Wargaming for International Relations Research

Forthcoming, *European Journal of International Relations*

Erik Lin-Greenberg, Reid B.C. Pauly, and Jacquelyn G. Schneider[1]

Political scientists are increasingly integrating wargames into their research. Either by fielding original games or by leveraging archival wargame materials, researchers can study rare events or topics where evidence is difficult to observe. However, scholars have little guidance on how to apply this novel methodological approach to international relations research. This article evaluates how political scientists can use wargames as a method of scholarly inquiry and sets out to establish a research agenda for wargaming in international relations. We first differentiate wargames from other methodological approaches and highlight their ecological validity. We then chart out how researchers can build and run their own games or draw from archival wargames for theory development and testing. In doing so, we explain how researchers can navigate issues of recruitment, bias, validity, and generalizability when using wargames for research, and identify ways to evaluate the potential benefits and pitfalls of wargames as a tool of inquiry. We argue that wargames offer unique opportunities for political scientists to study decision-making processes both in and beyond the international relations subfield.

[1] Authors listed in alphabetical order. The authors thank Valentin Bolotnyy, Amber Boydstun, Cole Bunzel, Peter Dombrowski, and Rose McDermott for helpful comments on drafts, as well as participants at the 2020 American Political Science Association Conference, the Georgetown University Wargaming Society, King's College London, and a 2019 MIT-NWC workshop, especially Richard Samuels, Eric Heginbotham, Stacie Pettyjohn, Ellie Bartels, and Andrew Reddie. Andrew Ortendahl provided excellent research assistance.

Human behavior and decision-making are at the core of the most enduring puzzles in international relations (IR). Yet data about decision-making, especially involving rare events and the elite decisionmakers associated with security and foreign policymaking, can be difficult to obtain. In recent years there has been a revival of interest in wargaming as a way to both generate and obtain these behavioral insights (Colbert et al., 2017; Schneider 2017; Pauly, 2018; Reddie et al., 2018; Jensen and Valeriano, 2019; Bartels, 2020, Dorn et al., 2020; Hirst, 2020; Williams and Drew, 2020).[2] Long the territory of policymakers, IR scholars are beginning to leverage archival data from historical wargames and have also fielded their own games to test theories on decision-making and conflict dynamics. Together, this up-and-coming scholarship uses wargames to explore the mechanisms and logics that underpin foreign policy decisions.

This burgeoning interest is a product of three factors. First, the declassification of Cold War-era defense wargames provides scholars with new and unique archival materials to better understand historical decision-making on topics like nuclear use and conflict escalation. Second, over the last two decades political scientists have increasingly turned to synthetic data generating processes like survey and lab experiments (Hyde, 2015). This behavioral turn has emphasized experimental design, which political scientists have applied when fielding their own games. Third, political scientists are increasingly interested in the microfoundations that underlie theories (Kertzer, 2017). By shedding light on decision-making processes, wargames provide researchers with a novel methodological tool for exploring and testing mechanisms upon which IR theories lie, potentially shedding deeper insights than other research approaches.

[2] Work on wargames within social science disciplines has increased four-fold since 2005. A Scopus search identified 44 books, articles, chapters, or reviews on wargaming in 2005 and 192 in 2017.

Wargames may offer scholars a promising tool to answer questions in creative ways, but before the field embraces wargaming methods or data, we need to better understand the promises and pitfalls of gaming for political science. How is wargaming different than other research approaches? What types of insights and data can wargames generate and how can researchers best use them for research? What should scholars consider when designing their own games? What methodological questions should be addressed to advance wargaming as a method for international relations research?

This article charts and evaluates how political scientists can use wargames as a method of scholarly inquiry and sets out to establish a research agenda for IR wargaming. We explore the development of researcher fielded wargames and the use of archival wargame material to generate insights on decision-making. We consider the utility of games for theory development and testing; examine issues of bias, validity, and generalizability; and describe how games can shed light on the microfoundations that underpin core IR theories.

The article proceeds in five sections. First, we define wargaming and identify different game types. Second, we review a series of propositions on the value of wargaming that differentiate it from other political science research approaches. Third, we discuss how researchers, using social science precepts as a guide, can evaluate the costs and benefits of design choices. Fourth, we describe the historical wargames material emerging in archives, how best to use this documentary evidence, and identify what historical wargaming can teach us about best practices for researcher-fielded games. We conclude by outlining a wargaming research agenda, exploring how wargames can complement other research approaches, contribute to ongoing debates, and propose specific questions that will help researchers better understand the inferences that can be drawn from wargames.

WHAT IS A WARGAME?

The use of wargames goes back millennia, with evidence of games in ancient Rome, early Iraq, and China (Caffrey, 2019). Wargames took on a central role in the modern conduct of war with the Prussian development of *Kriegspiel*, a boardgame that simulated combat to train officers (Schuurman, 2019; Wilson, 1968). A century later, the United States' embrace of wargames for military planning between World Wars I and II became a pivotal part of the Navy's success in the Pacific (Lillard, 2016). During the Cold War, the U.S. military turned again to wargames to understand the impact of the nuclear revolution (Schelling, 1987; Pauly, 2018). U.S. defense wargaming continued after the Berlin Wall fell, with games designed to test new ideas about warfare and aid acquisition decisions (Krepinevich and Watts, 2015).

Despite their long history, it is not always clear what constitutes a "wargame" (Sepinsky, 2021). While wargames emerged to prepare for combat, their use extends beyond the study of war. Governments use games to simulate natural disasters and to assess economic cooperation (Abbasi et al., 2012; Smith and Bell, 1992); consultants use wargaming to test new business strategies (Oriesek and Shwarz, 2008); and scholars have applied gaming to study how human behavior affects various social and political phenomenon (Banks et al., 1968; Camerer, 2011; Fiorina and Plott, 1978). Thomas Schelling's work on coercion, for instance, was inspired in large part by Department of Defense wargames he designed (Schelling, 1987), and Schelling's contemporaries used simulations to explore conflict and nuclear use (Brody, 1963; Bloomfield and Whaley, 1965; Hermann and Hermann, 1967). Scholars have since used experiments embedded in games to test explanations for conflict initiation (Johnson et al., 2006; McDermott, Cowden, and Rosen, 2008) as well as domestic political bargaining (Huckfeldt et al., 2014; Hamman et al., 2011). Most recently, political scientists have used wargames to study phenomena where data are scant, such

as the effect of emerging technologies on international relations (Schneider 2017; Reddie et al., 2018; Pauly, 2018; Jensen and Banks, 2018; Jensen and Valeriano, 2019; Williams, 2020; Lin-Greenberg, 2020, Schneider et al., 2021).

While often called "simulations" or "exercises," wargames are distinct from computer simulations of combat, field exercises featuring actual military forces, or organized brainstorming sessions. Moreover, most traditional wargames are also not lab experiments designed to study causal effects. Instead, wargames are interactive events that display four characteristics: human players, immersed in scenarios, bounded by rules, and motivated by consequence-based outcomes.

First, wargames involve human players. As Peter Perla (1990: 164) explains, "a wargame is an exercise in human interaction . . . its forte is the exploration of the role and the potential effects of human decisions." This human characteristic makes wargames ideal for research in which either the dependent variable or the hypothesized causal mechanism is about human behavior. Indeed, games can help shed light on the microfoundations, or lower-level mechanisms derived from individual human behavior, that underpin many scholarly theories (Kertzer, 2017). The human element of wargames differentiates them from computer simulations or econometric "games" in which models simulate assumed human behaviors.

Second, wargames place human participants into scenarios that simulate real-world decision-making (Pettyjohn, 2019). The representation of reality and the integration of context generates the thickness of wargaming scenarios and differentiates them from the lab and survey experiments increasingly used in IR research. These simulated decision-making environments, similar to those that participants regularly experience, can induce players to behave in ways that closely mirror their behavior when presented with similar real-world contexts.[3] Wargame

[3] Psychologists refer to this as "ecological validity." Egon Brunswik (1947), who developed the concept, defined it as the "degree of correlation between a proximal cue and the distal variable to which it's related." More recent scholarship

designers must carefully balance abstraction, which makes games easier to execute, and realism, which is unique to wargames and may ultimately increase the robustness of game findings.

Third, wargames feature rules that dictate how human players interact with the scenario. Rules may be rigid, in which players have a limited set of actions, or allow for free-play, where players are given few constraints. These rules can shape player behaviors and outcomes, ultimately affecting the conclusions observers can gather from a game. Rules therefore create complex design trade-offs. For example, free-play games can make replication difficult, whereas rigid games are more likely to unnaturally constrain outcomes. Although rules are a characteristic that games share with many simulations, models, and experiments, wargames (especially those with multiple moves, players, or teams) often use more complicated rules that govern how teams can interact while permitting a wider array of behavioral choices and, thus, more variance in outcomes. For instance, a survey experiment may ask subjects whether or not they wish to use military force, while a wargame may ask them how and when to employ the military forces at their disposal.

The fourth characteristic of wargames that distinguishes them from most other IR research approaches is the experiential nature of their consequence-based outcomes. As Bartels argues, a wargame must immerse human players "in a competitive environment based on a set of implicit or explicit rules, . . . [to] grapple with the potential consequences of their actions" (Bartels 2020). These consequences – such as "losing" a wargame or having decisions made in an earlier round affect a subsequent round – are thought to incentivize participants to consider their decisions more deeply. In more common research approaches such as survey experiments, participants generally do not face consequences, real or simulated. Wargames, at their best, transcend players beyond "gaming" outcomes to feeling and internalizing consequences of their behaviors. The success of

---

treats ecological validity as "a measure of how test performance predicts behaviors in real-world settings (Gouvier, Barker, and Musso, 2014)."

this final characteristic is tied to game designers making trade-offs within the previous three characteristics, including using the right players, creating appropriate scenarios, and building useful rules.

In sum, *wargames are interactive scenarios which immerse human players who make decisions in accordance with given rules and react to the consequences of their choices*. Variation in these four characteristics has led to a variety of "wargames" that look very different from one another. Wargames, for instance, include boardgames, tactical tabletop exercises with a handful of players, and political-military games with hundreds of participants. They may be played in-person, virtually, or using some hybrid combination, and feature different rules (Table 1). Scholars need to understand how these game characteristics affect the conclusions about international relations theory and decision-making that can be drawn from games; something we explore in subsequent sections.

Table 1: Characteristics of Games	
Players	At least two to thousands. Players include elites/experts or convenience samples.
Scenarios	Range from tactical simulations with limited geographic or substantive scope to strategic simulations featuring all of government decision-making.
Rules: Moves	Games can include one set of decisions (i.e., a move), multiple moves, or may occur continuously over an extended period. Moves may represent decisions taken within a set period of time (e.g., 30 days) or correlate with real time.
Rules: Sides	Sides are the number of teams in the game. In one-sided games, one team "plays" with no feedback. In one and a half-sided game, one team of players plays against adjudicators. Two-sided games usually include two teams playing against each other. Multi-sided games feature more than two teams playing in the game (e.g., allies).
Consequence-based Outcomes: Adjudication	Adjudication, which dictates consequence-based outcomes, may occur via probability tables, random distribution, or assessment by subject matter experts.

WHY WARGAMES?

Above we outlined what wargames are, but why might a researcher choose wargames over other methods or data sources? Below, we identify four propositions about the usefulness of wargames as a research tool to study decision-making: (1) wargames are more immersive for research subjects than other approaches, (2) elite participants who often play wargames more closely resemble actual decisionmakers than the public or convenience samples common to other methods, (3) interaction between participants better represents real-world decision-making, and (4) wargames present players with the consequences of their own decisions. Together, these propositions suggest that the primary value of using and analyzing wargames is not in generating new or better data about outcomes, but is instead in understanding behaviors and choices leading to these outcomes. Wargames do not predict what will happen in conflict or crisis, but they can tell us why and how one outcome or another occurred. While widely accepted within practitioner communities (Perla and McGrady, 2011; Oberholtzer et al., 2019; Wong et al., 2019; Perla, 1990; Bartels 2020), these assumptions about the value of wargaming as a research tool are mostly untested. We lay them out in this section to begin outlining a forward-looking research agenda on the unique role of wargames alongside other methodologies and archival data sources. In the conclusion, we assess how researchers might study these propositions and describe the types of questions scholars might tackle using wargames.

Overall, each of these four propositions improves the *ecological validity* of wargaming as a research approach. Ecological validity—a concept common in psychology research—concerns the extent to which behavior under test conditions mirrors real-world behavior. Put differently, more ecologically valid research designs should offer more robust insights on actual behavior. To attain high ecological validity, psychologists focus on three key dimensions. First, the *test*

environment should include features—such as time constraints and distractions—that occur in natural settings, rather than exhibiting the more sterile and unrealistic nature of stripped-down laboratory settings. Second, the *stimuli*—such as information injects—in a simulated environment should bear resemblance to real-world stimuli. Third, the *behavioral responses* and actions a participant can make in a test should be representative of those that they could make in the real-world. Information from tests involving environments that are too unrealistic or involving unnatural stimuli and behavioral responses may limit the conclusions that can be drawn from a study (Gouvier, Barker, and Musso, 2014).

Achieving high ecological validity in wargames requires simulation conditions that reflect the type of pressures, incentives, and information environment that real policymakers would have to contend with in an actual crisis. These conditions then allow participants to propose solutions similar to those they could propose in the real-world. If wargames can feature high ecological validity, scholars should be able to use wargames to realistically simulate and study foreign policy decision-making processes. Indeed, ecological validity enables professionals in other fields to use games and simulations for training purposes. A flight simulator programmed with accurate real-world parameters, for instance, has high ecological validity and is a cheaper, easier, and safer way to train pilots and learn about their decision-making.[4] While literature across disciplines continues to debate its definition (Brunswik, 1947; Baumeister and Vohs, 2007: 276; Schmuckler, 2001), we consider ecological validity to be a key element of external validity—the generalizability of research findings beyond the research context (Findley, Kikuta, and Denly, 2020).[5]

[4] We thank an anonymous reviewer for this example.
[5] John Kihlstrom (2021) summarizes this debate: "Egon Brunswik coined the term ecological validity to refer to the correlation between perceptual cues and the states and traits of a stimulus. Martin Orne adapted the term to refer to the generalization of experimental findings to the real-world outside the laboratory. Both are legitimate uses of the term[.]"

Proposition 1: Wargames are more immersive than other methods and therefore more ecologically valid.

As wargaming experts Perla and McGrady (2011: 113) assert: wargames "draw players into both participating in and constructing their narratives; they literally place the players inside the narratives." This is an argument for wargames over other approaches that do not replicate real-world decision-making environments. In terms of the dimensions of ecological validity, immersion purports to offer a valid test environment with valid stimuli. In the ideal, players are so immersed that they temporarily forget or ignore the fact that they are being studied and care only about their progress in the wargame. Accordingly, wargames seek to create immersive environments that feature stimuli in which participants act not as game players but instead internalize how they have in the past and would in the future react to similar real-life scenarios. Historically, wargames have reflected the real-world experiences of government participants. During the Cuban Missile Crisis, a defense official who had previously participated in wargames run by Thomas Schelling remarked, "This crisis sure demonstrates how realistic Schelling's [war]games are," to which a colleague responded, "No," the wargames "demonstrate how unrealistic this Cuban crisis is" (Schelling and Ferguson, 1988: 10).

Of course, a wargame could be stripped of enveloping detail and lose its immersive quality, but then it becomes questionable whether the activity can still be considered a wargame. Survey experiments, for instance, prioritize internal validity and control but often lack such immersive interaction or stressors of actual decision-making settings (Barabas and Jerit, 2010). In contrast, games—which usually last hours or days and feature extensive detail—can elicit participant buy-in by providing realistic scenarios, creating conditions where participants can win or lose vis-à-vis another team, and by allowing extended interaction between participants. Players, who invest time

and energy to participate, may better comprehend the scenario and care more about its outcomes than less immersed research subjects. As a result, they may respond more thoughtfully to a given scenario. Further, wargames often require players to make decisions with too much (or too little) information, time constraints, and emotional burdens, creating what McDermott (2002) terms "experimental realism." Indeed, scholars find that "synthetic experiences," which present research subjects with immersive fiction or videos, trigger cognitive processes akin to real-world decision-making (Daniel and Musgrave, 2017; Miller, 2020). Finally, immersion in wargames goes beyond many survey experiments by asking participants to play the role of decisionmakers and to answer what *would I do*, rather than *what would I support others doing?*[6]

Proposition 2: More representative samples make wargames more ecologically valid.

Wargames may offer greater insights than other research methods simply because they have traditionally recruited expert participants including policymakers and military officers. Scholars commonly argue that research offers the most useful insights when the study sample reflects the population of interest (McDermott, 2002; Hyde, 2015; Dietrich et al., 2021). IR scholars, however, have increasingly turned to larger online and student convenience samples for empirical studies. While this approach allows for repeatable, statistical analysis that overcomes the fundamental problem of causal inference and enables the study of public preferences, convenience samples may yield limited insights about government decision-making if subjects are not representative of actual policymakers (Oberholtzer et al., 2019; Dietrich et al., 2021).

In contrast, elite wargames typically feature the opposite of convenience samples—participants are deliberately recruited for their substantive knowledge or their experience in real-

[6] To be sure, an increasing number of survey experiments fielded on elites and practitioners ask respondents what they would do in a given scenario. See, for example, Tomz et al., 2020 and Chu and Recchia, 2021.

world decision-making. This recruitment strategy can result in highly realistic samples. However, even these realistic samples may include variations in experience and worldviews that significantly affect decision-making. For example, would national security experts from the Obama administration consider the same factors when making decisions as national security experts from the Trump administration? The representativeness of these elite research subjects may therefore improve the ecological validity of wargames, but—if only a limited number of participants are recruited to play a small number of games—researchers still need to clearly explain the inferences drawn from findings (and associated limitations).

Even if the number of elite participants is small and recruitment is targeted, the uniqueness of elite wargame participants can provide important analytical insights. Participant deliberations during games, for instance, might shed light on what factors elites emphasize or deemphasize when making decisions. How important, for example, were norms or ethics to decisions about conflict? What beliefs about international politics did participants bring into decision-making? Did they discuss mental models, historical analogies, or other heuristics when making decisions? Since elite participants can draw from their substantive knowledge and expertise when playing both researcher-fielded and government-sponsored games, the insights from these games may be more useful for IR theory testing than games played by non-experts.

Proposition 3: Group interaction is more representative of real-world decision-making than experiments or surveys that collect individual-level preferences.

A significant difference between most wargames and other synthetic data generating processes is the role of groups in decision-making. Wargames are inherently multi-player efforts, whereas most survey experiments and many lab experiments collect responses from individual

participants. The interactions of players within and across teams that ultimately shape decisions during wargames are important because real-world foreign policy decisions are rarely made by a single individual (Saunders, 2017; Mintz and Wayne, 2016; Kerr and Tindale, 2004). Group-level interaction in most wargames provides a unique opportunity to study how decision-making unfolds, and potentially enhances ecological validity by better simulating actual decision-making processes and behavioral responses than other research approaches. Indeed, factors such as emotions, hubris, miscommunication, status, reputation, diversity, gender, experience, and hawkishness can influence group dynamics and decision-making during wargames, allowing researchers an opportunity to explore how these important (but difficult to collect) variables affect foreign policy (Wang et al., 2020). Wargaming discussions can also reveal how groups self-sort and assign or defer decision-making responsibility according to the dispositions or characteristics of team members. Indeed, one rapporteur of a 1960 wargame shrewdly noted that policy heavyweight Walt Rostow "did an estimated 75% of the talking" on the U.S. team (Bloomfield, 1960). In contrast, surveys and many experiments often overlook group dynamics and make generalizations about foreign policy decisions by measuring individual-level preferences.

Proposition 4: Wargames present players with consequences, creating more ecologically valid data about outcomes and decision-making.

Games may be more likely to mirror real-world decision-making because they ask players to make choices that respond to or result in consequence-based outcomes.[7] This experiential quality of wargames, which requires players to adjust their strategy in the wake of simulated challenges, goes beyond concerns about the consequences of iteration or the shadow of the future.

[7] Many lab and field experiments incorporate consequences (e.g. points or money) for certain types of behavior. Yet, recent IR studies generally do not offer these types of incentives. For an exception, see Quek 2017.

Opposing teams engaged in political-military signaling as a "process of feeling around for what the other side might accept or reject" (Schelling and Ferguson, 1988: 1) are doing far more than deciding how to divide a dollar in a lab. Indeed, these decision-making logics could be akin to what Hayward Alker described as "inner monologues" of players involved in prisoner's dilemma games, shedding light on how humans interpret their own and others' actions (Alker, 1985).

First, wargames often allow players to "win" or "lose," at least relative to other participants. As a leading wargaming practitioner asserts, "wargames are a human activity . . . when people lose in games, they feel the loss. When they win, they get excited" (McGrady, 2019). Here, again, the introduction of consequences offers a more ecologically valid test environment in ways that can shape behavioral responses. Second, this proposition asserts that the intensity of such feelings of loss or excitement increase with the amount of effort research subjects have invested in their strategies. Placing a group of people together over an extended period of time can increase the salience of these consequences by investing players in games more so than survey experiments conducted online, via telephone, or mail. Games, therefore, allow researchers to examine the trade-offs, choices, and risks that participants take to win, helping to explore the microfoundations of IR theories.

Together, these four propositions suggest that wargames provide researchers with valuable insight into decision-making in situations where real-world data are limited. Critically, the value of wargames is not in determining outcomes, but in shedding light on how decisionmakers arrived at those outcomes. Although wargames are inherently simulations of reality, we believe their immersive nature, group interaction, consequences, and the use of elite samples more accurately model real-world decision-making environments than other research methods, boosting the ecological validity of findings relative to other approaches. Any individual game design may

accentuate some propositions while diminishing others—for instance, a one-sided game could sacrifice some competitive spirit while privileging group interaction—but wargames maintain some of the value of each proposition. In the sections that follow, we map out how scholars can field original wargames and use archival wargame data for research, and identify how to navigate these four propositions when designing games and analyzing game data.

SCHOLAR-GENERATED WARGAMES

Scholar-generated wargames are best used to answer questions about human decision-making, either regarding rare events or topics where real-world data are difficult to obtain. Accordingly, existing research using scholar-generated games tends to answer questions about emerging technology and nuclear weapons (Reddie et al., 2018; Schneider, 2017; Schechter et al., 2021; Jensen and Valeriano, 2019; Lin-Greenberg, 2020, Schneider et al., 2021). However, wargames can also be useful for studying a range of international relations topics, including group dynamics in foreign policy decision-making, the strength of norms in policymaking, the role of treaty commitments in decisions on the use of force, the development and utility of economic sanctions, perceptions of the comparative effectiveness of deterrence strategies, and the fidelity of crisis signaling.

In this section we integrate best practices from professional wargaming and political science research design to offer a how-to framework for scholars developing their own research wargames. In doing so, we outline the tradeoffs between ecological validity, internal validity, and feasibility of implementation. Figure 1 summarizes our key design recommendations.

Game Design and Iteration

The first step in wargame development is to determine whether the research question can best be answered using an observational or experimental design. Observational games are generally standalone events that manipulate neither the players nor the scenario they confront. A single observational game typically reveals a possible outcome in a defined scenario, making this type of game best suited either for exploring general decision-making processes or generating hypotheses.[8] In contrast, experimental games test hypotheses by varying key factors of interest – such as details about the scenario – creating "treatment" and "control" games that allow researchers to study how specific variables affect decision-making.[9]

The type of game typically affects the number of iterations required. Experimentally designed games may require more iterations than observational games in order to assess whether experimental manipulations lead to trends in decision-making. Researchers are increasingly fielding dozens to hundreds of experimental game iterations (Reddie et al., 2018; Schechter et al., 2021; Jensen and Valeriano, 2019, Schneider et al., 2021) to identify trends across games and help ensure findings are not the result of chance.

Participants

Even more important than iteration for both observational and experimental wargames is player selection. When making choices about sample, scholars should ask two questions: (1) Is my

[8] While many practitioner games are observational, we focus on designing experimental games since these games feature many of the same elements as observational games. For more on the use of scenarios for research question/hypothesis generation see, Barma et al. (2016).

[9] To be clear, experimental games may both randomly assign stimuli and randomly divide (non-randomly) recruited elite players into teams.

research question about decisions made by a particular entity or is it about human decisionmaking? (2) Who will the players represent?

If the research question is about decisions made by a particular entity, then, in the ideal case, real-world decisionmakers would "play" themselves in wargames. Such a construct would be the most ecologically valid. However, since senior officials rarely have time to participate in even high-profile government-sponsored games, practitioner wargames often rely on proxies including former policymakers or serving lower-level officials, who have sufficient subject matter and organizational expertise. Several researcher-fielded games have relied on this type of sample, drawing players from the military, private sector, and government (Schneider, 2017; Lin-Greenberg, 2020; Schneider et al., 2021). Researchers have also shortened gameplay or fielded virtual games to reduce the burden to elite participants. Regardless of the elite recruiting approach, researchers need to identify whether demographic or ideological characteristics of their sample might limit the conclusions that can be drawn from findings. For instance, a team comprised primarily of participants that served in government decades ago might behave differently than more recently serving officials.

Alternatively, if the research question is about more general decision-making or human behavior—e.g., how humans respond to different signals or threats—researchers may be able to justify recruiting more easily accessible convenience samples (Goldblum, Reddie, and Reinhardt, 2019). Indeed, a growing body of research suggests convenience samples often hold preferences similar to those of more representative or elite samples (Berinsky et al., 2012; Kertzer, 2020). Ideally, researchers should recruit samples that are more representative of the target population of interest whenever possible. However, given both the challenges of elite recruitment (Dietrich et al., 2021; Kertzer and Renshon, 2021) and important variations even within elite populations,

researchers should not rely on "eliteness" as a sufficient characteristic for player selection. Researchers should instead identify characteristics within both elite and convenience samples that might affect wargaming behaviors and interrogate data for the effect of these characteristics within wargame play.

One particular recruitment challenge for researchers is finding players to represent foreign decisionmakers in wargames that feature specific allied or rival actors. These participants should ideally have deep country knowledge on the actor they are asked to represent. This helps ensure their actions in the wargame remain in the realm of plausible decisions the actor might actually take. Yet even experts may mirror image their own frames of reference onto foreign actors (Jervis, 1976). To reduce this risk, researchers can attempt to recruit participants who are actually from the state they are asked to represent. Since this is not always possible, practitioner games often rely on regional experts, including academics and foreign service officers. Or, wargame designers could provide non-regional experts playing foreign actors with detailed pre-game preparatory materials or even a rule book that spells out plausible strategies or doctrines the foreign actor might follow. To be sure, expert players do not possess a crystal ball for foreign crisis behavior, but that is not the purpose of games. Designers should strive for realism, not prediction. When analyzing data, researchers must acknowledge how these recruitment challenges might affect participant behavior.

The number of participants is influenced by whom the players represent in a game and, therefore, how teams are constructed. This choice should be informed by whether the research question focuses on decisions made by a group or the role of specific individuals. In some games, for example, the research question asks about the role of organizations or groups—such as the interaction of military commands—and therefore requires enough players to emulate the functions of those organizations. In others, players represent specific roles—such as a president or cabinet

minister—or more abstract "officials" with undefined positions. When deciding how to devise these teams and whether to assign specific roles, researchers should consider how the research question affects the types of decisions players are asked to make in the game (Bartels, McCown, and Wilkie, 2013: 42–46). For instance, a game studying a response to a nuclear attack likely needs players representing multiple agencies, not just the defense establishment.

Game design also affects sample size. Experimental games with multiple treatment games, for instance, typically need more participants than less complicated games. Similarly, games featuring multiple sides or simulating detailed organizational processes generally require more players than one-sided or highly abstract games. There is no hard and fast rule about the ideal number of participants, but games should include enough players to allow for the interaction that distinguishes wargames from other research approaches.

Rules: Moves, Sides, Adjudication

Researchers must next determine the rules that define their game's structure—how many moves, sides (i.e., teams), and how much adjudication will the game include? First, in order to determine how many rounds (i.e., moves) the game requires, scholars should ask whether they are interested in one-off decisions (for instance, a choice to retaliate or not) or the result of multiple decisions (for instance, protracted crises or shifts of power over time). Additional rounds can enhance realism by introducing tangible consequences, but can also diminish control over confounding factors, particularly in experimental games that feature multiple parallel iterations.

Researchers must also make decisions about sides (i.e., how many teams play the game). A one-sided game may be sufficient for answering questions that are not contingent on another actor's immediate reaction—for instance, what is my immediate response to a terrorist attack?

These games require fewer participants and a less complicated adjudication process. In contrast, for research questions contingent on the other side's reaction, like the effectiveness of a signaling strategy on deterrence, scholars should consider either one and a half-sided games, where the other's actions are scripted or played by game adjudicators, or two/multi-sided games that involve multiple teams of players. These offer greater dynamism and allow researchers to explore interaction between actors. The greater the number of sides, however, the less control researchers have between multiple game iterations.

Most games with more than one move will require adjudication by game organizers. This "refereeing" of outcomes between rounds can affect how subsequent rounds unfold. In projects with multiple game iterations, this can introduce variation across games, resulting in games that are no longer directly comparable. In some cases, cross-game differences might be useful to researchers—for instance, to study how variation in the early stages of a crisis generates divergent downstream effects. Yet, these differences can introduce confounders that make it difficult to isolate the effects of additional manipulations introduced after the first round.

Scholars can draw from a range of adjudication techniques depending on their research goals. Researchers interested in comparing large numbers of games might use formulaic adjudication—like probability tables or randomly generated outcomes. This approach allows for standardized rules across games; however, it can diminish realism. Scholars hoping to maximize player buy-in may choose free play adjudication where experts determine outcomes based on subject matter knowledge. This may create a more dynamic game for players that bolsters ecological validity, but can introduce adjudicator bias, making it difficult to replicate adjudication across multiple games, and increasing the number of adjudicators needed to field games. Free play also increases stochasticity, which can reduce comparability across multiple iterations of a game.

Scenario Design

As with survey and lab experiments, researchers must balance control and realism in their scenario designs to construct a practical yet ecologically valid test environment. This often entails making tradeoffs between abstraction and detail when deciding how much information to provide about the scenario and environment. Wargames must be sufficiently realistic to capture elements of real-world decision-making and be ecologically valid, while simultaneously simple enough to answer research questions (Mutz, 2011: 65). For instance, how much background should participants receive about the lead-up to the crisis? How much information is available to participants about the opposing team's capabilities and intentions?

One critical design choice is whether to name actual countries in the scenario (Dafoe, Zhang, and Caughey, 2018). On the one hand, identifying real states may create a more realistic scenario that subsequently influences decision-making. This realism, however, could lead policymakers to refrain from participating in wargames for fear of revealing classified information. Or, participants might bring in biases about those countries. On the other hand, using fictional or unnamed states might increase participation by national security practitioners but limit the inferences that can be drawn from findings.

In general, scholars should lean toward realism and specificity for research questions about particular cases (e.g., how might the U.S. respond to an Iranian-sponsored cyberattack) and make more abstract scenario choices for broader questions that are applicable across a wide swath of cases (e.g., are cyberattacks viewed differently than conventional attacks). To be sure, overly abstract vignettes may lead participants to make assumptions that could diminish researcher control relative to scenarios that offer more contextual detail. While recent studies suggest the

tradeoff between abstraction and detail might be overstated, researchers should remain cognizant of these issues when designing wargame scenarios.[10]

Data Collection and Analysis: Capturing Motivations, Interactions, and Decisions

Finally, researchers must also develop a strategy to collect and analyze data generated during wargames. Wargaming data can be divided into two types: outcome and deliberative. Outcome data identifies decisions players make in a game, often captured in moves, response plans, or other formal data inputs. Outcome data is generally easier to collect than deliberative data, which records participant interactions. Outcome data is, however, incomplete without deliberative data. Deliberative data sheds light on how and why decisions were made, which can help researchers explore the microfoundations of theories. Rich deliberative data can be process traced to understand how ideas were raised, how players reacted, and how teams came to decisions. Together, outcome data and deliberative data can provide explanations for how and why phenomenon occur, vice the more probabilistic assessments of what could occur that are common features of much experimental IR research.

In the ideal case, researchers would record all participant interactions and decisions verbatim. In their online wargame, Goldblum et al. (2019) accomplish this by digitally capturing player decisions and chat messages between participants. Researchers could also video or audio record wargames to capture participants' tone and body language. Digital collection, however, is not always feasible. Participants may not consent to recording, wargames are frequently conducted in facilities where electronic devices are prohibited, and ambient noise and crosstalk can make

[10] For an in-depth discussion of navigating the tradeoffs between abstraction and detail, see Brutger et al. (2020). Using survey experiments, they find relatively few tradeoffs between abstraction and detail, a finding that could be validated and further explored using wargames.

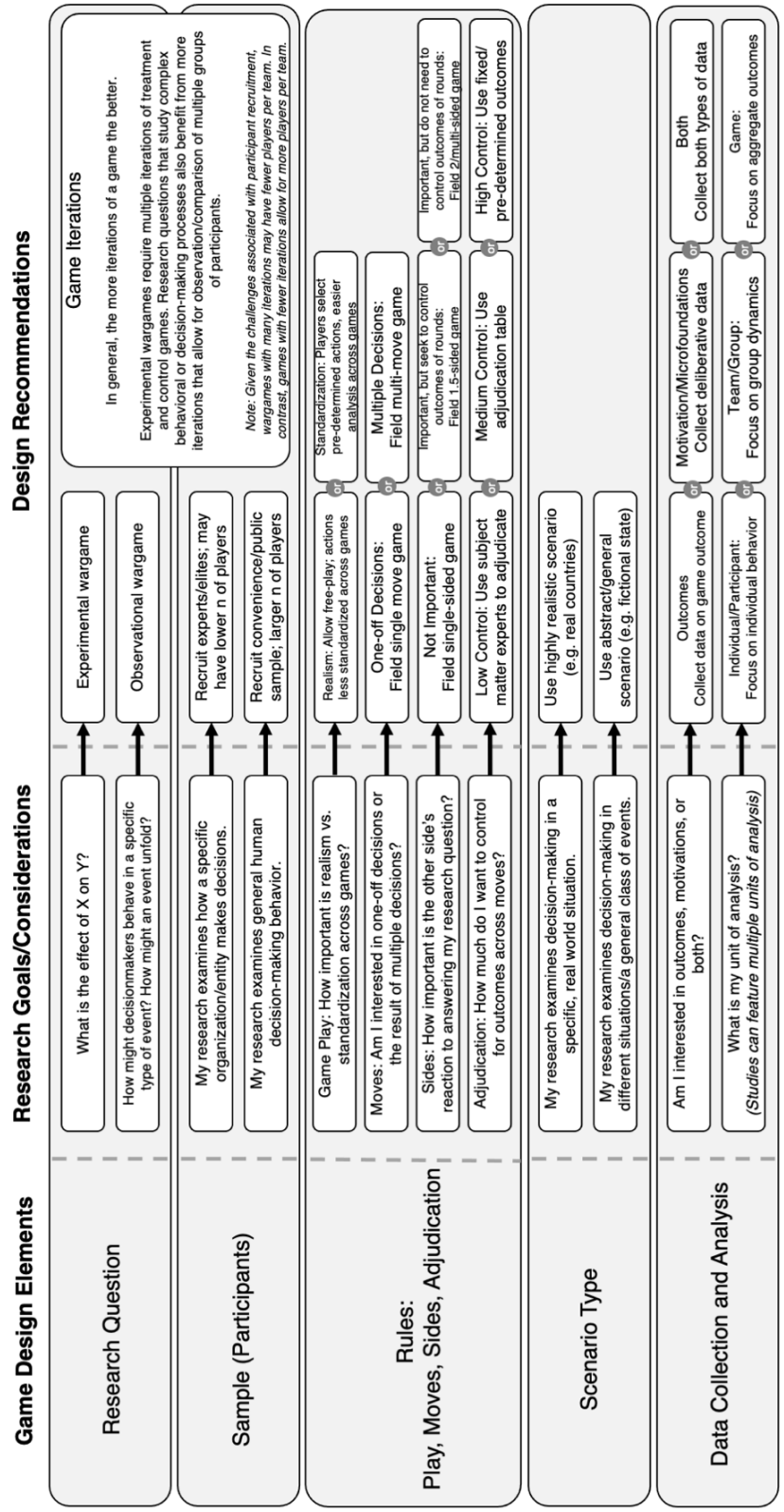
recording difficult. As a result, researchers often rely on research assistants to take notes on deliberations and to manually record team decisions, a process common in practitioner games.

Data collection by humans, however, is an inherently biased process. Because of their backgrounds, or because of the speed of discussions, notetakers will write down certain observations and omit others (Emerson, Fretz, and Shaw, 2011: 13). Additionally, data collection may generate a Hawthorne effect, in which participants alter their behavior because they are under observation (Wickström and Bendix 2000). To mitigate these risks, researchers can assign multiple research assistants to observe each game, allowing for triangulation while trying to make notetakers as unobtrusive as possible. To accurately capture game outcomes, researchers can instruct participants to submit forms that identify final decisions. These forms might also ask participants to list options they considered and to briefly explain why they selected the action they did, generating written data on the participants' perceptions of their own decision-making process. Finally, researchers may run post-game interviews or surveys either at the team or individual level to solicit information on the logics that guided decision-making. To gauge ecological validity and to improve future games, these interviews and surveys might also ask participants to describe how well the wargames simulated real-world decision-making environments.

When designing and analyzing wargames, researchers must determine the unit(s) of analysis. Units should typically be situated at the same level of analyses as the hypotheses under study (Gerring, 2012: 90–91). If the theory being tested is concerned with individual-level beliefs or behaviors (e.g., about internalized norms), researchers might use the wargame player as the unit of analysis. This allows for within game analysis to assess how a player's background or affiliation shapes her behavior. Similarly, projects using wargames to study theories of group dynamics might treat the team as a unit of analysis, allowing researchers to account for the mediating role of the

group in linking individual beliefs to team behavior. If, however, the theory deals with national security decision-making more broadly, researchers may consider using the game as the unit of analysis. Some studies, particularly those involving several game iterations might include multiple units of analysis that make both within and cross-game comparisons.

Figure 1: Game Design Recommendations



How might this work in practice? To demonstrate the wargame design process, we walk through how researchers designed and fielded the International Crisis Wargame Series (Schechter et al, 2021). The researchers began with a research question: "How do cyber operations affect nuclear stability?" Specifically, they wanted to assess whether cyber vulnerabilities and exploits within a rival's nuclear command, control, and communication networks (the primary independent variables) affected decisions on the use of force (the dependent variable). Based on the desire to study whether variation in vulnerabilities and exploits shaped decision-making, the researchers decided to use an experimental method. Researchers varied the availability of vulnerabilities and exploits across "control" and "treatment" games.

To generate a sample, the researchers initially sought elite expertise in foreign policy, cyber, and nuclear policy but later expanded their population to include students and convenience samples to explore how different types of expertise and demographic variables affected decision-making. In terms of rules, the researchers chose a relatively simple game play structure, with one move, one side, and a hypothetical scenario. This simple structure allowed the team to skip adjudication. As the researchers explained, "ICWG prioritized internal validity and control but also sought to iterate over time with a large and heterogeneous sample to create generalizable findings (Schechter et al, 2021: 6)."

Finally, the researchers collected both quantitative and qualitative data for analysis. Response plan "move sheets" completed during games captured group decisions, while surveys collected data to understand individual player motivations and explanations for actions taken in the game. As the researchers detail, "Although the Response Plan is developed collectively by the group, individual players may have different perceptions of the crisis or beliefs about the best

course of action. The surveys are intended to capture those perceptions and beliefs. Additionally, the survey attempts to capture how group dynamics may have influenced the completion of the Response Plans (Schechter et al, 2021: 10)."

When designing their game, the researchers made several explicit tradeoffs. To increase their total number of games and control play between iterations that took place over three years, the researchers made choices that may have decreased realism and immersion. For instance, using one-sided instead of multiple-sided games increased the number of game iterations, but traded off with realism. Similarly, using a one-move game instead of a multiple-move game helped enhance control, but perhaps limited the researchers' ability to explore more complicated questions of escalation. The aim of this illustration is not to identify one set of choices as right or wrong, but to help future scholars think deliberatively about the trade-offs inherent within wargames. In the following sections, we assess best practices and challenges associated with analyzing wargame-produced data.

ARCHIVAL WARGAMING DATA

In addition to fielding their own games, scholars may use data from historical games. In the late 1950s, two researchers did precisely what we discuss in this article: M.I.T. political scientist Lincoln Bloomfield and Harvard economist Thomas Schelling set out to design social scientific wargames (Bloomfield, 1984: 784-785). However, these innovations returned behind the shroud of government classification when, in 1961, the Joint Chiefs of Staff created a wargaming office and imported Schelling and Bloomfield's method (Bloomfield, 1963).

This data is increasingly available to international relations researchers (Emery, 2021). Today, declassified records of early American wargames can be found at presidential libraries, the

CIA's CREST archive, U.S. Declassified Documents Online, the RAND Corporation, and the MIT Archives. This data, from a golden age of senior policymaker participation in wargames run by social scientists, is ripe for political science theory testing. Data from more recent games run by government agencies, think tanks, NGOs, and scholars are also often publicly available, including those published by the RAND Corporation, Naval War College, Naval Postgraduate School, Harvard Belfer Center, and increasingly in periodical replication materials (Pauly, 2018; Schneider, 2017).

In either case—Cold War or modern games—how should scholars using archival game data think about its comparative advantages, its internal, ecological, and external validity, and biases? What can past games teach us about best practices for designing and analyzing future games?

Analyzing Archival Game Data

As with wargame design, a scholar's research question will inform their game selection. In some cases, researchers may seek to explain a particular historical policy or crisis decision, generate hypotheses to be tested against the historical record (Levine, Schelling, and Jones, 1991), or to generate historical counterfactuals. In other cases, scholars may seek to test theory.

Scholars trying to understand specific historical decisions can use wargames to study inputs to the policy process, what contingencies decisionmakers considered, or how agencies lobbied. One key reason scholars should pay more attention to practitioner wargaming is that many policymakers use games to inform planning and decision-making. Game selection will therefore be tied to the historical context in which games were played. For example, understanding nuclear policy or decision-making under the Johnson administration would benefit from games played by

members of his administration. In contrast, if a researcher is interested in the effects of a new technology on conflict or crisis, the political affiliation of the players may matter less but the expertise of the players may remain important.

If the research question pertains to more generalizable patterns of state behavior—such as when deterrence works, when leaders escalate, or how crises spiral spin out of control—selecting specific archival wargames records becomes akin to qualitative case selection. Open-ended political-military wargames, in which players have discretion over an array of statecraft tools, may be better for answering questions on or testing theories of the causes and consequences of war, or escalation and its limits. Operational wargames, in which players make tactical battlefield moves, may be more appropriate to test security studies theories on the conduct of war. However, scholars must ensure the games they select actually allow for the variation of interest, since some wargames can preclude certain player actions. For instance, researchers studying nuclear escalation using archival games should ensure the games they draw from do not prohibit the use of nuclear weapons.

Scholars must also ensure that the type of data presented in the archive is conducive to addressing their research question. For example, if scholars are interested in parsing the microfoundational mechanisms of theories, certain historical wargames with well-captured deliberative data are most appropriate. Compared to large-N games and even records from real-world events, small elite wargames typically offer exceptionally granular qualitative evidence of motivations and logics behind player choices. Many Cold War wargame records even capture transcripts of private discussions held after the game, sometimes moderated by a scholar. These participant reflections provide evidence generally not available after real-world events.

In rare instances, researchers may find games in the archives which were designed to ask research questions similar to their own. Even then, researchers must exercise caution in analyzing findings since wargames are still simulations of reality. For instance, do players view consequence-based outcomes as accurate representations of the real-world or as "game-isms" that limit ecological validity? Whether a participant thinks she is competing to win a game or to resolve a crisis can shape her behavior in ways that have meaningful implications for theory development and testing. When possible, therefore, researchers interested in theory testing and mechanisms should gather multiple archival wargames for cross-game meta-analysis that can reveal patterns of behavior despite significant differences between individual games in design, context, or players.

How *deliberative* and *outcome* data is collected, reported, and summarized varies across wargames, and the more the researcher knows about the design choices the better. Nonetheless, even with transparency in data generation, researchers must be attuned to several common biases. Much like interviews, diaries, or memoirs, archival wargaming records do not represent a complete and unbiased accounting of an event. Instead, the data provides good evidence which must then be evaluated and triangulated. To do so, scholars should recognize the difference between *raw* and *processed* wargaming data. Raw data includes quantitative and qualitative accounting of game actions or outcomes, transcripts of player discussion, or surveys or interviews of player experiences. Raw wargaming data is less likely to suffer from systemic bias than processed data, but also struggles with completeness (Bartels, 2020: 23-25). As discussed above, deliberative data, such as the transcripts of participant conversations are rarely a complete recording of discussions and reflect bias about what conversations observers thought were important (or audible).

In contrast to raw data, processed data presents a more finished picture of a wargame. Processed data includes game summaries or reports in which a game designer or administrator

documents game design, player actions, outcomes, conclusions, and policy recommendations. Because of its completeness, researchers may at first prefer processed data. However, processed data is more likely than raw data to exhibit some crucial biases. Game reports are often highly politicized documents that reflect the bureaucratic incentives of administrators which may not be transparent to scholars post-hoc. Indeed, tracking who was briefed (or not briefed) on a wargame's results has proven to be potent data for process tracers (Pauly, 2020). Greenstein and Burke (1989/1990: 576), for instance, found that the pessimistic conclusions of Vietnam wargames never made it to the Oval Office in the 1960s.

This bias stems from many practitioner games being "sponsored" by an agency or institution which uses game reporting to validate existing programs of record or doctrine or to justify budgets and authorities. It would be no surprise, for instance, to see an Air Force-sponsored wargame conclude that Congress should pay for more bombers. On the other hand, processed data is still valuable to researchers because it can reveal decisions made about wargame design: the scenario, adjudication, and subjects, as well as what original designer(s) sought to learn. Moreover, biases created by sponsors present an opportunity for scholars to ask research questions about the politics of organizational and bureaucratic rivalry in foreign policy development.

Other wargames may be biased not by their design or sponsorship but by their players. Researchers must endeavor to understand the relationships between and among hosts and players. Some may introduce acute Hawthorne observer effects that undermine the ecological validity of the test environment. Consider, for example, the Naval Postgraduate School's crisis games conducted with Indian and Pakistani players (Khan et al., 2016). At first blush these simulations seem an excellent opportunity to study escalation in South Asia. With American hosts in the room, however, the games risked becoming performative, with each regional nuclear power striving not

for victory but to cast the other as irresponsible. Thus, while some games serve important purposes of convening and educating policymakers, researchers who seek to test theory need to understand player incentives and potential observer bias.

Finally, whether raw or processed, much archival wargames data suffers biases of omission from the filter of declassification. This problem is not unique to wargames, and declassification bias does not affect wargame records more heavily than comparable qualitative sources such as the classified meeting minutes, policy reviews, and intelligence assessments frequently used in case study analysis (though defense organizations often choose to declassify games that support their budget or organizational priorities). Classification may also improve the quality of data if players speak more openly in private or anonymized classified recordings. Nonetheless, it is a problem for scholars if an outcome of interest in a wargame affects its declassification. American wargame records, even those in which the "Blue" (U.S.) team "lost," are able to be declassified, but their availability in archives beyond the 1970s is sparse. Scholars contributing to the wargaming research agenda must continue to file FOIA and Mandatory Declassification Review requests for documents.

AN AGENDA FOR WARGAMING RESEARCH

National security practitioners have long relied on wargames to inform policy. By drawing from existing games or by fielding their own, researchers can also use games to test IR theories—particularly to explore the microfoundations and mechanisms that underlie decision-making. As a tool of scholarly inquiry, wargames have the potential to better approximate the messiness of real-world decision-making and produce deeper insights about human decision-making than other commonly used methods. Researchers might use wargames as a standalone research design or

incorporate them into mixed-method research designs, where wargames help compensate for the shortcomings of other research approaches. At the core, wargames emphasize processes over outcomes by providing scholars with insight into why certain perceptions developed or decisions were taken. They let researchers explore how decisionmakers interact, strategize, process information, and perceive or misperceive their allies and adversaries.

Wargames provide an opportunity for researchers to parse the evidence in player deliberations, for instance, from intra-team dialogues about perceptions of a rival's signals and intent. This could yield insights on how humans understand their roles, the roles of others, or interpret the meaning and contexts of decisions. The constructed social environment of wargames also allows researchers to explore how characteristics such as gender, identity, hierarchy, and experience influence interactions within and across teams. Interactions during wargames can shed light on the mechanisms that underlie decisions, helping scholars to study a host of substantive topics. Indeed, many core concepts in IR, such as deterrence, crisis signaling, and the initiation of war are based on decision-making and interaction of policymakers. Accordingly, insights drawn from wargames can help researchers unpack theories in ways that go beyond empirical tests that focus solely on variation in outcomes.

Beyond using wargames for substantive research on IR theory, scholars might pursue research on the real-world impact of policy wargames. Future studies could, for instance, examine whether and when decisionmakers learn from government-fielded games. The Pentagon's SIGMA wargame series in the 1960s featured senior policymakers and foretold the quagmire in Vietnam (McDermott, 2002; Pauly, 2018), but its results were sidelined during the foreign policy-making process. How common are such dismissals of wargaming lessons? Conversely, selectively declassified wargames have played an outsized role in recent public budget and weapons

acquisitions discussions in the United States, with legislators even calling for more wargaming to inform decisions (Gallagher, 2020). How do wargames interact with organizational politics and how might the politicization of wargames resemble other political attempts to influence budgetary or policy choices? Researchers might couple archival wargame reports with process tracing and elite interviews to study these questions.

To effectively use wargames for substantive IR research, however, scholars must also examine whether and how various elements of wargame design and execution can affect their overall validity (internally, externally, and ecologically) and the conclusions that can be drawn from games. To this end, scholars might more deeply investigate the four propositions we presented in this article. This will help scholars better employ wargaming as complements to other research approaches.

First, future projects might study whether the immersive nature of wargames produces different behaviors than less immersive approaches. For instance, do subjects use different decision-making logics or invoke the same heuristics when participating in wargames than they do when completing surveys? What does this mean for ecological and external validity of wargame findings? And, what can this tell us about the types of questions that scholars can tackle using wargames?

To do this, researchers could turn to archival gaming data to investigate the implications of immersion. A survey of 77 players in political-military wargames held at M.I.T. between 1958 and 1964 found that 64.9% reported an "extreme" or "intense" degree of emotional involvement. Schelling recalled that participants "virtually lived" wargames and that it was difficult to spend so many hours in a scenario "without its beginning to seem either real or as one that could be real" (Department of Defense 1966, D3). But while these games may have created immersive and

realistic environments, some players also reported an inclination to behave aggressively (Barringer and Whaley 1965: 440).

If immersion in wargames bolsters ecologically validity, we might expect to see player behavior in games paralleled in the records of real-world crisis decision-making. Archival wargames offer several anecdotal examples—Vietnam, Cuba, Berlin Crises—where participant behavior in games and actual crises were similar. Short of such historical validation, researchers might ask elite participants days, weeks, or years later whether their wargaming experiences informed their real-world decision-making.

There is some evidence for such influence on players after they have left the gaming environment. In the same M.I.T survey described above, 56% of players who were "engaged in policy planning, formulation, or implementation" could recall an instance in which their wargame experience had been of practical value in their job. While contemporary data is scarcer, some players have recalled profound effects. Condoleezza Rice, for instance, reported that as National Security Advisor on September 11, 2001, she thought to notify Moscow of U.S. military alerts and explain to friends and foes alike "that the United States has not been decapitated" based on her experience with misperception and escalation during Cold War crisis simulations (Rice and Zegart, 2018: 178). Similarly, former Deputy Secretary of Defense Robert Work and Under Secretary of Defense for Intelligence Michael Vickers recount the pivotal role that a series of future wargames designed by the Office of Net Assessment in the 1990s and early 2000s played in developing strategies and weapons procurement almost twenty years later (Krepinevich and Watts, 2015).

Beyond analyzing archival data, researchers might design and field mixed-method projects featuring wargames conducted alongside alternative methods within a parallel research study. This would allow researchers to assess how immersion affects participant behavior and decision-

making. Recent projects that have fielded parallel survey experiments and wargames to address the same research question provide a useful starting point (Reddie et al., 2018). Schneider et al., 2021, for instance, finds that participants immersed in virtual wargames demonstrate far higher comprehension of wargame vignettes than survey experiment respondents provided with the identical scenario and preparatory reading materials. Indeed, 97.5% of wargame participants answered a scenario comprehension question correctly, compared to just 73% of survey experiment respondents. Additional research might also more systematically explore immersion by varying the structure and setting of wargames. Researchers could, for instance, vary the length of wargames (e.g., hours versus days) or the physical settings in which they are held, and assess whether there are changes in decision-making processes or outcomes.

Second, additional research might help scholars better understand whether and how experts behave differently than non-experts during wargames. This line of research would directly contribute to the lively methodological debate about the utility of different types of samples in modern empirical IR research. On the one hand, some scholars suggest that using highly representative, but small, expert samples to play fewer wargames limits the degree to which findings can be generalized (Reddie et al. 2018). They believe large—often convenience—samples playing multiple game iterations, allow for statistical analysis that overcomes generalizability issues and enables replication. Other scholars and practitioners, however, believe convenience samples limit the conclusions of games (Oberholtzer et al, 2019). Specifically, non-experts may lack the technical or policy knowledge needed to make realistic decisions that mirror those that might play out in the real-world. Political scientists using other research approaches have long debated whether convenience samples are adequate proxies for more representative, expert samples (Hyde, 2015; Dietrich et al., 2021). Some studies find divergence between the behavior

of convenience and elite samples (Mintz, Redd, and Vedlitz, 2006; Pauly, 2018), while others find congruence between elite and non-elite preferences (Kertzer, 2020).

Again, insights from archival games offer opportunities to explore whether our second proposition affects what we can learn from wargames. The striking 'eliteness' of their players makes some archival wargames excellent points of comparison. Schelling, for instance, directed wargames in which the Chairman of the Joint Chiefs of Staff, the Chief of Staff of the Army, and the Attorney General participated (Schelling and Ferguson, 1988). These games can be compared to other archival games played by Pentagon guests including celebrities, journalists, and business executives. Researchers interested in contributing to methodological debates about research samples might consider fielding wargames that compare decision-making processes in wargames played by experts with those from identical games played by non-experts.

Third, future research might contribute to studies on group dynamics by exploring whether and how group interaction affects decision-making and behavior during wargames. If wargames are ecologically valid, the lessons gleaned from games, should be applicable to actual decision-making contexts. For instance, do participants worry teammates will judge them for what they say and do? Are groups more likely to mitigate or amplify individual risk propensities? How does team composition affect group dynamics? Transcripts from both researcher fielded and archival games often include players justifying their choices to one another. This data could be parsed for language associated with hierarchical, deferential, combative, emotional, or gendered decision-making, and researchers could explore how different team compositions affect dynamics across multiple iterations of a given game.

Recent projects that field wargames alongside less interactive synthetic data generating processes offer a starting point for this type of future research (Reddie et al., 2018). Lin-Greenberg

(2020), for instance, finds that decisions of wargaming teams frequently evolve during deliberations among participants. In some cases, participants change their position after discussing an issue with teammates. Or, participants holding a particular view may simply be outnumbered by other members of the team and defer to the majority position. The dynamic nature of wargame deliberations therefore provides an advantage over less interactive research approaches that often only capture individual-level preferences at a specific moment in time, leaving researchers with less understanding of how ideas develop and evolve.

Fourth, scholars should assess whether and how consequence-based outcomes shape behavior and decision-making during wargames. Wargaming experts struggle to delineate between consequences that accurately reflect real-world decision-making and those that are "game-isms" that might limit the conclusions that can be drawn from games. Do players take greater risks in games than they would during actual crises? Do participants act honestly, or do they take actions to support their employers' institutional interests? A similar debate on whether incentives and rewards affect behavior during simulations and experiments remains unresolved (Karagozoglu and Urhan, 2017; Andersen et al., 2011). To address these questions researchers might turn to archival games to see if elites took similar risks during wargames as in real-world crises. Or, researchers could assess whether participant behavior changes as stakes vary across multiple wargames.

The use of wargaming for international relations scholarship is still in the early stages of a renaissance, but we believe the approach has significant potential for researchers seeking to understand how foreign policy and national security decisions are made. As scholars explore the benefits and limitations of wargaming as a tool of inquiry, we see the exciting possibilities of wargaming research helping to tackle otherwise difficult to address theory and policy-motivated questions.

References:

Abbasi M, Kumar S, Augusto J, Filho A, and Liu H (2012) Lessons learned in using social media for disaster relief-ASU crisis response game. *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*: 282-289.

Alker, H (1985) From Quantity to Quality: A New Research Program on Resolving Prisoner's Dilemmas. Paper given at APSA, New Orleans, August 1985. https://dornsife.usc.edu/assets/sites/556/docs/from_quantity_to_quality.PDF

Andersen S, Ertaç S, Gneezy U, Hoffman M, and List J (2011) Stakes matter in ultimatum games. *American Economic Review* 101(7): 3427-39.

Banks M, Groom A, and Oppenheim A N (1968) Gaming and simulation in international relations. *Political Studies* 16(1): 1-17.

Barabas J and Jerit J (2010) Are survey experiments externally valid? *American Political Science Review* 104(2): 226-242.

Barma N, Durbin B, Lorber E, and Whitlark R (2016) 'Imagine a World in Which:' Using Scenarios in Political Science. *International Studies Perspectives* 17(2): 117-135.

Barringer R and Whaley B (1965) The MIT Political-Military Gaming Experience. *Orbis* 9(2).

Bartels E (2020) Building Better Games for National Security Policy Analysis: Towards a Social Scientific Approach. RAND Corporation.

Bartels E, McCown M, and Wilkie T (2013) Designing Peace and Conflict Exercises: Level of Analysis, Scenario, and Role Specification. *Simulation & Gaming* 44(1): 36–50.

Baumeister, R and Vohs, K (2007) Ecological Validity. In: *Encyclopedia of Social Psychology*, Thousand Oaks, CA: SAGE Publications.

Berinsky A, Huber G, and Lenz G (2012) Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* 20(3): 351-368.

Bloomfield L (1960) Political Exercise II-Working Papers 1960. Box 9. Bloomfield Papers. MC 326. Institute Archives and Special Collections, MIT Libraries, Cambridge, Massachusetts.

Bloomfield L (1984) Reflections on Gaming. *Orbis*, 28(4): 783-790.

Bloomfield L (1963) Four Political-Military Exercises. Box 9, Bloomfield Papers, MC 326, Institute Archives and Special Collections, MIT Libraries, Cambridge, Massachusetts.

Bloomfield L and Whaley B (1965) The Political-Military Exercise: A Progress Report. *Orbis* 8(4): 854-869.

Brody R (1963) Some systemic effects of the spread of nuclear weapons technology. *Journal of Conflict Resolution* 7(4): 663-667.

Brutger R, Kertzer J, Renshon J, Tingley D and Weiss C (2020) Abstraction and Detail in Experimental Design. *American Journal of Political Science* (forthcoming).

Brunswik E (1947) Systematic and Representative Design of Psychological Experiments. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*: 143–202.

Caffrey M (2019) On Wargaming. *The Newport Papers*. 43

Camerer C (2011) *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Chu J and Recchia P (2021) Does Public Opinion Affect the Preferences of Foreign Policy Leaders? Experimental Evidence from the UK Parliament. *Journal of Politics* (forthcoming).

Colbert E, Sullivan D, and Kott A (2017) Cyber-Physical War Gaming. *Journal of Information Warfare* 16(3): 119-133.

Dafoe A, Zhang B, and Caughey D (2018) Information Equivalence in Survey Experiments. *Political Analysis* 26(4): 399–416.

Daniel J and Musgrave, P (2017) Synthetic Experiences. *International Studies Quarterly* 61(3):503–516.

Department of Defense (1966) Final Report of NU I and II-66, Two Interagency Politico-Military Games Played by Officials of the Executive Branch during 1/11–2/8/66. February 28, doc. no. GALE|CK2349234646, U.S. Declassified Documents Online.

Dietrich S, Hardt H, Swedlund HJ (2021) How to make elite experiments work in International Relations. *European Journal of International Relations* 27(2): 596-621.

Dorn A, Walter S, and Pâquet S (2020) From Wargaming to Peacegaming: Digital Simulations with Peacekeeper Roles Needed. *International Peacekeeping* 27(2): 289-310.

Emerson R, Fretz R, and Shaw L (2011) *Writing Ethnographic Fieldnotes.* University of Chicago Press.

Emery J (2021) Moral Choices Without Moral Language: 1950s Political-Military Wargaming at the RAND Corporation. *Texas National Security Review* 4(4).

Findley M, Kikuta K, and Denly M (2020) External Validity. *Annual Review of Political Science* (forthcoming).

Fiorina M, and Plott, C (1978) Committee decisions under majority rule: An experimental study. *American Political Science Review* 72(2): 575-598.

Gallagher M (2020) "Bringing Congress to the [Wargaming] Table for a Bigger, Better Navy." *War on the Rocks*. October 20. https://warontherocks.com/2020/10/bringing-congress-to-the-wargaming-table-for-a-bigger-and-better-navy/.

Gerring J (2012) *Social Science Methodology*. Cambridge University Press.

Goldblum B, Reddie A, and Reinhardt J (2019) Wargames as Experiments: The Project on Nuclear Gaming's SIGNAL Framework. *Bulletin of the Atomic Scientists*. May 29. https://thebulletin.org/2019/05/wargames-as-experiments-the-project-on-nuclear-gamings-signal-framework/.

Gouvier W, Barker A, and Musso M (2014) Ecological validity. *Encyclopedia Britannica*. https://www.britannica.com/science/ecological-validity

Greenstein F and Burke J (1989/1990) The Dynamics of Presidential Reality Testing: Evidence from Two Vietnam Decisions. *Political Science Quarterly* 104(4): 557-580.

Hamman J, Weber R, and Woon J (2011) An experimental investigation of electoral delegation and the provision of public goods. *American Journal of Political Science* 55(4): 738-752.

Hermann C and Hermann M (1967) An Attempt to Simulate the Outbreak of World War I. *American Political Science Review* 61(2): 400-416

Hirst A (2020) States of play: evaluating the renaissance in US military wargaming. *Critical Military Studies*: 1-21.

Huckfeldt R, Pietryka M, and Reilly J (2014) Noise, bias, and expertise in political communication networks. *Social Networks* 36: 110-121.

Hyde S (2015) Experiments in International Relations: Lab, Survey, and Field. *Annual Review of Political Science* 18 (1): 403–24.

Jensen B and Banks D (2018) *Cyber Operations in Conflict: Lessons from Analytic Wargames.* Report, Center for Long-term Cyber Security.

Jensen B and Valeriano B (2019) Cyber Escalation Dynamics: Results from War Game Experiments. International Studies Association. Toronto, Canada.

Jervis R (1976) *Perception and Misperception in International Relations*. Princeton University Press.

Johnson D, McDermott R, Barrett ES, Cowden J, Wrangham R, McIntyre MH, Peter Rosen S (2006) Overconfidence in Wargames: Experimental Evidence on Expectations, Aggression, Gender, and Testosterone. *Proceedings of the Royal Society* 273(1600): 2513–2520.

Karagözoglu E and Urhan U (2017) The effect of stake size in experimental bargaining and distribution games. *Group Decision and Negotiation* 26(2): 285-325.

Kerr N and Tindale R (2004) Group performance and decision making. *Annual Review of Psychology* 55:623-655.

Kertzer J (2017) Microfoundations in International Relations. *Conflict Management and Peace Science* 34(1):81–97.

Kertzer J (2020) Re-Assessing Elite-Public Gaps in Political Behavior. *American Journal of Political Science* (forthcoming).

Kertzer J and Renshon J Experiments and Surveys on Political Elites. *Annual Review of Political Science* (forthcoming).

Khan F, Wueger D, Giesey A, and Morgan R (2016) South Asian Stability Workshop 2.0: A Crisis Simulation Report. *Report No. 2016-001*, Naval Postgraduate School.

Kihlstrom J (2021) Ecological Validity and "Ecological Validity." *Perspectives on Psychological Science* 16(2):466-471.

Krepinevich A and Watts B (2015) *The Last Warrior*. Basic Books.

Levine R, Schelling T, and Jones W (1991) *Crisis Games 27 Years Later*. RAND Corporation.

Lillard J (2016) *Playing War: Wargaming and US Navy Preparations for World War II.* Potomac Books.

Lin-Greenberg E (2020) Wargame of Drones: Remotely Piloted Aircraft and Crisis Escalation. Working Paper. https://papers.ssrn.com/abstract=3288988.

McDermott R (2002) Experimental Methodology in Political Science. *Political Analysis* 10(4): 325–42.

McDermott R, Cowden J, and Rosen S (2008) The Role of Hostile Communications in a Crisis Simulation Game. *Peace and Conflict: Journal of Peace Psychology* 14(2): 151-168.

McGrady E (2019) Getting the Story Right about Wargaming. *War on the Rocks*. November 18. https://warontherocks.com/2019/11/getting-the-story-right-about-wargaming/

Miller A (2020) The Information Game: Police-Citizen Cooperation in Communities with Criminal Groups. PhD Dissertation, M.I.T.

Mintz A, Redd S, and Vedlitz A (2006) Can We Generalize from Student Experiments to the Real-world in Political Science, Military Affairs, and International Relations? *The Journal of Conflict Resolution* 50(5): 757–76.

Mintz A and Wayne C (2016) *The Polythink Syndrome.* Stanford University Press.

Mutz D (2011) *Population-Based Survey Experiments*. Princeton University Press.

Oberholtzer J, Doll A, Frelinger D, Mueller K, and Pettyjohn S (2019) Applying Wargames to Real-World Policies. *Science* 363(6434).

Oriesek D and Shwarz J (2008) *Business Wargaming*. Ashgate.

Pauly R (2018) Would U.S. Leaders Push the Button? Wargames and the Sources of Nuclear Restraint. *International Security* 43(2): 151–192.

Pauly R (2020) What to Do When Predicting Pandemics. *Foreign Policy.* September 11.

Perla P (1990) *The Art of Wargaming: A Guide for Professionals and Hobbyists*. Naval Institute Press.

Perla P and McGrady E (2011) Why wargaming works. *Naval War College Review 64*(3):111-130.

Pettyjohn S (2019) The Promise and Pitfall of Wargames: Differentiating Good from Bad Games. Working Paper.

Reddie A, Goldblum B, Lakkaraju K, Reinhardt J, Nacht M, and Epifanovskaya L (2018) Next-Generation Wargames. *Science* 362(6421):1362–64.

Rice C and Zegart A (2018) *Political Risk.* Twelve.

Saunders E (2017) No Substitute for Experience: Presidents, Advisers, and Information in Group Decision Making. *International Organization* 71(1): 219-247.

Schechter B, Schneider J, and Shaffer R (2021) Wargaming as a Methodology: The International Crisis Wargame and Experimental Wargaming. *Simulation & Gaming* (forthcoming).

Schelling T (1987) The Role of Wargames and Exercises, in Carter A, Steinbruner J, and Zraket C, eds., *Managing Nuclear Operations*. Brookings Institution Press.

Schelling T and Ferguson A (1988) Remarks at John F. Kennedy School of Government, Harvard University, Cambridge, Massachusetts, November 22.

Schmuckler M (2001) What is Ecological Validity? A Dimensional Analysis. *Infancy* 2(4): 419-436.

Schneider J (2017) Cyber Attacks on Critical Infrastructure: Insights from War Gaming. *War on the Rocks*. July 26. https://warontherocks.com/2017/07/cyber-attacks-on-critical-infrastructure-insights-from-war-gaming/.

Schneider J, Schechter B, and Shaffer R (2021) Cyber Operations and Nuclear Use: A Wargaming Exploration. Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3956337

Sepinsky J (2021) Is it a wargame? It doesn't matter: rigorous wargames versus effective wargaming. *War on the Rocks.* February 24. https://warontherocks.com/2021/02/is-it-a-wargame-it-doesnt-matter-rigorous-wargames-versus-effective-wargaming/.

Smith J and Bell P (1992) Environmental concern and cooperative-competitive behavior in a simulated commons dilemma. *The Journal of social psychology* 132(4): 461-468.

Schuurman P (2019) A Game of Contexts: Prussian-German Professional Wargames and the Leadership Concept of Mission Tactics 1870–1880. *War in History*: 1-21.

Tomz M, Weeks J, and Yarhi-Milo K (2020) Public Opinion and Decisions About Military Force in Democracies. *International Organization* 74(1): 119-143.

Wang R, Fishkin J, and Luskin R (2020) Does Deliberation Increase Public-Spiritedness? *Social Science Quarterly* 101(6): 2163-2182.

Wickström G and Bendix T (2000) The 'Hawthorne Effect' — What Did the Original Hawthorne Studies Actually Show? *Scandinavian Journal of Work, Environment & Health* 26(4): 363–67.

Williams H and Drew A (2020) Escalation by Tweet: Managing the New Nuclear Diplomacy. *Centre for Science and Security Studies*: Kings College London.

Wilson A (1968) *The Bomb and the Computer: Wargaming from Ancient Chinese Mapboard to Atomic Computer*. Delacorte Press.

Wong Y, Bae S, Bartels E, and Smith B (2019) *Next Generation Wargaming for the US Marine Corps*. Rand Publications.