REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED *(From - To)*
03-02-2024	Final Report	*6/3/2019 - 8/31/2023*

4. TITLE AND SUBTITLE

Empirical Analysis for Meeting Great Power Challenges

5a. CONTRACT NUMBER
N00014-19-1-2466

5b. GRANT NUMBER
GRANT12695347

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Stephen Biddle, Eli Berman, Eric Min, Laura Samotin, Meyer Thalheimer, John Severini

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Columbia University School of International and Public Affairs,
420 W 118th Street, New York, NY 10027;
University of California at San Diego Department of Economics,
9500 Gilman Drive # 0508, La Jolla, CA 92093-0508;
University of California Los Angeles Department of Political Science,
4289 Bunche Hall, Los Angeles, CA 90095-1472

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

ONR HUMAN & BIOENGINEERED SYSTEMS
875 N. Randolph Street
Arlington VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Cleared for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The project is intended to support the National Defense Strategy by identifying factors that will shape the ability of great powers such as China or Russia to use new technology and new strategies effectively. It does this by exploiting new data sources and novel research methodologies to build stronger theories of combat outcomes in modern great power warfare. From these findings, we draw implications for U.S. policies to deter such rivals, and to project power successfully if deterrence fails.

15. SUBJECT TERMS
Naval warfare, submarine warfare, technology, numerical preponderance, skill, regime type, GDP per capita, human capital, historical data, machine reading

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	None	13	Stephen Biddle
Unclassified	Unclassified	Unclassified			19b. TELEPHONE NUMBER 212-854-1496

From: Stephen Biddle

Re: Final Report, Empirical Analysis for Meeting Great Power Challenges

Date: 3 February 2024

This memo presents the final report for work under the Minerva Initiative grant "Empirical Analysis for Meeting Great Power Challenges," award number N00014-19-1-2466.

Major Goals

The project is intended to support the National Defense Strategy by identifying factors that will shape the ability of great powers such as China or Russia to use new technology and new strategies effectively. It does this by exploiting new data sources and novel research methodologies to build stronger theories of combat outcomes in modern great power warfare. From these findings, we draw implications for U.S. policies to deter such rivals, and to project power successfully if deterrence fails.

In particular, the project consists of two sub-tasks that take different empirical strategies to these ends.

The first sub-task exploits and expands the new NAVBATTLE dataset. These NAVBATTLE data cover all 573 great-power surface naval battles fought between the dawn of the Age of Sail in 1649 and the end of the last major great-power war at sea in 1988, and include observations on losses in ships, personnel, and tonnage; codings for victories, defeats, and draws; battle duration; material balances in ships, guns, and displacement; battle dates and geo-spatially coded locations; and identification of commanders and nationalities. The project has expanded these data, and merged them with a series of available cross-sectional time-series datasets on the internal political, social, and economic makeup of states. It then analyzed the merged data to identify correlates of successful performance at sea; to distinguish more from less significant contributors; and to draw implications for scholarship and current U.S. defense policy. The second phase of this sub-task used the results to select a small sample of theoretically critical battles for in-depth study.

The second sub-task is a pilot study designed to assess the feasibility of large-scale military effectiveness data collection via machine coding of archival records, and to produce an initial dataset for exploratory empirical analysis. Novel methodologies for machine coding have great promise for translating qualitative source material into quantitative data in a rigorous way. The efficiency of any given machine coding effort, however, is a function of the specific format, content, and consistency of the records to be coded. The purpose of this project is to assess whether the war diaries and other archival materials available are sufficiently consistent in format and content to permit cost-effective coding – and if so, to compile a pilot dataset for a substantively meaningful subset of the available records as a proof of principle.

This pilot study began with archival records on the conduct of all 6,272 U.S. submarine attacks on Japanese shipping during World War II. The goal was to machine-code narrative records on the outcomes and circumstances of these attacks, and to machine-code the narrative service records of all 421 submarine commanders who led those attacks. The result is a new dataset covering the tonnage sunk; US losses; tactical and material circumstances; and training experience, educational experience, personality traits, age, marital status, socio-economic status at entry, prior assignments, and service

performance of the US commanders involved. We then analyzed the resulting data statistically to explore causal interrelationships among these variables - and used the results to assess the viability of the approach for follow-on research to examine other combat domains.

This work has produced a body of research composed of separate, yet interrelated, journal articles and essays designed to provide deep understandings of the preceding topics and their implications for policy. In particular, we have completed a draft scholarly journal article presenting the findings of the large-*n* empirical analysis on naval warfare from phase I of the first sub-task; a draft scholarly journal article presenting the findings of the in-depth case studies on naval warfare from its second phase; and a published paper in the Empirical Study of Conflict project working paper series presenting the new data coding methodology and initial descriptive data analysis from the second sub-task.

We also plan to prepare policy-oriented briefings on the results for presentation to interested audiences in the policy community and Department of Defense, and to prepare scholarly presentations for interested academic audiences to increase the visibility of our findings among fellow researchers.

Accomplishments and Major Findings

All three deliverable papers are complete in draft and were delivered to the sponsor on August 31, 2023, together with a related paper that uses some of the same data to address other aspects of naval warfare and its implications for U.S. defense policy.

Victory at Sea: What Explains Effectiveness in Naval Warfare?

The large-*n* empirical paper is titled "Victory at Sea: What Explains Effectiveness in Naval Warfare?" by Stephen Biddle, Eli Berman, and Meyer Thalheimer. In it, we present and describe the expanded NAVBATTLE dataset, then quantitatively address four questions about the effectiveness of military technology at sea:

1) How important is an advantage in materiel? Our central finding here is that a heavier squadron of ships has a large advantage, though less so since the introduction of steam power in the 1850s. In the age of sail, doubling displacement predicts about four times more enemy casualties per friendly casualty. Since then, doubling displacement predicts only twice more enemy casualties per friendly casualty.
2) How valuable is superior technology? For battles after the introduction of steam power in the mid-19th century we use launch year as a proxy for technological sophistication. By this measure, a decade of technological advantage is worth about twice the displacement.
3) Do democracies have an advantage? We find that historically they have. Our modeling suggests that in the past, a democracy gap equal to that between the U.S. and China today has produced an advantage equivalent to about an 80 percent larger force. That said, our data do not enable us to disentangle empirically the institutional advantage of democratic political organization per se from the predicted advantages of higher GDP/capita, better human capital, or greater international trade – all of which cluster with democracy historically, but might not in the future.
4) How important was the transition to steam powered warships? Many see this transition as a revolution in military affairs; was it, and if so, in what sense? Our analysis suggests a more

complex picture, as the transition to steam was quickly followed by a sequence of other changes, both technological and societal, not all of which can be traced to the technological innovation of steam propulsion.

From these empirical findings we draw a series of implications for U.S. defense policy and international relations scholarship. China's GDP is widely expected to exceed America's in coming years, and this GDP growth has fueled a major expansion in the material size and sophistication of the Chinese navy. This growth is likely to continue, and could easily shift the material balance of naval power between the United States and the PRC in China's favor over time. China's political institutions, however, are radically less democratic than those of the United States, and those institutions are unlikely to become much more democratic any time soon. Hence the relative influence of material and political variables for naval power is a central question for understanding the future of the U.S.-China competition at sea. Material and nonmaterial variables both matter, of course, but their relative weight is an issue that could shape the military balance profoundly given the contrasting strengths and weaknesses of the United States and China as naval powers.

While our findings suggest that the marginal influence of democracy relative to material strength shrank after 1860, it is still strong enough to exert an important effect on dyads with very dissimilar political systems, such as the United States and the PRC. In fact, for the United States and the PRC in 2023, our statistical analysis implies that the difference in political systems might outweigh as much as a 2:1 disadvantage in aggregate tonnage, or a 17-year difference in the average age of the deployed fleets.

If so, this may be grounds for guarded optimism on the future of the naval competition in the Western Pacific. Other readings of the data could be less optimistic; in particular, the GDP/capita gap will likely continue to shrink in the future, and it is an alternative predictor of naval battle outcomes. Nevertheless, the empirical analysis above implies that structural U.S. advantages can, in principle, compensate for significant growth in Chinese material strength.

To realize this potential, however, requires prudent investments in the material size and sophistication of the U.S. Navy. It also requires attention to the personnel, training, and professional military education accounts that create and retain the proficient leaders and crews that American political institutions enable. The observed results in NAVBATTLE are the product of states that have usually striven to maximize their naval potential within the constraints of budgets, technology, and political institutions, *inter alia*. Variance in institutions, in particular, has been associated with important variance in naval performance – but these institutions create potential that must be realized with prudent investment and attentive policy making. It would be a serious mistake for the United States to allow the accounts that create proficient sailors and officers to atrophy in ways that leave the country short of the potential its institutions enable. Of course, there are important interactions between material investment and nonmaterial human performance – an undersized fleet that must operate at a tempo that exhausts crews and precludes training can reduce performance just as underinvestment in training or personnel can do. And there may be grounds for concern that current U.S. Navy operating tempos are exceeding what crews can sustain. But the answer probably does not lie in materiel acquisition alone. Recruiting, educating, training, and retaining skilled people is an important structural advantage for the United States in its competition with the PRC – but only if U.S. policies exploit this.

That said, these findings rest on statistical analysis of the past. Historical study is the only way to observe real warfare, but the nature of naval history creates constraints that cannot be entirely

escaped. The most recent battle that meets our inclusion criteria was the U.S. engagement with Iran in Operation Praying Mantis of 1988 – there have been no real-world examples of such naval battles available for study in the last 35 years. We thus cannot exclude the possibility that changes in technology or other features of naval warfare may have changed causal relationships in ways that no empirical analysis of actual warfare can detect.

Statistical analysis is also limited to variables whose values can be collected for large-*n* analysis. NAVBATTLE contains far more data than was previously available, but variables such as aggregate tonnage or the difference in average launch year are clearly imperfect measures of material naval strength. Navies with many small ships and a few large ones may have equal aggregate tonnage but different values in combat from those with the opposite makeup; navies with many combatants and few auxiliaries may have the same aggregate tonnage as others with fewer combatants and more auxiliaries but very different performance in battle. In any given year, states vary from one another in their technology base, and thus in the sophistication of the ships they launch in that year; Japan, for example, lagged behind the United States in the deployment of shipboard radar early in World War II. Average launch year is thus at best an approximation of naval technological sophistication.

That said, it is striking that several of these aggregate approximations perform so well as predictors of observed combat outcomes – and that they sometimes do so over such enormous spans of time. Aggregate squadron weight has clear shortcomings as a measure of naval strength, yet it is statistically significant at levels in excess of .01 as a predictor of victory and casualty ratios over fully three centuries of combat experience, and in itself explains 20-25 percent of the variance over that period. It is not an ideal measure, but it is an informative one all the same.

The dataset also lacks detailed information on crew training, shipboard procedures, task force organization, unity of command, officer talent, combat motivation in the age of steam, or a variety of other nonmaterial traits that are surely relevant to a squadron's performance in battle. New techniques for coding archival documents for large-*n* data collection, as addressed in our third deliverable described below, may offer increasing potential to measure and collect values for such variables in the future. But for now, this paper is restricted to studying the state-level socio-political attributes that, by hypothesis, create the enabling conditions for these more granular nonmaterial contributors to success in battle. Socio-political aggregates such as regime type, moreover, can overlook important outliers or exceptions to the general trend: Japan in World War II, for example, was an autocracy that atypically punched above its weight. Here, too, however, though the available measures are imperfect they nonetheless offer valuable insight: regime type, for example, is significant as a predictor of victory and casualty ratios over fully three centuries of combat experience at significance levels in excess of .01 for the pooled data and in excess of .05 for the age of steam data, notwithstanding its aggregate nature and the presence of occasional exceptions such as Imperial Japan. Better measures have the potential to improve our understanding and to sharpen our models, but the data we have are informative all the same.

Historical large-*n* analysis thus has limitations that must be kept in mind. But in exchange it offers a source of insight that cannot be obtained in other ways – and it enables the relative contribution of effects that are all plausibly relevant to be weighed in light of actual combat experience in ways that shed unique, if partial, light on contemporary policy concerns.

For scholars, our analysis challenges the validity of international relations scholarship that treats national power solely as a matter of material strength. Almost all empirical studies in international

security involve some treatment of military capability, whether as the focus or as a control. But this literature is overwhelmingly materialist in its operationalization of capability. Material variables clearly matter for naval warfare. And there is reason to believe that materiel is more influential for naval outcomes than for war on land. But nonmaterial variables matter for naval warfare, too. Theories and empirical studies that treat power solely through material proxies such as population, GDP, or steel production thus miss important aspects of international political reality. To explain the Pax Britannica, for example, solely as a matter of British material power is to miss the critical nonmaterial contributions to British naval dominance in the 19th century. The conduct of war, whether on land or at sea, is a social undertaking in which the decisions and actions of people – not just the weight of materiel – matter for outcomes. To understand the international politics of sea powers, and of the continental states with which they contend, it is thus important to consider more than just materiel, important as that is.

Technology, Behavior, and Effectiveness in Naval Warfare:
The Battles of Savo Island and Cape Saint George

The small-n case study paper is titled "Technology, Behavior, and Effectiveness in Naval Warfare: The Battles of Savo Island and Cape Saint George," by John Severini and Stephen Biddle. The paper explores two battles from the NAVBATTLE dataset with unusual leverage for testing ideas about material and non-material contributions to effectiveness at sea: the battles of Savo Island and Cape Saint George.

Savo Island, fought on the night of August 8-9, 1942, was a surface engagement between an Imperial Japanese Navy squadron and a combined force of American and Australian warships in waters off the island of Guadalcanal in the Solomons chain. Cape Saint George, fought after midnight on the night of November 25, 1943, was waged about 500 miles northwest of Guadalcanal near Bougainville, and pitted Japanese and American destroyer squadrons. In both actions, the Japanese enjoyed superior torpedoes but were outnumbered, outgunned, and without the critical technology of radar for search and fire direction. In both battles, U.S. and allied forces enjoyed numerical superiority and the critical technological advantage of radar. In material terms, both actions should have been U.S. victories – and especially so at Savo Island, where the allied advantage in tonnage over the Japanese exceeded forty percent. The outcomes, however, were radically different. Savo Island was a decisive victory for the Japanese, who sank four allied heavy cruisers and damaged another allied cruiser and two destroyers, while losing no ships themselves, suffering only moderate damage to two heavy and one light cruiser in exchange. In fact, Savo Island has been described as the U.S. Navy's second greatest defeat in its history, exceeded only by the Japanese attack on Pearl Harbor. Cape Saint George, by contrast, was a decisive U.S. victory in which the Americans sank three of the five Japanese destroyers committed and damaged another, while suffering no losses themselves, notwithstanding a material balance that was less advantageous than in the American defeat at Savo Island.

An in-depth focus on two battles enables a degree of detailed analysis of issues such as organization, leadership, or combat dispositions that are difficult to address with large-*n* modeling. Case method also allows detailed process tracing to help distinguish real causation from mere coincidence. But in exchange it poses an inherent risk of selection bias: an argument's success (or failure) in a small sample of cases might be an artifact of having chosen misleading or unrepresentative examples, making case

selection a critical issue for validity. Savo Island and Cape Saint George offer a number of important properties that help to mitigate this risk.

In particular, they offer an unusual degree of control for extraneous variation that might otherwise confound causal inference. Both were night surface actions fought in the same Solomon Islands theater of World War II. Both pitted the Imperial Japanese Navy against allied opposition dominated by the U.S. Navy. Both involved attempts by the Japanese to reach a disputed island in support of their ground force defense of the island against American invasion forces. Technology varied little between the first battle and the second. And both involved outnumbered, outgunned Japanese squadrons with superior torpedoes but otherwise inferior equipment – and especially, inferior sensors for night engagement. The cases differ, however, in the choices and behavior of the respective commanders, especially on the allied side. Perfect control is impossible in case method, but these battles enable an unusually controlled examination of the effects of varying behavior under material circumstances that were uncommonly similar.

They also present extreme outcomes that are unlikely to be products of chance alone. The Japanese victory at Savo Island was remarkably one-sided. Not only was it the second-most-severe defeat in the history of the U.S. Navy, it was more one-sided than almost 70 percent of all battle outcomes by all states in the history of naval warfare since the dawn of the age of sail in the 17th century. And the Cape Saint George outcome was even more one-sided: its casualty-exchange ratio is the 13th-most lopsided result in the last 375 years of naval history. A difference this extreme is unlikely to be a mere artifact of random chance; for two cases whose material circumstances were this similar to produce results this different from differing nonmaterial behaviors is to suggest a systematic causal effect for variance in behavior at sea.

Of course, no analysis based on a handful of cases can in itself prove or disprove causal claims. But the particular features of these two cases thus offer an unusual degree of explanatory leverage, and warrant a greater shift in understanding than such a small sample would otherwise enable.

That shift in understanding focuses in particular on the relative importance of material and nonmaterial variables in naval battle. The radically divergent results at Savo Island and Cape St. George correlate poorly with the material balance of forces. Instead, we argue that the critical difference was a series of very different non-material choices by the commanders in the respective squadrons.

At Savo Island, a splintered allied chain of command, training deficiencies in night operations, and a series of tactical decisions that left allied ships divided into detachments located beyond mutual supporting distance of one another left the allied fleet unable to exploit its material advantages. In particular, the crucial technological advantage of radar could not be exploited by crews and commanders whose skills in using the new equipment were inadequate to realize its theoretical potential. The materially inferior Japanese, by contrast, had a unified chain of command and extensive training in night surface battle, which enabled choices that took maximum advantage of the equipment at their disposal and enabled that equipment to overwhelm a materially superior foe.

At Cape Saint George, by contrast, a re-trained and re-organized American squadron made systematically different choices, maintaining concentration, coordinating fires, and maneuvering to maximize the firepower available to them. And the same radar technology that had failed in practice at Savo Island was now used effectively by crews and leaders who were trained and organized to make the

most of its potential. Systematically different behavior with less-advantageous materiel created a vastly different and far more successful result.

None of this is to suggest that material advantage is irrelevant at sea; of course it is not. But our analysis does suggest that to assess naval balances on the basis of material variables alone is to risk serious error – and error that could have important implications for assessing the balance between U.S. and Chinese naval capability in the Western Pacific and its future.

For policy makers, the paper's chief implication is the need for sustained investment in the personnel, training, and professional military education accounts that create and retain proficient leaders and crews. Of course materiel also matters in naval warfare; in fact, there is good reason to believe that the marginal influence of materiel is stronger for outcomes at sea than on land. But in neither land nor maritime warfare does materiel alone account for all observed variance in outcomes. People and equipment are both expensive. But there can be important incentives to overinvest in materiel, which is manufactured in influential congressional districts, and which has a service lifetime that can be far longer than the transient, perishable effects of personnel training and readiness. There are important interactions between material investment and nonmaterial human performance – an undersized fleet that must operate at a tempo that exhausts crews and precludes training can reduce performance just as underinvestment in training or personnel can do. And there may be grounds for concern that current U.S. Navy operating tempos are exceeding what crews can sustain. But the answer probably does not lie in material acquisition alone. Recruiting, educating, training, and retaining skilled people can make the difference between outcomes like those at Savo Island and those at Cape St. George, and warrant sustained policy attention to ensure.

A second class of policy implication concerns net assessment of the challenges posed by a rising China. Savo Island and Cape St. George suggest that superior materiel alone is not sufficient to ensure success in naval warfare. A growing economy can underwrite a large, well-equipped fleet, and this is clearly important. But navies are exceptionally complex combinations of technology and human organization, and they operate in environments where decision timelines on the order of minutes or even seconds can make the difference between victory and defeat. The behavior of leaders and crews is a complex product of a host of underlying causes, from the structure of societies to the organization of institutions to the nature of economies and more. It is beyond our scope in this small-*n* paper to establish the causal connection between underlying political and social variables and the kinds of crew and officer behavior we observe at Savo Island or Cape St. George. But if this connection holds at sea as it apparently has on land, then Chinese material preponderance may not necessarily underwrite a commensurate payoff in real, usable naval power.

A third class of policy implication involves the variability of outcomes in naval warfare. The paper considers a counterfactual variation in which Allied errors at Savo Island are remediated and the Allies use their materiel to something like its full potential. The most important net consequence of this is to eliminate just a few minutes of unopposed firing for the Japanese and to substitute a few minutes of unopposed firing for the Allies – these differences are the product of a variety of changes in Allied behavior, but their most important single effect is to alter the timeline of the battle on a scale that is nevertheless far smaller than hours, much less days or weeks. These differences produced radical changes in combat outcomes: improved Allied behavior increases Japanese losses from the historical outcome of two ships damaged to all eight ships sunk in the counterfactual – a loss rate in the

counterfactual of 100 percent, without any change in the materiel engaged on either side. Wartime behaviors of these kinds, however, are difficult to anticipate in peacetime. Russian behavior in their February 2022 invasion of Ukraine, for example, was famously different from that expected by most Western analysts; in general, it is easier for intelligence agencies to count ships or missiles or aircraft than to anticipate sometimes subtle differences in human behavior, and especially so for war at sea, where the differences that matter can unfold on timelines of just a few minutes in combat. Nor can intelligence analysts safely conclude that if a technology has been deployed then it will be used to anything like its full potential. U.S. radar offered a powerful potential advantage in 1942, but an analyst who assumed its theoretical performance to estimate real capability at Savo Island would have been wildly off the mark. This sensitivity of outcomes to variations in behavior that are hard to anticipate in peacetime makes net assessment of likely outcomes at sea unusually challenging. And this in turn should counsel humility in our ability to project the likely outcome of naval combat: the same materiel can produce enormous variations in outcomes as a function of behavioral differences that are challenging to anticipate in advance.

For scholars, these results reinforce what is perhaps the central finding of the last generation of research in military effectiveness: the importance of nonmaterial variables. The military effectiveness literature has heretofore been dominated by studies of continental warfare; by extending its reach into the maritime domain, we demonstrate that at least some of that continentally-focused literature's findings hold for war at sea as well.

But perhaps our most important scholarly implication is to show that the study of naval warfare is inherently a social science undertaking – engineering and physics are obviously important for understanding war at sea, but the physical sciences alone present an incomplete picture. There is also a critical role for the study of human behavior and human choices and how these interact with the nature of the materiel in use. The greater the degree to which physical and social scientists can interact with one another the greater the likely payoff in our ability to understand this critical domain.

Quantitative Archival Analysis in Strategic Studies: The World War II Pacific Submarine Campaign

Our third paper is titled "Quantitative Archival Analysis in Strategic Studies: The World War II Pacific Submarine Campaign," by Stephen Biddle, Eric Min, and Laura Samotin. It responds to the second subtask in the project, assessing the feasibility of large-scale military effectiveness data collection via machine coding of archival records. In it, we present a method that uses optical character recognition (OCR) software and a process of crowdsourced data conversion to code narrative archival records at scale, producing large-*n* data on a variety of important nonmaterial variables. We illustrate this method by coding a new dataset on the determinants of success and failure in submarine warfare, and demonstrate its utility with a preliminary analysis of the resulting data.

The results demonstrate that it is possible to extract quantitative data from archival records on topics of interest to strategic studies. Archived personnel records on US submarine commanders in the Pacific campaign of World War II enable us to study the effects of a wide variety of non-material variables that are absent from previously available large-*n* datasets, including the effects of leadership, initiative, judgment, industry, forcefulness, and comportment, as well as a variety of intellectual abilities such as officers' accomplishment in engineering, mathematics, languages, and government. By coding these

data and merging them with a preexisting dataset on the outcomes of combat action with Japanese surface ships we have demonstrated that it is possible to discover statistically significant relationships between nonmaterial variables – especially the traits of individual commanders – and combat outcomes.

These data are incomplete – our work was interrupted by the COVID-19 pandemic, which closed the relevant archives; the re-opened archives are still working through deep backlogs of requests, which has restricted our access to the archival records on which the method turns. But we have secured access to enough records to demonstrate the method, its techniques and requirements, its strengths and weaknesses, and its potential to shed light on nonmaterial variables and their role in war.

These new data offer a number of substantively important insights. First, reinforcing the findings of the other papers in the project, we show that nonmaterial variables are likely to matter for naval combat in ways that would render analyses based on materiel alone suspect. Even the partial dataset available in 2023 indicates that traits such as officers' academic performance as cadets had important consequences for their subsequent success in combat. As our access to records expands and as the statistical power available to the models grows, we think this finding is likely to prove robust, and we are confident that the data will also suggest a range of other important nonmaterial contributors to combat effectiveness.

Second, not all individual traits are equally important. Many believe that initiative and leadership, in particular, are important determinants of success in combat – and indeed they may well be. But our initial analysis suggests that for submarine combat in World War II they may have been less important than the intellectual and personal qualities that produced strong performance at the Naval Academy.

That said, we should emphasize that these are tentative findings pending completion of data collection. These findings are also dependent on superior officers' ability to assess the leadership and initiative of their subordinates in the years prior to World War II. Some of the most important archival records for the analysis are senior officers' prewar FITREPs, or Fitness Report evaluations of their subordinate officers who then commanded submarines in the war. The evidence on which these findings are based is necessarily the subjective assessments of the human beings who supervised the rated officers, and of course these human beings were subject to a range of potential biases and perceptual filters. These assessments, however, were made by qualified naval professionals who had worked with the rated officers and who were responsible for evaluating them to the best of their ability. Promotion and development of officers are among the central functions of any military organization. They shape the careers of every individual in the Service, and they ultimately determine the Service's future and the identity of the officers who will run it. These decisions are taken with the utmost seriousness by all military organizations. The assessment process may well be flawed – indeed a central purpose of our analysis is to help identify any such flaws in the interest of improving the evaluation process – but they are not flawed because the officers doing the rating were unqualified or unserious about the undertaking.

As we complete our data collection, we anticipate a variety of extensions and expansions of the models considered above. Among these will be a more detailed treatment of the military circumstances of the engagements (such as the nature of Japanese escorts, the relative strength of American submarine detachments and Japanese warships, the class of American submarine engaged and its technology, the tactics employed, the weather, and so on). We also anticipate modeling a wider range of officer traits, including the incidence of commendations and disciplinary actions, family structure (a possible

contributor to risk tolerance), the nature of prior service assignments, age and experience, or socio-economic status at the time of enlistment, among others.

In the meantime, our work to date has also suggested a variety of best practices for the conduct of such research. When we began the project, we had anticipated that machine-reading using the ABBYY FineReader software would enable us to code the entirely of officer personnel files automatically, with human intervention limited to providing some guidance regarding the structure of each image and reviewing the results for accuracy and occasional correction. This was indeed possible for some records, (especially the Academy Register that presents the grades earned by officers as cadets at the Naval Academy), and would be possible for some other archival records where the information is typed in a well-structured format. In many cases, however, the nature of the records in the archives makes this impossible for the current state of the art in OCR technology. Some information is hand written. Many documents' condition is poor in ways that confuse the OCR system. Stamps or annotations confuse the OCR system. And document formats changed frequently as the Navy changed its forms for officer evaluation, increasing the coding effort that would be required to machine-read them. Some of these challenges can be overcome with sufficient labor (such as writing new codes for each version of Navy forms). Others will be overcome in time as artificial intelligence technology improves – it will eventually be possible to machine-read handwritten text, for example.

For now, however, we believe that crowd-sourcing offers a superior methodology given what we have discovered about the nature of the available documents. Scholars have already noted the potential benefits to using crowdsourced techniques to correct poorly converted OCR text, but actual implementation of crowdsourcing has involved the creation of bespoke software and is limited to only reviewing text that has already undergone an automated OCR process. Amazon's Mechanical Turk (MTurk) service, by contrast, offers a more streamlined, versatile, and cost-effective means for coding large volumes of such material. MTurk is a popular and commonly used tool that matches Requesters (those who want to assign a task) with Workers (those who want to perform tasks for money) to complete tasks online. We were able to code 4,198 FITREPs in this way at a total labor cost of under $7,000. Effective use of MTurk requires meaningful preparatory effort by researchers; we discovered, for example, that the workers' efforts could be made far more efficient by developing standardized frames for data entry. These preparatory efforts will be unique to the project undertaken, but our experience indicates that early investment of effort in such mechanisms is rewarded by a substantial increase in data collection efficiency overall. As we complete our data collection we anticipate that other best practices will also emerge.

On balance, however, we are convinced that it is both possible and intellectually productive to create new strategic studies datasets from archival records using such methods. And we are convinced that this approach offers a new and unique window into the essential role of nonmaterial variables, in particular, for military effectiveness and the outcomes of armed conflict. Commander traits in the World War II submarine campaign offer an interesting starting point for demonstrating these methods – but much more can be done, and much more can be learned, from a broader exploration of the wealth of knowledge in the nation's narrative archival records for the study of war via large-*n* data collection.

Military Effectiveness and Naval Warfare

An additional paper that was not part of the grant but which was written by members of the project team and which contributes to the understanding of similar issues, and which was delivered in draft to the sponsor in August 2023 as an additional deliverable, is Stephen Biddle and John Severini, "Military Effectiveness and Naval Warfare," now forthcoming in the journal *Security Studies*. The paper uses the NAVBATTLE dataset to compare patterns in warfare on land and at sea, and to derive implications from this for defense policy and international relations scholarship.

The paper finds important differences deriving from the contrasting nature of the sea and land as military environments. In particular, the greater exposure of naval battle makes battle outcomes at sea typically faster, more decisive, more one-sided, and more sensitive to differences in materiel than battle outcomes on land.

Critical similarities arise from the technical complexity of all modern warfare and the challenges this creates for human organizations. Material variables are relatively more influential for outcomes at sea than on land, but in neither domain is materiel a sufficient explanation in itself. War in either domain is an increasingly complex enterprise – the relatively simpler physical environment at sea is offset by the more complicated nature of the machines used there and the greater time pressure in their interactions, not to mention the moral and psychological challenges of all mortal combat. Both on land and at sea, some states have proven able to field combat organizations with the skills and motivation needed to master dangerous and highly complex enterprises, but others have not. As the nominal lethality of skillfully employed weapons has grown in both land and sea warfare, this variation in state organizational capacity has created a growing gap in the real military effectiveness of states that can and states that cannot field the necessary organizations. And this in turn has placed an increasing premium on all actors' ability to field organizations that can master complexity under danger. Materiel matters more for war at sea than on land, but in both domains its effects interact with the non-material variables that produce skilled, motivated execution.

These features pose implications for policy and scholarship. As for policy, an important issue in the recent debate has been the growing importance of land based anti-access area denial (A2/AD) capability for naval combat. While a complete analysis of A2/AD campaign dynamics is beyond the paper's scope, it is nevertheless likely that these dynamics will differ significantly from traditional naval combat by virtue of A2/AD's reliance on land basing. Whereas many of A2/AD's targets are surface ships, the missiles that create A2/AD capability are chiefly land based. These land-based missiles are increasingly able to target shipping many hundreds or perhaps thousands of kilometers at sea. The analysis in the paper demonstrates a number of important differences between land and maritime combat; mature A2/AD capability would create a hybrid operating environment in which maritime targets that choose to operate in large parts of the south and east China Seas would suffer the exposure typical of naval warfare whereas their land-based assailants would be embedded in the complexity of the terrestrial environment. The terrestrial environment's complexity increases the duration and reduces the decisiveness of operations against land forces, whereas the exposure of naval combat conduces to shorter, more destructive, and more one-sided combat dynamics. And this in turn suggests that it will be difficult, other things being equal, for naval forces to survive the prolonged campaign they would require to root out an extensive system of land-based A2/AD. Given the cost and slow replacement time

of modern warships, it may be that sailing away rather than slugging it out could be the wisest course for navies facing a mature land based A2/AD threat.

But this depends on the skills of the respective combatants. The importance of human behavior for outcomes in either domain suggests that highly-lethal modern weapons might empower a highly skilled navy to disarm an unskilled A2/AD opponent at tolerable cost. The analysis in the paper suggests that error in the face of skillfully-wielded modern weapons is increasingly costly, and the complexity of modern A2/AD could promote error for states whose armies function more like those of France in 1940 or Iraq in 1991 and less like Germany in 1944 or the United States in 1991.

Given this, it would be a mistake to allow U.S. skills to atrophy in the interest of increasing modernization spending for the U.S. Navy. Materiel matters more at sea than in land warfare, but it determines outcomes in neither. And martial skill is expensive to create, expensive to maintain, and perishable. As a result, it can slip in peacetime under pressures to modernize technology, expand force structure, or maintain a demanding operations tempo. In fact, there may be reason to believe that skills in important parts of the U.S. Navy may have decayed in recent years as a result of such pressures. If so, the analysis above suggests potentially grave consequences in a high-lethality confrontation with a skillfully employed Chinese A2/AD system even if U.S. technology keeps pace. To compete successfully with China will require astute naval modernization and an adequate force structure – but it will also require the training and personnel retention needed to preclude a meaningful skill deficit. Given the greater relative importance of materiel at sea, it is reasonable to spend relatively more on modernization for maritime than terrestrial forces – but if this happens at the price of an unskilled navy that price will likely prove too high.

For scholars, the results above suggest an important opportunity for research. The effectiveness literature in political science has produced important findings, but to date it has focused largely on continental conventional warfare. Maritime conflict is increasingly important, and its dynamics are different enough that effectiveness at sea cannot simply be assumed to follow the patterns identified for warfare on land. There are important similarities – but there are also crucial differences. These differences warrant sustained examination that exploits the findings of recent political science research on effectiveness but which applies, extends, adapts, and modifies them for the unique environment of war at sea.

Training

All research assistants on the project were included in all project meetings and participated fully. This included a weekly project update meeting, in which progress is discussed and the latest statistical results are presented for group discussion.

Professor Biddle also provided for all new project staff an extensive on-boarding tutorial in the project's goals, design, and intended policy role, as well as a refresher and tutorial on the statistical methods used in the project.

Professors Biddle and Berman also worked closely offline with the PhD candidates who played the primary role in carrying out the statistical modeling and case research, providing guidance and instruction.

All project publications will be co-authored with one or more graduate student participants in the project.

Plans

The grant is complete, but our work on the subject matter continues. For now, this consists in presenting the draft papers at scholarly conferences and obtaining comment and feedback on the drafts, which we will use to improve the manuscripts and to revise the drafts for publication. We anticipate that revised versions of "Victory at Sea" and "Technology, Behavior and Effectiveness in Naval Warfare" will be submitted for publication this summer. "Quantitative Archival Analysis in Strategic Studies" has been published in the ESOC Working Paper series, but we are continuing to compile and code archival records as the archives make these available. This process is slow but is under the control of the archives and largely out of our hands; we hope to complete this document collection in the next year to two years, at which point we will re-analyze the data, revise the paper, and submit the revised version for journal publication – but this timetable depends on the archives' ability to work their way through their backlog and is thus subject to considerable uncertainty.

We also anticipate presenting our findings to scholarly and policy audiences as opportunity affords. The first such presentation will be this February, at the Western Naval History Association Annual Meeting in San Diego.

Honors and Awards

Nothing to report.

Technology Transfer

Nothing to report.

Dissemination since last interim report

- Stephen Biddle, "The Use of History in Policy Analysis and Security Scholarship," panel presentation, Colgate University, 11/8/22
- Stephen Biddle, "Military Strategy in the Contemporary World," online, 10/25/23 (over 175 people in audience)
- Stephen Biddle, "Future Warfare in the Western Pacific," presentation to the Stanford University Workshop on US-China Competition, Palo Alto CA, 11/4/23
- John Severini and Stephen Biddle, "Technology, Behavior, and Effectiveness in Naval Warfare," paper presentation to the Dartmouth College Military Force Analysis Seminar, online, 1/26/24