTERS (Nov. 7, 2022), https://www.reuters.com/legal/government/us-supreme-courts-gorsuch-says-justices-should-require-12-person-juries-2022-11-07 (stating that only six states allow for fewer than twelve jurors in felony cases: Arizona, Connecticut, Florida, Indiana, Massachusetts, and Utah).

[15] Burch v. Louisiana, 441 U.S. 130, 134 (1979).

[16] Ballew v. Georgia, 435 U.S. 223, 234-36 (1978).

[17] Ramos v. Louisiana, 140 S. Ct. 1390, 1397 (2020).

[18] *Id.* at 1418.

[19] *See* Major Jeff Walker, *The Practical Consequences of a Court-Martial Conviction*, ARMY LAW., Dec. 2001, at 1.

[20] 10 U.S.C. § 825(e).

[21] 10 U.S.C. § 816(c).

[22] 10 U.S.C. § 852(a)(3).

[23] *See* United States v. Daniel, 73 M.J. 473 (A.F. Ct. Crim. App. 2014).

appellate court, the U.S. Court of Appeals for the Armed Forces (CAAF), which is a Federal court of record staffed by civilian judges appointed by the President and confirmed by the Senate for fifteen-year terms,[24] has refused to reverse those lower Service court decisions,[25] and the Supreme Court has refused to intervene.[26]

The research conducted thus far concerning the reliability of verdicts reached by small and nonunanimous juries has uniformly cast that question in the context of civilian mock trials—with jurors being instructed on civilian standards of law, civilian criminal procedure, civilian cultural references, and civilian fact patterns. This paper details the recent efforts of a multi-disciplinary team of two Air Force military lawyers (judge advocates), two psychologists, and an applied mathematician to explore whether small panels suffer the same deficiencies when the mock trial they participate in is presented as being a court-martial—with the panel members instructed on military standards of law, using military lexicon and rank designations, with military cultural references, and military fact patterns. More specifically, the team developed an experimental paradigm to contrast the deliberation outcomes of an eight-member panel and a six-member panel for a mock sexual assault court-martial case.

The findings from this research are particularly important now because, of late, Congress has shown a willingness to reassess its composition of courts-martial. In December 2016, Congress enacted changes to the controlling body of law: the Uniform Code of Military Justice (UCMJ) (Title 10, Chapter 46, U.S. Code).[27] Those changes took effect in January 2019 and raised the number of members required to serve on a general court-martial panel from five to eight, and the number required for a special court-martial from three to four.[28] These incremental changes, while a step in the right direction, still have not brought the

[24] 10 U.S.C. § 942.

[25] *See, e.g.*, United States v. Daniel, 76 M.J. 473 (C.A.A.F. 2014). *But see also* United States v. Strong, 83 M.J. 392 (C.A.A.F 2023) (pending decision, but petition for review was recently granted, and briefs ordered, on the issue of "whether [a]ppellant was deprived of her constitutional right to a unanimous verdict").

[26] *See*, *e.g.*, Daniel v. United States, 574 U.S. 1079 (2015) (cert. denied).

[27] National Defense Authorization Act for Fiscal Year 2017, Pub. L. No. 114-328, sec. 5161, § 816, 130 Stat. 2000, 2897 (2016).

[28] *Id.*

military justice system into alignment with civilian practice. A panel of four members is still well below what the Constitution requires for civilian trials. Even eight members falls short of the historical and cultural standard Americans traditionally expect of criminal trials of twelve members. Moreover, the eight members required of a general court-martial can be reduced to six if issues during trial necessitate the release of panel members.[29]

Further, although Congress increased the quorum required for a conviction of most offenses from a two-thirds majority vote to a three-fourths concurrence, that is still well short of the unanimity required for a conviction of a serious offense in American civilian jurisdictions.[30] Congress also chose to preserve the practice of allowing the officer who ordered the court-martial to proceed to also select the panel members.[31] Despite these differences from the civilian criminal justice system, the enactment of this legislation shows that Congress is trying to make courts-martial more closely match their civilian counterparts, while also making them more consistent and reliable as fact-finding entities.[32] This study offers valuable data to inform that effort.

II. Previous Research

Since 1967, as many as seventeen civilian empirical studies concerning the difference between six- and twelve-member juries have occurred.[33] The takeaway from those efforts is summarized as follows: "In short, there still are no ideal studies of jury size effects. All of them are compromises of one kind or another."[34] In 1997, the California legislature mandated a study that would have used rigorous methodology "because of frustrations resulting from equivocal findings generated by flawed

[29] 10 U.S.C. § 829(d)(1)(B).
[30] *See* National Defense Authorization Act for Fiscal Year 2017, Pub. L. No. 114-328, sec. 5235, § 852, 130 Stat. 2000, 2916 (2016).
[31] *See id.* sec. 5182, § 825, 130 Stat. at 2900.
[32] *See* Fred L. Borch III, *Military Justice in the Army: The Evolution of Courts-Martial from the Revolutionary War Era to the Twenty-First Century*, ARMY LAW., no. 2, 2023, at 35.
[33] Michael J. Saks & Mollie W. Marti, *A Meta-Analysis of the Effects of Jury Size*, 21 L.& HUM. BEHAV. 451, 452 (1997).
[34] *Id*. at 454.

studies."[35] That effort failed, however, when a court official forbade employment of the statute and allowed parties who were to receive smaller juries to "opt out of that assignment in favor of a [twelve]-person jury."[36]

Despite these shortcomings, in 1997, Michael Saks and Mollie Marti separately reviewed sixteen studies in existence to that date concerning the effect of jury size and conducted a meta-analysis.[37] The findings they published are a remarkable and concise compendium of the body of research relating to this topic. Their work marshals a wide variety of data regarding each study, including factors such as sample size, the pool from which study participants were acquired, whether the cases being studied were civil or criminal in nature, whether the study was conducted in a courtroom or in a laboratory, and the medium used to present the trial to the study participants.[38]

Saks and Marti then assigned each study a weighted value.[39] For example, "studies employing stimulus cases that were so extreme that all verdicts were the same, and which therefore were inherently incapable of detecting any effects of jury size on verdicts, received a weight of zero."[40] Only two studies received a weighting of zero.[41] Six studies were rated as either eight or nine, four received rating between four and seven, and four received a rating of one.[42] The studies that received a one rating constituted "uncontrolled correlational studies, which allowed the parties to self-select cases into jury size conditions, thereby tending to put more complex and higher stakes cases in front of larger juries."[43]

Ultimately, Saks and Marti concluded that the research showed significant differences between six- and twelve-member juries.[44] First, the "largest effect of any of the variables studied" is that "[twelve]-person juries are more likely than [six]-person juries to contain at least one

[35] *Id.*
[36] *Id.*
[37] *See id.* at 453.
[38] *See id.*
[39] *See id.* at 454.
[40] *Id.*
[41] *Id.* at 453.
[42] *Id.*
[43] *Id.* at 454.
[44] *Id.* at 457.

member of whatever minority group is under consideration."[45] Reduced jury size decreases the opportunity of minority "representation from about 63-64 [percent] to about 36-37 [percent]."[46]

Only eleven of the studies Saks and Marti reviewed captured data on the length of deliberations, and only two of those provided statistics necessary to determine whether jury size significantly affected that factor.[47] However, much data regarding factors that favor more thorough deliberations was captured. For example, Saks and Marti found that "[t]rial testimony was discussed more accurately in the deliberations of larger juries than in the deliberations of smaller juries."[48] Further, members of larger juries "remembered more of the facts in evidence, measured by a post-deliberation test of their recall."[49]

Saks and Marti's work dovetails nicely with the findings of the U.S. Supreme Court on the subject. The Court, reviewing many of the same empirical studies that Saks and Marti relied upon, found that progressively smaller juries are "less likely to foster effective group deliberation" and are prone to "inaccurate fact-finding and incorrect application of the common sense of the community to the facts."[50] Additionally, the Court found that the research shows individual members in smaller panels are "less likely . . . to make critical contributions necessary for the solution of a given problem," and "as juries decrease in size . . . they are less likely to have members who remember each of the important pieces of evidence or argument."[51] Further, according to the research "the smaller the group, the less likely it is to overcome the biases of its members to obtain an accurate result."[52] In contrast, larger panels benefit from "increased motivation and self-criticism."[53]

[45] *Id*.

[46] *Id*.

[47] *Id*.

[48] *Id*. at 459.

[49] *Id*.

[50] Ballew v. Georgia, 435 U.S. 223, 232 (1978).

[51] *Id*. at 233.

[52] *Id*.

[53] *Id*.

The Court held these deficiencies "suggest that the risk of convicting an innocent person . . . rises as the size of the jury diminishes."[54] The Court assessed that "the verdicts of jury deliberation in criminal cases will vary as juries become smaller, and that the variance amounts to an imbalance to the detriment of one side, the defense."[55]

It should be noted that the Court's conclusion that smaller juries impose an imbalance against the defense does not answer the next logical question of whether that imbalance causes incorrect verdicts. Whether such a detriment drives decisional errors in any given case is not easy to determine because the definition of correct is largely subjective. For example, Saks and Marti's study defined "correct" as being the verdict that they thought the public at large would have likely reached had they, collectively, had the opportunity to decide the case.[56] Under that standard, Saks and Marti concluded that "meta-analysis of the [ten] relevant studies of simulated trials [found that jury size had] no significant effects [on the jury's ability to reach the correct verdict]."[57] But that standard necessarily assumes Saks and Marti's assessment of the public's inclinations were accurate. There is no way to test that assumption because it is based on criteria that is non-empirical and non-replicable – the researcher's personal belief as to what the public would have done had it had the chance. Such a definition of correct is unscientific and unhelpful.

Further, even if a non-subjective standard for correctness could have been formulated, it was probably impossible for Saks and Marti to reach a reliable conclusion regarding correctness from the data set they were given. They aggregated data from studies that mixed data from civil and criminal cases, involving different standards of proof, different evidence, and potentially even different community mores.[58] For example, a correct outcome in a civil case may differ significantly from the outcome that would be deemed correct in a criminal trial given higher burdens of proof that are customarily placed on the prosecution in criminal proceedings.

Although the correctness of a verdict may be resistant to scientific measurement, that does not mean that the question of correctness is

[54] *Id.* at 234.
[55] *Id.* at 236.
[56] *See* Saks & Marti, *supra* note 33, at 461.
[57] *Id.* at 461-62.
[58] *See* Saks & Marti, *supra* note 33.

unimportant to the question of the ideal size of a jury or court-martial panel. Rather, the size of a jury has long been thought to be an important factor driving the risk of an incorrect verdict.[59] Specifically, Condorcet's jury theorem, coined in 1785 by the Marquis de Condorcet in *Essay on the Application of Analysis to the Probability of Majority Decisions*,[60] posits that the ideal size of a jury varies in proportion to the relative likelihood of each individual juror reaching the correct vote.[61] For example, if under the circumstances individual jurors are more likely than not to vote correctly, then the more jurors on the court, the better. If, in contrast, each juror is more likely to vote incorrectly, then the ideal number of jurors for society's sake is one. Of course, applying that principle requires an accurate definition of what a correct vote looks like. For reasons explained above, reaching an accurate, scientific, non-subjective definition of correct in all but the most clear-cut of cases is an exceptionally challenging endeavor.

To meet the challenge of defining correctness of a verdict, study participants were presented a mock military justice sexual assault case that, evidentiarily, was designed to be a close call on the questions of consent and whether the accused harbored a reasonable mistake of fact as to consent. Pains were taken to ensure that evidence was presented to the undergraduate participants of the study that could support a finding of either guilty or not guilty. The evidentiary presentation was video recorded and played for each of the participating panels of undergraduate students. The military judge's instructions on the evidence and on the conduct of deliberations were likewise recorded and played for each panel. The goal was to make the mock case a neutral variable so as to test the effect of having panels of varying size. The research team then used statistical modeling to predict, mathematically, the probability of a guilty verdict of

[59] *See* Ballew v. Georgia, 435 U.S. 223, 233-34 (1978) ("[R]ecent empirical data suggest that progressively smaller juries are less likely to foster effective group deliberation. At some point, this decline leads to inaccurate factfinding and incorrect application of the common sense of the community to the facts. Generally, a positive correlation exists between group size and the quality of both group performance and group productivity.").

[60] MARIE-JEAN-ANTOINE-NICOLAS DE CARITAT, MARQUIS DE CONDORCET, ESSAI SUR L'APPLICATION DE L'ANALYSE À LA PROBABILITÉ DES DÈCISIONS RENDUES À LA PLURALITÉ DES VOIX [ESSAY ON THE APPLICATION OF ANALYSIS TO THE PROBABILITY OF MAJORITY DECISIONS] (1785) (Fr.).

[61] Franz Dietrich & Kai Spiekermann, *Jury Theorems*, STAN. ENCYC. OF PHIL. (Nov. 17, 2021), https://plato.stanford.edu/archives/spr2023/entries/jury-theorems.

an eight-member panel as opposed to panels of lesser size. By comparing the actual results of each mock trial to the statistical probability data that such a panel would render a guilty verdict, our study was able to assess the likelihood that the correct verdict would be rendered by a panel of that size.

III. Methods

A. Preliminary Analyses

As a preliminary step in determining how to test the efficacy of an eight-member panel (as compared to panels of lesser size), we examined the statistical probability of guilty verdicts for panel compositions of five members through eight. As shown in Table 1, the largest spread exists between panels of eight and six (a 12 percent difference in the percentage of non-guilty votes needed for acquittal).

Table 1

Number of and Percentage of Members Needed for Two-Thirds Majority and Acquittal

Number of Members	Two-thirds Majority	Number and Percentage of not guilty votes needed for acquittal
5	4	(2) 40%
6	4	(3) 50%
7	5	(3) 43%
8	6	(3) 38%

It was recognized that those statistics were relevant for only a single case, and that probabilistic modeling was needed to determine whether the probability of convictions *over time* is influenced by the fact that the number and percentage of not guilty votes needed to acquit varies depending on the size of the court-martial. Therefore, an applied mathematician developed probabilistic models of conviction based on different jury composition sizes, focusing on the two jury compositions with the greatest spread (i.e., six- and eight-member panels) and in two different conditions (i.e., all votes are possible and at least two members of the panel vote not guilty).

B. Probabilistic Modeling Applied to Panel Size

A basis of this modeling involves a simple probability calculation. The probability of an event is a number between zero and one (including zero and one), that measures the likelihood that the event will occur. It is defined as the number of cases favorable for the event to occur, divided by the total number of cases possible, that is:

$$P = \frac{Number\ of\ favorable\ cases}{Number\ of\ cases\ possible}\,.$$

As an example, the probability of rolling a five on a die is one-sixth because there is only one favorable outcome out of six outcomes possible. However, the probability of an event does not predict the exact outcome; it is only an estimate of what to expect will happen, and it gets more and more accurate in the long run.

A second basis of this calculation involves combinatorial mathematics. The number of groups of *k* objects that could be formed from a total of *n* objects is denoted $\binom{n}{k}$, and it is called *the number of combinations of n objects taken k at a time* (often read as "*n* choose *k*"). It can be calculated using the formula:

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots\ (n-k+1)}{1\cdot 2\cdot 3\cdots (k-1)k}\,.$$

For example, if a committee of three is to be formed from a group of twenty people, there are $\binom{20}{3} = \frac{20\cdot 19\cdot 18}{1\cdot 2\cdot 3} = 1140$ possible ways of choosing the committee.

When we apply these formulae to the six-member panel,[62] we need at least (2/3) * 6 = 4 guilty votes for conviction or at most two not guilty votes. This can occur in the following scenarios (in the diagrams below, the panel members' votes are represented by either a *g* or an *ng*):

Member:	**1**	**2**	**3**	**4**	**5**	**6**
(i) All members vote guilty:	g	g	g	g	g	g

which represents one of the *favorable cases* for conviction (see definition of probability).

(ii) All but one vote guilty:	ng	g	g	g	g	g
or	g	ng	g	g	g	g
or	g	g	ng	g	g	g
or	g	g	g	ng	g	g
or	g	g	g	g	ng	g
or	g	g	g	g	g	ng

Thus, there are six more *favorable cases* for conviction, when exactly one panel member votes not guilty. This number could have also been found by applying the combinations formula above for finding the number of *k* = 1 person groups that can be formed out of an *n*=6 persons: $\binom{6}{1} = \frac{6}{1} = 6$.

[62] In this discussion, we assume that for every member of the panel, the probability of a guilty or a non-guilty vote is the same; however, in practice, this might not be true.

Member:	**1**	**2**	**3**	**4**	**5**	**6**
(iii) All but two vote guilty:	ng	ng	g	g	g	g
or	ng	g	ng	g	g	g
or	ng	g	g	ng	g	g
.	.	.	.	.	.	.
or	g	g	g	g	ng	ng

Instead of enumerating all the possibilities, the combinations formula is applied, which gives a total of $\binom{6}{1} = \frac{6\cdot 5}{1\cdot 2} = 15$ possible scenarios in which exactly two panel members vote not guilty. In total, there are 1 + 6 + 15 = 22 scenarios possible for conviction, in which at least four members vote guilty. These represent the *favorable cases* in the definition of probability above. The total number of *cases possible* is $2^6 = 64$, because each one of the six panel members has two choices. Thus, the probability of a guilty verdict is:

$$P = \frac{22}{64} = 0.34375.$$

This means that, in the long run, we can expect an approximate rate of conviction of 34.375 percent.[63]

If we assume that at least two panel members always vote not guilty, then the number of *favorable* cases for a conviction drops to fifteen (as cases (i) and (ii) cannot happen anymore), and the number of *possible* cases also decreases to 64 - 1 - 6 = 57, for the same reason. Therefore, the probability of a guilty verdict under this restriction is:

$$P = \frac{15}{57} = 0.26316,$$

[63] In practice, we should expect the actual conviction rate to start getting close to this value only after a large number of trials.

which means that, in the long run, the approximate rate of conviction is expected to be 26.316 percent, when at least two of the panel members vote not guilty.

Meanwhile, for an eight-member panel, because (2/3) * 8 = 5.33, we need at least six guilty votes for conviction, or at most two not guilty votes. This can occur in the following scenarios:

Member:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
(i) All vote guilty:								
	g	g	g	g	g	g	g	g

which represents 1 of the *favorable cases* for conviction in the definition of probability.

Member:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
(ii) All but one vote guilty:								
	ng	g	g	g	g	g	g	g
or	g	ng	g	g	g	g	g	g
or	g	g	ng	g	g	g	g	g
or	g	g	g	ng	g	g	g	g
or	g	g	g	g	ng	g	g	g
or	g	g	g	g	g	ng	g	g
or	g	g	g	g	g	g	ng	g
or	g	g	g	g	g	g	g	ng

Thus, there are eight more *favorable cases* for conviction, when exactly one panel member votes not guilty. Again, we can apply the combinations formula for finding the number of *k* = 1-person groups that can be formed out of an *n* = 8 persons: $\binom{8}{1} = \frac{8}{1} = 8$.

Member:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
(iii) All but two vote guilty:								
	ng	ng	g	g	g	g	g	g
or	ng	g	ng	g	g	g	g	g
or	ng	g	g	ng	g	g	g	g
.	.	.	.	.	.	.	.	.
or	g	g	g	g	g	g	ng	ng

This gives a total of $\binom{8}{2} = \frac{8 \cdot 7}{1 \cdot 2} = 28$ possible scenarios in which exactly two panel members vote not guilty.

Therefore, we have a total of 1 + 8 + 28 = 37 scenarios possible for conviction, in which at least six panel members vote guilty. Again, these represent the *favorable cases* in the definition of probability. The total number of *cases possible* is now $2^8 = 256$, because each one of the eight panels has two choices. Thus, the probability of a guilty verdict is

$$P = \frac{22}{64} = 0.14453.$$

This means that, in the long run, we can expect an approximate rate of conviction of 14.453 percent.[64]

In this case, if we assume that at least two panel members always vote not guilty, then the number of *favorable* cases for a conviction drops to twenty-eight (as cases (i) and (ii) cannot happen anymore), and the number of *possible* cases also decreases to 256 - 1 - 8 = 247, for the same reason. So, the probability of a guilty verdict under this restriction is

[64] Keep in mind that the percentage of convictions should get close to this 14.453 percent value only after a very large number of trials.

$$P = \frac{28}{247} = 0.11336,$$

which means that, in the long run, we can expect an approximate rate of conviction of 11.336 percent, when at least two of the panel members vote not guilty.

In sum, these probabilistic models show that over time, including cases where there were at least two dissenting not guilty votes, there is a significant imbalance in the likelihood of a conviction:

SIX-member panel	**EIGHT-member panel**
≈34.375 convictions	≈14.45 convictions
Out of 100 trials:	
≈26.316 convictions (if at least two vote not guilty)	≈11.33 convictions (if at least two vote not guilty)

Combining these two estimations within each group gives an average of 30.35 percent convictions expected in six-member panels and 12.89 percent convictions expected in an eight-member panel, which is a difference of 17.46 percent. Therefore, increasing the number of required members from six to eight would shift the balance (at least from a mathematical perspective) substantially towards verdicts favoring the defendant. However, although this offers statistical support for Congress's decision to increase the number of members required for a general court-martial from five to eight,[65] there was no known empirical evidence, until the research discussed in this paper, that a panel of eight would be more likely to acquit than a panel of six, especially in the types of criminal cases that are commonly seen in the U.S. military justice system (e.g., sexual assault cases).

[65] National Defense Authorization Act for Fiscal Year 2017, Pub. L. No. 114-328, sec. 5161, § 816, 130 Stat. 2000, 2897 (2016).

C. Pilot Testing

In order to calibrate the case and test the effectiveness of the protocol, we piloted a mock criminal military trial scenario, involving an allegation of sexual assault committed by a military member against another military member, and tried by a court-martial, with eighteen panels (eleven panels containing six members and seven panels containing eight members) consisting of 122 undergraduates. The undergraduates were randomly assigned to a six- or eight-member panel. In this pilot testing, 36 percent of the six-member panels, and 0 percent of the eight-member panels determined that the accused, "Airman Abis," was *guilty*. Using the average of probabilistic modeling statistics of the verdicts with no restrictions and the verdicts with at least two not guilty votes as a benchmark (30.35 percent in six-member panels and 12.89 percent in eight-member panels), we determined that the case as presented likely contained too many exculpatory facts to return guilty verdicts in the eight-member panels. Therefore, to ensure the case would be more balanced towards conviction, we removed two statements made by Airman Kinsey's (the alleged victim) roommate from the trial script, "I overheard her say something about masturbation. Airman Abis asked whether he should close the door and she said, "I don't fucking care." Removing these sentences resulted in a more balanced case, with more panels finding the accused guilty. The following methods and results pertain to all of the panels conducted, using the calibrated case presentation, subsequent to this pilot testing.

D. Participants

Participants were 265 university students (162 women, 103 men) enrolled in a psychology subject pool at a midwestern university who received course credit for participating in the study. Their average age was 20.38 (*standard deviation (SD)* = 4.48, *range* = 18 to 55), and most described their sexual orientation as heterosexual (94 percent), followed by bisexual (3 percent), gay/lesbian (2 percent), and other (1 percent). Most participants (96 percent) were not currently and had never been in the military; however, five participants (2 percent) identified as veterans, four participants (2 percent) were current reservists, and one student identified as being a member of the National Guard.

E. Pre-Trial Procedure

Prior to trial day, participants completed a demographic survey and a pre-trial panel member attitudes survey.[66] Upon arrival on the day of the mock trial (to a classroom assembled as mock panel member room), participants were randomly assigned to panels of either six or eight members. They completed consent forms and were then told via a three-minute video that they would be participating in a mock trial. The gravity of the task was emphasized via the video, where they were encouraged to take seriously their roles as fact finders. Additionally, after viewing the opening instructions, they were all required to stand as a group (with their right hand raised) and go through traditional jury instructions. The experimenter read the following:

> Will the jury please stand and raise your right hand? Do each of you swear that you will fairly try the case before this court, and that you will return a true verdict according to the evidence and the instructions of the court, so help you, God? Please say "I do." [Experimenter waited for participants to say "I do."] You may be seated.

F. The Case

After all participants said, "I do," and took their seats, each participant was also provided with a pen and legal pad and was encouraged to take notes (all notes were shredded after the mock trials). The experimenter then played a twenty-minute video containing a fictitious sexual assault case involving two Air Force members: Airman Roberto Abis (accused) and Airman Ellen Kinsey (victim). In the case, Airman Roberto Abis was charged with sexual assault by causing bodily harm. Namely, "In that he did, at or near Scott Air Force Base, Illinois, on or about 14 August 2016, commit a sexual act upon Airman Ellen Kinsey, to wit: penetrating her vagina with his penis, by causing bodily harm to her, to wit: penetrating her vagina with his penis without her consent." The video, narrated by a U.S. Air Force judge advocate with experience serving as a trial defense counsel, included an introduction (including preliminary instructions), presentation of the evidence (including descriptions of testimony from the

[66] *See infra* Section III.H (Measures).

accused, the victim, several witnesses, and a forensic psychologist), substantive instructions on the law (including descriptions of key terms like bodily harm, mistake of fact as to consent, and reasonable doubt), and procedural rules*,* which described the rules they were required to follow during deliberation (a physical copy of the procedural rules was also given to each of the participants before they started their deliberation). The introduction, substantive instructions on the law, and procedural rules were fashioned using Department of the Army Pamphlet 27-9, *Military Judges' Benchbook* (2014).[67]

G. Deliberation

For the deliberation, participants were seated around a rectangular deliberation table with their panel member numbers on the table in front of them (so that they could be identified when commenting). After viewing the case, the experimenter distributed a deliberation packet to each panel member and read them standard instructions about the materials contained in the packets. The experimenter also ascertained who was senior in rank (first in terms of class standing and second in terms of age) and appointed that person the foreperson. The foreperson was given a set of written instructions detailing the steps of the deliberation: 1. participants complete pre-deliberation individual verdict sheet, 2. participants discuss all relevant facts of case, 3. anonymous vote is taken whereby participants write "guilty" or "not guilty" on legal pad paper and hand it to the foreman, 4. foreperson counts votes aloud, 5. foreperson asks if any more deliberation or revote is necessary, 6. participants complete post-deliberation individual verdict sheet, and 7. foreperson completes group verdict sheet). Importantly, the experimenters did not know the true nature of the study. Once the packets were distributed and the foreperson was appointed, the experimenter told the mock panel members to begin their deliberations and to come to the hallway if they had any questions or when their deliberations were complete. Then, the experimenter left the room. All deliberations were recorded using an iPad and large table microphone and uploaded to an online password-protected archive. Overall, thirty-

[67] U.S. DEP'T OF ARMY, PAM. 27-9, MILITARY JUDGES' BENCHBOOK (29 Feb. 2020).

eight of the forty deliberations were successfully recorded throughout the entire deliberation and then transcribed.

H. Measures

1. Pre-Trial

Prior to the mock-trial, participants completed a demographic survey and the previously validated Pretrial Juror Attitudes Questionnaire (PJAQ).[68] Participants responded on a five-point Likert scale (1 = *strongly disagree*, 5 = *strongly agre*e) to indicate their agreement with twenty-nine items in six different categories (for example: "Defense lawyers are too willing to defend individuals they know are guilty:" cynicism (CYN) towards the defense, and "If a suspect runs from police, then he probably committed the crime:" system confidence (CON)). For this study, only these two subscales (CON: Cronbach's alpha = .67, and CYN: Cronbach's alpha = .61) were used.[69]

2. Pre-Deliberation – Individual

Prior to leaving the room, the experimenter advised all participants to complete a pre-deliberation form before engaging in any discussion and to leave it in their personal folder so no one else could see it. This step was also listed in the instructions packet that was given to the foreperson. This verdict sheet, adapted from Ruva and Guenther,[70] asked participants to indicate whether, before any deliberation occurred, they found the defendant guilty or not guilty and then rate their confidence in the verdict on a seven-point Likert scale (1 = *I'm certain he is not guilty*, 7 = *I'm certain that he is guilty*).

[68] Lecci, Len & Myers, Bryan, *Individual Differences in Attitudes Relevant to Juror Decision Making: Development and Validation of the Pretrial Juror Attitude Questionnaire (PJAQ)*, 38 J. APPLIED SOC. PSYCHOL. 2010 (2008).

[69] Cronbach's alpha is a measure of internal consistency of a test or scale. It is a test of reliability (whether responses are consistent between questions).

[70] Christine L. Ruva & C. C. Guenther, *From the Shadows into the Light: How Pretrial Publicity and Deliberation Affect Mock Jurors' Decisions, Impressions, and Memory*, 39 L. & HUM. BEHAV. 294, 297 (2015).

3. Post-Deliberation – Individual

After the verdict was declared final, participants completed another individual verdict sheet. However, this time the participants were asked to indicate, after all deliberation occurred, whether they found the defendant guilty or not guilty and then rate their confidence in the verdict on a seven-point Likert scale (1 = *I'm certain he is not guilty*, 7 = *I'm certain that he is guilty*).

4. Post-Deliberation – Group

The foreperson completed the group verdict sheet, which was modeled after a verdict sheet from criminal court contexts. On this sheet, they were required to enter whether their panel found the defendant, Airman Abis, guilty or not guilty, and the final vote count. To be certain the two-thirds vote was used appropriately, the calculation was provided on the verdict sheet (such as: "Two-thirds majority vote is required for a [g]uilty verdict (for example: six-eighths or four-sixths)").

5. Deliberation Times

Deliberation times were computed by inspecting the electronic video files, calculating the time between when the experimenter left the room and when the foreperson left to retrieve the experimenter at the end of the group's deliberation.

6. Deliberation Comments

Deliberation comments were evaluated on their content as per Horowitz and Bordens.[71] After all deliberations were transcribed, the deliberations were segmented into single units of information (propositions)—resulting in a total propositions measure. Using the same classification scheme as Horowitz and Bordens,[72] two independent raters

[71] Irwin A. Horowitz & Kenneth S. Bordens, *The Effects of Jury Size, Evidence Complexity, and Note Taking on Jury Process and Performance in a Civil Trial*, 87 J. APPLIED PSYCH. 121, 125 (2002).
[72] *Id.* at 125.

then classified each proposition as probative (case-related information, like "she had a boyfriend, she called her boyfriend to try to get him to come over" and "Justin said when he came over he did smell alcohol on her breath"), non-probative (not case-related, irrelevant, or incorrect, such as "three [drinks] – I counted three," and "I kinda felt my personal way about this situation"), and evaluative (evidence or case-based opinions, "but I think that it shouldn't be, like he shouldn't be charged with rape," and "he was more in a right mind than she was"). These raters were unaware of the true purpose of the study. The interrater reliability of the coding was acceptable (Kappa = .78). For final coding, the two raters resolved any differences through discussion.

IV. Results

Overall, forty mock trials were conducted (twenty-seven with six members and thirteen with eight members). One participant (from a six-member panel) did not complete the pre-trial questionnaire and that person's data was excluded; however, because he participated in the trial, the group's results were still presented. Prior to the group analyses, we analyzed the individual data to determine whether age or sex of participants was related to pre-trial attitudes or individual pre- and post-deliberation verdicts. Age was not significantly related to pre- or post-deliberation verdicts (*p*s > .05); however, age was inversely related to PJAQ scores for cynicism towards defense ($r = –.13$, $\underline{p} = .03$) and system confidence ($r = –.23$, $p < .001$), reflecting a more negative view towards the legal system and less cynicism towards defense counsel among older participants. Meanwhile, in terms of sex, women were significantly more likely than men to indicate that Airman Abis was guilty on the pre-deliberation form (64.6 percent of women vs. 49.5 percent of men, $X^2(n = 265) = 5.89$, $p = .02$). However, this difference disappeared in the post-deliberation verdicts; after deliberation, 45.6 percent of men and 42.6 percent of women indicated that Airman Abis was guilty (X^2 ($n = 265$) = 0.24, $p = .63$). Meanwhile, men and women did not differ significantly in their pre-trial attitudes towards defense or their system confidence (*p*s > .25).

We also examined whether the six- and eight-member panels were similar on these demographic characteristics. As shown in Table 2, the panels were similar in terms of age and their pre-trial attitudes towards defense counsel and the legal system. However, there were significantly more men in the six-member panels than the eight-member panels. As women were more likely to indicate that Airman Abis was guilty on their pre-deliberation forms, this sex difference could potentially translate into a slight bias towards guilty verdicts for the eight-member panels. However, this was not the case; the individual pre-trial verdicts of the six- and eight-member panels did not differ significantly. The guilty votes in the individual pre-deliberation verdict sheets for the six- and eight-member panels, were 62 percent and 53 percent, respectively (*X2* ($n = 264$) = 1.97, $p > .05$).

Table 2

Descriptives and Significance Tests for Demographic Characteristics of Six- and Eight-Member Panels

	6 members	**8 members**	
	M (*SD*)/ *N* (%)	*M* (*SD*)/ *N* (%)	*t*/X^2
Age	20.40 (4.23)	20.35 (4.86)	0.91
Male Sex	72 (69.9)	31 (30.1)	$X^{2=}$ 5.91, *p* = 0.02
Cynicism towards Defense	2.97 (0.50)	2.99 (0.57)	–0.40
System Confidence	2.94 (0.75)	2.95 (0.56)	–0.13

A. Verdicts

Among six-member panels, thirteen of twenty-seven (48 percent) returned a guilty group verdict, whereas in eight-member panels, four of thirteen (31 percent) returned a guilty group verdict. Thus, the probability of the accused being convicted dropped 17 percent in cases when two additional panel members were added to the panel.

As Table 3 shows, in the groups who returned a guilty verdict, most of the individual members (in both the six- and eight-member panels) thought the accused was guilty before the deliberation began, and this number increased after the deliberation. However, there were no significant differences in the average individual pre- or post-deliberation verdicts for the six- or eight-member panels. Additionally, after they returned their individual final verdicts, there were no significant differences between the six- and eight-member panels in how confident they were in their ratings. In both groups, participants were, on average, “pretty sure he is guilty.”

Table 3

Descriptives and Significance Tests for Pre- and Post-Deliberation Guilty Votes and Verdict Confidence Ratings by Panel Size When Group Verdict was Guilty

	6 members		**8 members**		
	M (*SD*)	% guilty votes	*M* (*SD*)	% guilty votes	*t*
Pre-deliberation verdict	0.75 (0.43)	75%	0.63 (0.49)	63%	1.28
Post-deliberation verdict	0.91 (0.29)	91%	0.84 (0.37)	84%	0.99
Confidence	5.75 (1.49)	38%[a]	5.72 (1.41)	28%[a]	0.09

Note. Six-member panel *n* = 77, eight-member panel *n* = 32. For verdicts, 0 = not guilty, 1 = guilty. For confidence ratings, 0 = certainly not guilty, 7 = certainly guilty. [a]Percentage of those who indicated that they were "certain" that Airman Abis was guilty; there were no significant differences between six- and eight-member panels in these percentages $X^2(n = 109) = 0.91, p > .05$).

Meanwhile, in the groups that returned a not guilty verdict, most of the participants began the deliberation with the belief that the accused was not guilty, and the percentage who believed he was not guilty increased after the deliberation was finished.[73] Again, there were no significant differences in the average individual pre- or post-deliberation verdicts for the six- or eight-member panels. However, this time, there was a difference between the six- and eight-member panels in how confident they were in their post-deliberation verdict ratings. Those in eight-member panels had significantly greater confidence in their not guilty verdicts than those in six-member panels. While the average eight-member panelist

[73] *See infra* Table 4.

voting for acquittal was "pretty sure he is not guilty," the average six-member panelist voting for acquittal was "not sure but think he is guilty."

Table 4

Descriptives and Significance Tests for Pre- and Post-Deliberation Guilty Votes and Verdict Confidence Ratings by Panel Size When Group Verdict was Not Guilty

	6 members		**8 members**		
	M (*SD*)	% guilty votes	*M* (*SD*)	% guilty votes	*t*
Pre-deliberation verdict	0.50 (0.50)	50%	0.49 (0.50)	49%	0.71
Post-deliberation verdict	0.13 (0.34)	13%	0.11 (0.32)	11%	0.67
Confidence	2.83 (1.61)	17%[a]	2.29 (1.42)	26%[a]	2.22*

Note. Six member $n = 84$, eight member $n = 72$. For verdicts, 0 = not guilty, 1 = guilty. For confidence ratings, 0 = certainly not guilty, 7 = certainly guilty. [a]Percentage of those who indicated that they were "certain" that Airman Abis was not guilty; there were no significant differences between six- and eight-member panels in these percentages $X^2(n = 156) = 2.20, p > .05$).

B. Quality of Deliberation

The deliberation times in the six-member panels ranged from 4.42 minutes to 44.13 minutes, and the deliberation times in the eight-member panels ranged from 7.23 minutes to 31.20 minutes. There was no significant difference between the six- and eight-member panels in how many minutes they spent deliberating, on average ($M = 21.41$, $SD = 9.84$ versus $M = 18.53$, $SD = 7.41$, respectively, $t(37) = 0.93$, $p = 0.34$). However, there was a trend for those in the six-member groups to spend more time in deliberations. As Table 5 shows, there were also no significant differences in neither the number of total propositions elicited by members in six- and eight-member panels nor in their percentage of probative, non-probative, and evaluative propositions. However, there was a trend for the six-member panels to produce more non-probative propositions during deliberation than the eight-member panels. Additionally, although the length of deliberation time was significantly and positively correlated to the percentage of probative statements in the eight-member panels (r =.63, $p < .05$); the length of deliberation was not significantly related to the percentage of probative comments in the six-member panels (r = .39, $p > .05$). In other words, the length of deliberation appeared to produce more substantive deliberations in the eight-member panels but not in the six-member panels.

Table 5

Descriptives and Significance Tests for Number and Percentage of Deliberation Propositions by Panel Size

	6 members		**8 members**		
	M (*SD*)	%	*M* (*SD*)	%	t^a
Total propositions	284.04 (171.10)		234.38 (97.19)		0.96
Probative proposition	76.76 (49.08)	26%	65.61 (33.03)	27%	–0.50
Non-probative proposition	171.16 (102.50)	62%	133.54 (60.34)	58%	1.26
Evaluative proposition	36.12 (28.53)	12%	35.23 (15.13)	15%	–1.51

Note. [a]t-tests were performed to compare the two panel sizes on the percentage statistic (such as proposition/total propositions for each category).

V. Discussion and Recommendations

Previous studies concerning civil trials have already found that increasing panel size to twelve members results in a more diverse panel that engages in more robust deliberations.[74] Our goal was to test whether a more modest increase in panel size, corresponding to recent changes in military court-martial panel composition, has similar effects. Our study's findings confirm the validity of Saks and Marti's aggregation—progressively larger panels are more likely to acquit than smaller panels. However, we did not find, as Saks and Marti predicted, that such a relatively small increase in panel size resulted in more meaningful deliberations, at least not significantly (although eight-member panels were marginally less distracted by non-probative propositions). Of note, the smaller six-member panels were 17 percent more likely than eight-member panels to convict an accused. This aligns with the difference predicted by the probabilistic modeling (17.46 percent). The fact of that alignment supports the practical applications of probabilistic modeling for making predictions in human behavior in the context of trials. As discussed earlier, whether the increased risk of conviction posed by smaller panels represents an increase in the risk of "incorrect" verdicts is not a question easily amenable to scientific measurement. Nonetheless, exposing an accused to a risk of conviction that is higher than can be explained by statistics is a matter of concern for any criminal justice system. Justice requires verdicts to be reliable—in both fact and appearance.

This study made some additional findings that were not explored by Saks and Marti. For example, the discovery that older panel members tended to hold more negative views of the legal system while, conversely, viewing defense counsel with less cynicism, warrants further study. Likewise, more research could be focused on this study's finding that women were more likely than men to start deliberations with a view that the accused was guilty, but that their differing views at the start of the deliberative process were washed away by the end. This study's findings regarding the relative confidence that individual panel members had in the verdict their court reached also could be a fruitful area for further research. Panel members who were part of larger eight-member courts where the verdict reached was not guilty were significantly more confident in that

[74] Horowitz & Bordens, *supra* note 71, at 125-28.

verdict than their counterparts in six-member courts who had reached the same verdict. Confidence in verdicts is an important goal for any system of justice. Future research also should study the extent to which race and ethnicity of the panel members, as well as the accused and victim, might influence outcomes for panels of varying sizes, especially if unanimity is not required for the panel to render a verdict of conviction.

The implications of this research on military justice policy are profound. Increasing the size of panels, conclusively, increases the chance that the accused will be found not guilty of the Government's allegations. This may be especially important for a system of justice that has, of late, been criticized for pursuing adult penetrative sexual assault prosecutions all the way to verdict even though, in 31 percent of those cases, at the time the charging decision was made, the Government lacked sufficient evidence to support a reasonable expectation of obtaining or sustaining a conviction on that allegation.[75] In 10 percent of those cases, the Government lacked even probable cause.[76] In such a system, where contemporary standards of prosecutorial discretion[77] are not employed to prevent weak cases from going forward to trial, it is important to ensure that the trial forum is exceptionally reliable. Smaller panels lack such reliability. Further, allowing those smaller panels to render non-unanimous verdicts amplifies the risk that the voice of racial and ethnic minorities on those panels will be diluted. Such dilution is not only potentially dangerous to innocent minority defendants but is also incongruent with the goal of ensuring that racial and ethnic minority communities have the equal opportunity to participate in our system of government, including judicial systems.

For these reasons, abolishing court-martial panels that are smaller than eight members (as exists in special courts-martial and general courts-

[75] DEF. ADVISORY COMM. ON INVESTIGATION, PROSECUTION, AND DEF. OF SEXUAL ASSAULT IN THE ARMED FORCES, REPORT ON INVESTIGATIVE CASE FILE REVIEWS FOR MILITARY ADULT PENETRATIVE SEXUAL OFFENSE CASES CLOSED IN FISCAL YEAR 2017, at 56 (2020).
[76] *Id*. at 57.
[77] The Department of Defense has prescribed a standard for prosecution but has expressly made that standard non-binding. *See* MANUAL FOR COURTS-MARTIAL, UNITED STATES app. 2.1, ¶ 2.3 (2024) ("This Appendix provides non-binding guidance issued by the Secretary of Defense [. . .] [convening authorities and special trial counsel] should not refer a charge to a court-martial unless the admissible evidence will probably be sufficient to obtain and sustain a finding of guilty when viewed objectively by an unbiased factfinder.").

martial where members become unexpectedly unavailable after the court has been impaneled) should be a congressional priority, as should the abolishment of non-unanimous verdicts in all courts-martial.