



Strategy Is Only Partly an Illusion: “Relative Foresight” as an Objective Standard for Evaluating Foreign Policy Competence

Citation

Friedman, Jeffrey A., and Richard Zeckhauser. "Strategy Is Only Partly an Illusion: “Relative Foresight” as an Objective Standard for Evaluating Foreign Policy Competence." HKS Faculty Research Working Paper Series RWP24-004, May 2024.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37378528>

Terms of use

This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles (OAP), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.

Please share how this access benefits you. [Submit a story](#)



HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Strategy Is Only Partly an Illusion: “Relative Foresight” as an Objective Standard for Evaluating Foreign Policy Competence

Faculty Research Working Paper Series

Jeffrey A. Friedman
Dartmouth College

Richard Zeckhauser
Harvard Kennedy School

May 2024

RWP23-004

Visit the **HKS Faculty Research Working Paper Series** at: <https://ken.sc/faculty-research-working-paper-series>

The views expressed in the **HKS Faculty Research Working Paper Series** are those of the author(s) and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. Faculty Research Working Papers have not undergone formal review and approval. Such papers are included in this series to elicit feedback and to encourage debate on important public policy challenges. Copyright belongs to the author(s). Papers may be downloaded for personal use only.

Strategy Is Only Partly an Illusion:

“Relative Foresight” as an Objective Standard for Evaluating Foreign Policy Competence

Jeffrey A. Friedman

Associate Professor of Government, Dartmouth College

jeffrey.a.friedman@dartmouth.edu

Richard Zeckhauser

Ramsey Professor of Political Economy, Harvard University

richard_zeckhauser@harvard.edu

Forthcoming in *Foreign Policy Analysis*

Abstract

Foreign policy-makers must grapple with complexity, uncertainty, and subjectivity. As Betts (2000) puts it, these challenges raise the possibility that “strategy is an illusion”: that there is no reliable method for assessing skill at managing international politics. By contrast, we show that researchers can objectively evaluate a critical component of foreign policy competence using a standard we call “relative foresight,” defined as decision-makers’ ability to anticipate consequences of their choices as compared to alternative views based on similar information. Relative foresight can be measured without relying on value judgments or subjective probabilities. By contrast, other common frameworks for gauging foreign policy competence, such as comparing leaders’ behavior to the rational actor model or assessing procedural rationality, almost always leave room for reasonable disagreement. We demonstrate that relative foresight provides a useful tool for evaluating major foreign policy choices through case studies of Barack Obama’s decisions regarding the Afghan Surge and the raid on Osama bin Laden’s compound. Our framework has broad implications for research on normative, prescriptive, and descriptive dimensions of foreign policy analysis.

Acknowledgments

For helpful comments on earlier drafts, we thank Kathryn Schwartz, Mike Poznansky, Megan Stewart, two anonymous reviewers, and participants at the 2021 Pyrenean Interdisciplinary Research Event. Friedman conducted research for this paper while he was a fellow at the Institute for Advanced Study in Toulouse. Funding from the French Agence Nationale de la Recherche (under the Investissement d’Avenir programme, ANR-17-EURE-0010) is gratefully acknowledged.

Strategy Is Only Partly an Illusion

Was Napoleon a strategic genius, or did he just get lucky? Was Henry Kissinger an effective Secretary of State? How well did Indira Gandhi handle India's relationship with Pakistan? Does Vladimir Putin's decision to invade Ukraine suggest that he is irrational?

Answering these kinds of questions requires assessing *foreign policy competence*, defined here as a leader's ability to make foreign policy decisions that promote expected national interests. Foreign policy competence is notoriously challenging to evaluate, because it is generally impossible to know how well decisions maximize expected national interests without making value judgments or assessing subjective probabilities (Jervis 1997; Renshon and Larson, 2003; Burchill 2005). As Betts (2000) puts it, the complexity, uncertainty, and subjectivity of foreign policy decision-making raise the troubling prospect that "strategy is an illusion" – that there are no objective criteria for determining which foreign policy choices are more defensible than others.

Subjective evaluations can play valuable roles in structuring foreign policy analysis. For example, even if it is generally impossible to say that foreign policy decisions are objectively "right" or "wrong," careful reasoning can winnow complex debates down to the core assumptions one must adopt in order to justify high-stakes choices.¹ The balance of expert consensus also provides a useful heuristic for judging leaders' competence, subject to the

¹ For example, several scholars argue that U.S. President Lyndon Johnson made a reasonable decision to fight the Vietnam War (e.g., Mueller 1980; Mearsheimer and Rosato 2023). Though reasonable people disagree with that interpretation, careful articulations of that argument help readers to develop more sophisticated views regarding the costs, benefits, and uncertainties associated with U.S. decision-making in Vietnam.

caveat that widely-shared opinions do not always reflect rigorous analysis (Badie 2010; Schafer and Crichlow 2010).²

The subjectivity that surrounds foreign policy analysis nevertheless raises fundamental problems for scholarship and policy. If researchers lack objective methods for separating good choices from bad, then reasonable people will generally be able to disagree about whether foreign policy decisions were mistaken. That ambiguity makes it difficult to draw firm conclusions about how foreign policy can be systematically improved (Betts 2000). Moreover, any time scholars claim that foreign policy decisions were warped by factors such as domestic politics or personal psychology, they are implicitly claiming that competent leaders should have known that a different course of action was superior to the one they actually chose (Glaser 2010, 2-3; Stein 2013, 369). The subjectivity that is inherent to analyzing foreign policy decisions thus impedes making normative, prescriptive, and descriptive claims about the conduct of international politics.

This paper's principal contribution is to show how researchers can objectively evaluate one critical component of foreign policy competence. We call that component "relative foresight," defined as a decision-maker's ability to anticipate the consequences of their choices as compared to alternative views that were based on similar information. We explain why relative foresight plays an important role in making competent foreign policy decisions. We also show how this concept differs from other approaches that researchers commonly employ to assess

² For example, even though some scholars argue that Napoleon's battlefield successes stemmed largely from luck rather than skill (Connelly 2006), most experts believe that Napoleon was an unusually effective strategist (Leggiere 2023). Even if there is no way to objectively conclude whether this consensus is correct, it signals that one side of the debate is likely more plausible than the other.

foreign policy competence, such as the rational-actor model, procedural rationality, and Brier Scores. Though foreign policy decision-makers would ideally possess many attributes beyond relative foresight, we describe how reasonable people can almost always disagree about the extent to which leaders possess those other attributes, or about whether those other attributes are important for promoting national interests. The approach we propose for measuring relative foresight thus stands out from other frameworks, because it allows researchers to evaluate a critical dimension of foreign policy competence without relying on subjective judgments.

We also show that the concept of relative foresight provides a novel lens for assessing foreign policy choices. For example, the 2011 raid on Osama bin Laden's compound in Abbottabad, Pakistan is widely celebrated as being one of President Barack Obama's signature foreign policy successes. Yet, the fact that the raid succeeded does not prove that it was a reasonable risk to take. Even Obama, at the time, expressed ambivalence about whether the available intelligence was strong enough to justify launching the mission. If anything, most published accounts of this decision agree that Obama was more pessimistic than his advisers about the probability that bin Laden was living in Abbottabad.³ Obama's relative pessimism about a successful policy decision provides a negative signal of relative foresight, as that judgment proved to be less accurate than the expectations of other well-informed national security officials.⁴

³ See Friedman and Zeckhauser (2015) for a review of relevant sources.

⁴ As we discuss below, a "negative signal," though indicative, is not dispositive. As with any other uncertain endeavor, competent decision-makers will often make smart judgments that do not align with observed outcomes, and it is thus difficult to distinguish skill from luck in just one or a few cases. In the long run, however, more competent decision-makers should display greater levels of relative foresight, on average.

By contrast, even though the Afghan Surge is generally considered to be one of Obama's major foreign policy failures, Obama anticipated the downside risks of that decision more accurately than his military advisers did. Obama's perceptions of the Afghan Surge thus appear to have been better-calibrated than alternative views based on similar information. Of course, Obama's critics would say he should not have approved the Afghan Surge if he questioned that policy's prospects. But low-probability gambles can be worth taking if the benefits of success significantly outweigh the costs of failure (Gelb and Betts 1979). Unfortunately, there is no objective way to determine whether or not the Afghan Surge met that standard. Once again, focusing on factors that lend themselves to objective evaluation presents new insight for (re)evaluating major foreign policy choices.

Our analysis contributes to long-standing debates about the challenges of evaluating the quality of high-stakes policy-making (Vickers 1965; Betts 2000; George 2006; Mearsheimer and Rosato 2023). We show that foreign policy competence is only partly an "illusion": important elements of foreign policy competence can, in fact, be measured objectively. To our knowledge, this paper is the first work to show that. Yet, we readily acknowledge that relative foresight is just one component of foreign policy competence writ large. By distinguishing one aspect of foreign policy competence that lends itself to objective evaluation from others that do not, we probe the boundaries of empirical reasoning in foreign policy analysis, showing how some aspects of this enterprise permit clear resolution while others inherently rest on subjective judgments over which reasonable people can almost always disagree.

The paper proceeds in five parts. Section 1 explains why several prominent approaches to assessing foreign policy competence almost always rely on value judgments or subjective probabilities. Section 2 develops our concept of relative foresight and explains how researchers can objectively measure this attribute. Section 3 shows how the concept of relative foresight offers a novel lens for (re)evaluating foreign policy decisions through case studies of the bin Laden raid and the Afghan Surge. Section 4 discusses methodological challenges surrounding our framework. Section 5 concludes with broader implications for foreign policy analysis.

Section 1. Existing Frameworks for Evaluating Foreign Policy Competence

What does it take to draw objective conclusions about foreign policy competence? For the purposes of this paper, we define objectivity with respect to the properties of verifiability and unbiasedness. *Verifiable* inferences depend on observable facts. *Unbiased* inferences reliably correlate with foreign policy competence. Neither of these criteria requires possessing perfect information; indeed, measurement error is ubiquitous throughout the social sciences. The key question for our purposes is whether parameters of interest can be measured in principle. If no amount of empirical information could ever resolve reasonable disagreements when assessing foreign policy competence, then we would not consider those assessments to be objective.

Poker players' winnings provide an instructive example of how imperfect data can ground objective inferences about decision-making competence. Good poker players do not win every session in which they participate, much less every hand that they play. A few hands of poker thus convey limited information about which players are better than others. But, the more

hands of poker we observe, the more reliably poker winnings will reflect players' skill levels. This indicator therefore satisfies our criteria for providing a verifiable and unbiased metric of decision-making competence.

Unfortunately, foreign policy decision-making does not produce such clear outcome metrics. Whereas poker players' winnings are easy to observe and tally, there is no obvious way to quantify the kinds of national interests that foreign policy decision-makers seek to advance (Finnemore 1996; Burchill 2005). And, while it is relatively straightforward to estimate the chances that poker players will draw the cards they need to win a hand, even the most sophisticated statistical models can at best approximately forecast outcomes in international affairs (Jervis 1997). As a result of these problems, we will show that three prominent approaches to studying foreign policy competence – calculating success rates, comparing decision-makers' performance to the rational-actor model, and assessing procedural rationality – cannot generally produce objective conclusions. The paper's next section will then explain how our concept of relative foresight meets the standards of verifiability and unbiasedness.

Success rates and aggregate performance

One intuitive approach to gauging foreign policy competence is to examine leaders' track records for making successful choices. Thus, the bin Laden raid would serve as a positive signal of President Obama's competence, whereas the failed Iran hostage rescue mission would yield a negative signal of President Jimmy Carter's competence. A substantial volume of scholarship

assumes that this is how rational observers can and should evaluate the competence of foreign policy decision-makers (e.g., Downs and Rocke 1994; Smith 1998; Reiter and Stam 2002).

Yet, success rates cannot ground objective evaluations of decision-making competence without also accounting for the stakes that different choices involve. Consider a situation in which a decision-maker must choose between two lotteries, labeled A and B. Lottery A has a sixty percent chance of paying out \$10, and a forty percent chance of paying nothing. Lottery B has a twenty percent chance of paying out \$100, and an eighty percent chance of paying nothing. A competent decision-maker who is not highly averse to risk should prefer Lottery B over Lottery A, even though Lottery B has a far lower success rate. Relying on success rates to measure decision-making competence will thus reward leaders for taking high-probability bets, even when those bets have negative expected value, rather than making decisions that maximize expected utility.

As noted above, this problem is easy to solve in poker, where a player's net winnings are easy to quantify and they reliably reflect the quality of decision-making in large samples. By contrast, scholars lack clear criteria for assigning value to different foreign policy outcomes (Baldwin 2000). Jimmy Carter's handling of the Iran hostage crisis is a good example of this problem. The material stakes involved in that crisis – fifty-two embassy employees held hostage and eight U.S. servicemembers killed in a failed rescue mission – were relatively small by the standards of major military decisions. Yet, the Iran hostage crisis also involved important intangible considerations, such as preserving America's reputation for resolve in confronting hostile adversaries and maintaining national morale. There is no commonly-accepted method for

quantifying the value of those intangible factors. This means there is also no objective way to determine whether Carter's failure to rescue the Iran hostages outweighs his success in other areas of foreign policy, such as mediating the 1978 Camp David Accords. Any attempt to "sum up" the overall costs and benefits of Carter's foreign policy decision-making thus inherently depends on researchers' value judgments.

The rational-actor model

A second approach to evaluating foreign policy competence is to ask how well leaders' choices resemble the decisions that rational actors would have made. The rational-actor model plays a central role in empirical analyses of foreign policy decision-making (Geva and Mintz 1997).

Whenever researchers argue that a foreign policy decision was shaped by factors such as cognitive biases, political pressures, or bureaucratic procedures, they are implicitly claiming that a rational leader who was not subject to those forces would have made a different, better choice (Glaser 2010, 2-3; Stein 2013, 369).

The rational-actor model provides a good benchmark for gauging skill in realms, such as poker, where it is always possible to precisely estimate a player's chances of winning a hand and it is easy to quantify the costs and benefits associated with making different bets. But the rational-actor model provides limited guidance for defining the actions that competent decision-makers should take in foreign policy, where probability assessments and value judgments are inherently subjective, and where it is thus impossible to objectively quantify the degree to which any foreign policy decision should be expected to advance national interests (Mearsheimer and Rosato 2023, 19-100).

To illustrate, Lyndon Johnson's decision to send combat forces to Vietnam is among the most thoroughly-criticized military decisions in U.S. history. Yet, even here, it is difficult to make a clear case for strategic incompetence. Johnson and his advisers generally understood that escalating the Vietnam War was a long-shot gamble. They nevertheless believed that this risk was worth taking in order to stave off a "domino effect" of spreading communism in East Asia (Gelb and Betts 1979). Moreover, even though Johnson and his advisers knew escalating the Vietnam War ran a substantial risk of failure, it was hard to believe that prospects for defeating the communist insurgency were nil given the United States' massive material advantage (Mueller 1980; Hoffman et al. 1981, 9). To show that Johnson failed to make this risk-reward tradeoff correctly, one would need to estimate at least four parameters: the probability rational actors would have assigned to escalation preserving a non-communist South Vietnam, the costs that rational actors would have expected the war to incur, the probability rational actors would have assigned to Saigon's collapse triggering communist uprisings in other countries, and the costs rational actors would expect the United States to bear if those "dominoes" fell. Since none of those parameters can be estimated with any precision, it is impossible to prove that a competent decision-maker should have known to pursue a different course of action.⁵

⁵ Intuitively, one might expect there to be cases where competent leaders should pursue risky decisions simply because the stakes are so enormous. One example might be Winston Churchill's decision to reject a deal with Hitler in 1940, despite the overwhelming odds that Britain faced in fighting Nazi Germany. Yet, as Betts (2000, 13) explains, Churchill's reasoning that it was better for Britons to "choke in [their] own blood" rather than to sign a negotiated settlement with Hitler was very much open to dispute. Since even stakes this high do not justify national suicide, Churchill's decision to wage the Battle of Britain can only be defended by demonstrating that the balance of costs, benefits, and risks produced a positive expected value for the nation.

Procedural rationality

A third common approach to evaluating foreign policy competence involves assessing procedural rationality. Instead of asking how close leaders came to maximizing expected national interests, researchers use the concept of procedural rationality to examine how thoughtfully leaders processed information before making high-stakes choices. Examples of effective information processing include considering multiple options, encouraging dissenting opinions, and explicitly weighing tradeoffs among competing values (e.g., Herek, Janis, and Huth 1987; Renshon and Larson 2003; Badie 2010; Schafer and Crichlow 2010).

Yet, it is not obvious that higher degrees of procedural rationality systematically generate better foreign policy choices. For example, Egyptian President Anwar Sadat was known to base decisions on gut instincts rather than rigorous deliberation. When Sadat decided to launch a war against Israel in 1973, most of his advisers argued that it was foolish to attack an opponent who possessed conventional military superiority. Sadat discounted these well-sourced assessments of the war's risks based on his intuition that capturing a relatively small piece of Israeli-occupied territory would open up "psychological space" for conducting negotiations. Sadat's critics thus describe the October War as a reckless gamble – but Sadat's admirers say that his willingness to discard ostensibly-rational advice was precisely what allowed him to be a transformative figure (e.g., Israeli 1985; Kissinger 2022, 205-276).

Similarly, several studies identify "transformational" or "visionary" leadership as desirable attributes that are distinct from – and, arguably, in tension with – the capacity for rationalist,

cost-benefit analysis (Hermann, Preston, Korany, and Shaw 2001; Rathbun 2019).⁶ Procedural rationality may even degrade foreign policy competence in some cases. For instance, Johnson (2019) argues that cognitive biases such as overconfidence and fundamental attribution error can grant foreign policy decision-makers adaptive advantages over more procedurally rational adversaries. Some scholars of intelligence and national security similarly argue that attempts to promote procedural rationality through the use of structured analytic techniques can dampen valuable intuition and thereby impair analysts' abilities to support sound decision-making (Marrin 2012).⁷

If we cannot say with certainty what kinds of decision-making processes produce better foreign policy decisions than others, then the only way to resolve these debates is to conduct empirical tests demonstrating that some procedures are systematically more likely to yield outcomes that advance expected national interests. Yet, we have already seen that it is impossible to develop objective metrics for determining which foreign policy decisions advance expected national interests better than others.⁸

⁶ Some decision-makers, most notably Richard Nixon, believe that procedural irrationality can also be advantageous when coercing adversaries, who may be cautious to avoid the risks of escalation tensions with reckless leaders (McManus 2021).

⁷ For example, an experimental study by Dhami, Belton, and Mandel (2019) indicates that Analysis of Competing Hypotheses, which is perhaps the most widely-known structured analytic technique for intelligence analysis, may in fact marginally reduce the accuracy of intelligence estimation.

⁸ By way of analogy, political scientists have demonstrated that leaders who possess particular attributes are prone to different behaviors. For example, leaders appear to be more prone to escalating militarized disputes when they place more weight on preserving their personal reputations (Yarhi-Milo 2018), or when they lack combat experience (Horowitz, Stam, and Ellis 2015). But this descriptive pattern does not readily justify normative judgments. In order to say whether leaders who are more prone to escalation are "better" or "worse" decision-makers than their peers, it is necessary to identify what the optimal rate of escalating international crises entails. Defining that standard requires making value judgments and assessing uncertainty in a manner that almost always leaves room for reasonable disagreement.

Is strategy an illusion?

All three frameworks described in this section add important value to scholarship on foreign policy analysis. For example, the fact that Napoleon won a significant string of battles against materially-superior opponents does not directly prove that he was a superior strategist (especially since the Napoleonic Wars ultimately resulted in their namesake's defeat and exile). Napoleon's unusual success rate in conventional combat nevertheless indicates that at least some of his attributes as a military leader were worth emulating, and that observation played a fundamental role in grounding modern discourse on strategic studies (Gat 1989). Similarly, even if it is rarely obvious how a perfectly rational actor would make the difficult tradeoffs that foreign policy decisions require, careful analysis can at least help to focus analysts' attention on the assumptions that play the most important role in evaluating high-stakes choices like Lyndon Johnson's decision to escalate the Vietnam War. And though it is difficult to demonstrate that any particular conception of procedural rationality generates better foreign policy outcomes than others, research in this field has provided policymakers with a range of templates they can use to structure complex decisions (e.g., Mintz and Wayne 2016; Heuer and Pherson 2020).

Despite the important contributions that these frameworks have made to foreign policy analysis, it is important to recognize that none of them can ground objective inferences about the quality of foreign policy decision-making. Exceptions to this argument must involve leaders making choices that can be identified as errors without either invoking value judgments or assessing subjective probabilities. For instance, many historians blame the ill-fated Charge of the Light Brigade a commander who issued poorly-written orders that led a subordinate to

mistakenly attack an impregnable defense (Adkin 1996, 125-134). These kinds of obvious failures of decision-making are nevertheless rare in international politics; searching for them narrows the scope of foreign policy analysis in a manner that cannot ground a generalizable framework for evaluating leaders' capabilities.

Section 2. Relative foresight

This section thus proposes a new approach for evaluating foreign policy competence, using a concept that we call relative foresight. We define relative foresight as a decision-maker's ability to anticipate consequences of their choices as compared to alternative views that were based on similar information.

As stated at this paper's outset, relative foresight captures only one aspect of foreign policy competence. Yet, leaders' predictive capacities are undoubtedly crucial for making effective foreign policy decisions. Every foreign policy decision represents a prediction that one course of action should, in expectation, lead to better outcomes than its alternatives (Betts 2000). A decision-maker with better predictive abilities should thus be better-able to make choices that advance expected national interests. The framework we advance for measuring foreign policy competence is also, to our knowledge, the first tool that scholars have proposed that can consistently measure foreign policy competence in a manner that satisfies the properties of verifiability and unbiasedness. This section thus establishes that strategy is only partly an "illusion": that at least one important element of foreign policy competence lends itself to objective assessment.

To see why this is the case, we can denote the expected utility of a foreign policy decision as $EU = p_s \cdot U(\text{Success}) + (1 - p_s) \cdot U(\text{Failure})$. In this expression, p_s represents the chances decision-makers assign to achieving their strategic objective, $U(\text{Success})$ represents the utility decision-makers expect to receive, relative to the status quo, if they achieve their intended objective relative to the status quo, and $U(\text{Failure})$ represents the utility decision-makers expect to receive, relative to the status quo, if their decision fails to achieve its intended goal.⁹ Decision-makers who are more competent should make choices that yield higher expected utilities. But we have already seen that any attempt to measure the expected utility of a foreign policy outcome inherently rests on value judgments, as there is no objective way to determine the relative weight that leaders should assign to lives, money, and other elements of the national interest.¹⁰

This leaves leaders' assessments of uncertainty – specifically, the chances they assign to decisions producing different outcomes – as the only component of expected utility that researchers can assess without relying on subjective assumptions. While there is generally no way to know what assessments of uncertainty a rational actor should have made on the basis of information that was available at the time, it is possible to objectively measure a decision-

⁹ Here, we treat success and failure as mutually-exclusive and mutually-exhaustive options. In reality, most strategic decisions involve a much broader distribution of probability-weighted outcomes, which makes the prospect of objectively estimating expected utility even more tenuous.

¹⁰ Researchers can evaluate the accuracy of decision-maker's predictions about the amount of lives or money that a decision will cost. For example, the George W. Bush administration initially estimated that the invasion of Iraq would cost \$50 billion, a figure that was nearly two orders of magnitudes too low. This error reflects poorly on the Bush administration's predictive capabilities in a manner that can be evaluated within our framework of relative foresight. This is different from asking whether the Bush administration appropriately perceived the costs and benefits of the Iraq War from the standpoint of U.S. national interests. Answering that question requires making value judgments about how to weigh lives and money against other components of U.S. national security, as well as making counterfactual speculations about the degree to which leaving Saddam in power would have threatened U.S. interests.

maker's track record for making accurate predictions. The most common approach to measuring forecasting skill quantitatively is to compute a Brier Score, which measures the squared error between an individual's probability assessments, and the judgments they could have made if they knew the future with certainty.¹¹ Tetlock (2005), Mellers et al. (2015) and other scholars have deployed this method to rate the accuracy of geopolitical forecasts.¹²

The Brier Score nevertheless suffers an important drawback when it comes to objectively evaluating the quality of foreign policy decision-making: it is impossible to determine how "good" or "bad" a particular Brier Score is without defining the degree of difficulty associated with the predictions an analyst made. For example, a meteorologist forecasting the chances of rain in the Atacama Desert, which is one of the driest places on earth, will almost surely receive a better Brier Score than a meteorologist forecasting the chances of rain at the summit of Mount Washington, which experiences some of the world's most unpredictable weather patterns. Evaluating a meteorologist's skill thus requires adjusting measures of forecasting performance based on factors such as base rates and volatility. Such adjustments are straightforward to make in weather forecasting, where meteorologists can gather well-structured data about the frequency and consistency of rain in any geographical region.

By contrast, most high-stakes foreign policy decisions involve unique circumstances that have no close analogies, and thus lack large reference classes of similar events that researchers can

¹¹ Thus, if a meteorologist says there is a 20 percent chance of rain and it does in fact rain, then her Brier Score for that judgment would be $(1.0 - 0.20)^2 = 0.64$. If it did not rain on that day, then her Brier Score would be $(0.0 - 0.20)^2 = 0.04$.

¹² On how scholars can calculate foreign policy officials' Brier Scores even when those officials do not quantify their forecasts, see Mandel (2015) and Friedman, Lerner, and Zeckhauser (2017).

use to calibrate statistical models precisely (Neudstadt and May 1986; Khong 1992). As a result, there is rarely an objective way to estimate the chances that a foreign policy decision will generate specific outcomes. Thus, even if researchers computed the average Brier Score associated with every one of a foreign policy decision-maker's assessments of uncertainty, that statistic could not reliably indicate the decision-maker's competence unless those researchers also made unverifiable assumptions about what kind of Brier Score a skilled decision-maker should have received when assessing those same issues.

To overcome this challenge, we propose evaluating the accuracy of leaders' judgments relative to colleagues who assess the same issues based on similar information. The Appendix describes how to quantify that metric, which serves as our formal definition of relative foresight, but that formalization is not crucial to understanding our argument and its implications. Thus, if Meteorologist A works on Mount Washington and Meteorologist B works in the Atacama Desert, then there is little value in comparing their Brier Scores to each other: B's score will almost always be better than A's. Yet, if we gather enough data, we can credibly rank-order meteorologists who work in each region, respectively. We might then learn that Meteorologist A is in the 80th percentile of weather forecasters assigned to Mount Washington, whereas Meteorologist B is in the 20th percentile of forecasters assigned to the Atacama Desert. These scores provide verifiable and unbiased indications of each meteorologist's performance as compared to their peers in a given context.

Anwar Sadat's decision to attack Israel in 1973 illustrates how the concept of relative foresight can be applied to studying foreign policy competence. Though there is no objective way to

determine whether Sadat made the correct choice to launch the October War, the case offers a clear, positive signal regarding Sadat's relative foresight. At the time, most of Sadat's military advisers did not believe it was possible to coerce Israel into making strategic concessions. Egypt's war minister was so skeptical of Sadat's strategy that he refused to develop plans for it, and was thus relieved from command. Accounts of the case show that Sadat also received objections from his vice president, his naval commander, his defense intelligence staff, and two of his top ground forces commanders (Shazly 1980, 27-33, 177-182; El-Ghamasy 1993, 151-158; Bar-Joseph 2005, 11-12). Sadat did not inform his National Defense Council about early stages of planning for the war because he knew that his security team overwhelmingly believed the idea was infeasible (Israeli 1985, 83).¹³ Sadat's decision to pursue the October War thus provides a concrete signal of Sadat's ability to predict the observable outcomes of his choices more accurately than did his advisers.

We would evaluate Sadat's decision-making differently had he followed the consensus of his military and intelligence staff. If Sadat's advisers agreed that the October War was likely to extract concessions from Israel, then this case would no longer indicate that Sadat possessed unusual insight. We could even imagine a situation in which Sadat's decision to go to war indicated that he lacked relative foresight: that would be the case if Sadat had reluctantly chosen to attack despite being more pessimistic than the rest of his advisers about the strategy's chances of success. This logic illustrates how it is only possible to evaluate leaders' predictive abilities by examining their views in relation to other contemporary perspectives.

¹³ By contrast, Sadat believed the attack was likely to succeed – one account claims that he placed the chances of success as high as thirty percent (Israeli 1985, 84).

Section 3. Reconsidering the Bin Laden Raid and the Afghan Surge

If relative foresight captures just one element of foreign policy competence, then how much value can this framework provide for assessing leaders' skill at managing international politics?

In order to show how our theoretical framework adds nontrivial insights to foreign policy analysis, this section examines two of President Barack Obama's best-known foreign policy choices: his 2011 decision to raid Osama bin Laden's compound in Abbottabad, Pakistan and his 2009 decision to approve the Afghan Surge.

The bin Laden raid is generally considered to be one of Obama's signature foreign policy successes. We nevertheless argue that this case reflects negatively on Obama's relative foresight, because Obama offered a relatively pessimistic assessment of the chances that his intelligence services had found bin Laden's location. Conversely, although the Afghan Surge is generally considered to be one of Obama's major foreign policy failures, we explain how Obama appears to have anticipated the downsides risks of this policy better than his military advisers did.

Of course, these indications of relative foresight capture just a few of the many factors that mattered for shaping Obama's decisions. We explain below how Obama has received ample praise for other aspects of how he handled the bin Laden raid along with substantial criticism for pushing ahead with the Afghan Surge. We nevertheless show that these arguments inherently rely on value judgments or subjective probabilities. Distinguishing the elements of these cases that lend themselves to objective evaluation from those that do not thereby reveals how conventional approaches to judging foreign policy competence almost always lend

themselves to reasonable disagreement. By contrast, our framework for assessing relative foresight can establish objective, nontrivial insights about a decision-maker's performance.

The bin Laden raid

A U.S. special forces team raided Osama bin Laden's compound in Abbottabad, Pakistan on May 2, 2011. The operation killed bin Laden, seized a trove of documents regarding al-Qaeda's operations, and received immediate acclaim.

Journalists' accounts and participant memoirs provide a detailed narrative of how President Obama wrestled with uncertainty regarding bin Laden's suspected location.¹⁴ These sources describe how the Central Intelligence Agency (CIA) had several clues indicating that the Abbottabad compound was connected to al-Qaeda. However, there was no direct evidence that bin Laden was present at the compound. The head of CIA's bin Laden unit estimated a 95 percent chance that bin Laden resided there. CIA Deputy Director Michael Morell estimated those chances at 60 percent. Secondary sources and memoirs agree that most estimates clustered around 70 or 80 percent. A CIA "red team" designed to assess the information skeptically put those chances at roughly 40 percent. President Obama concluded this discussion by declaring that the chances bin Laden was in the compound were fifty-fifty.¹⁵

Obama thus appears to have been relatively pessimistic, relative to advisers who possessed similar information, regarding the chances that bin Laden was living in the Abbottabad

¹⁴ See, for example, Bowden (2012), Sanger (2012, 68-113); Morell (2015, 143-176), and Brennan (2020, 225-248).

¹⁵ These discussions are summarized in Friedman and Zeckhauser (2015).

compound. The simple average of estimates Obama received as described in the previous paragraph was 69 percent. The only reported estimate lower than Obama's was the red team's deliberately skeptical view, which could justifiably be excluded from an analysis of genuine opinions. The last paragraph also noted that reports on this case describe how several participants' assessments clustered in the 70-80 percent range. Given that bin Laden did, in fact, turn out to be living at the Abbottabad compound, this case offers a negative signal regarding Obama's relative foresight.

Of course, debates about the probability that bin Laden was living in the Abbottabad compound captured just one piece of Obama's broader deliberations about whether to authorize the raid. For example, Obama and his advisers reportedly devoted extensive discussion to whether the prospect of killing or capturing bin Laden was worth sparking a diplomatic crisis with Islamabad, which would almost surely object to U.S. forces conducting a combat mission on Pakistani soil. Obama also reportedly considered bombing the compound so as to minimize the risk of U.S. casualties, before concluding that a raid was superior because it minimized collateral damage and maximized opportunity to positively identify bin Laden on site. These considerations were doubtless important, yet they also depended on value judgments. There is no way to objectively measure the benefits of killing bin Laden relative to the cost of diplomatic fallout with Pakistan, the drawbacks of civilian casualties, or the importance of having special forces operators who could positively identify bin Laden's body.

Arguing that President Obama's deliberations about the bin Laden raid provide a negative signal of his foreign policy competence hardly implies that Obama's decision was harmful in

expectation. Most accounts of this case suggest that Obama's advisers were divided about whether to strike the Abbottabad compound. It is thus reasonable to conclude that Obama's choice had a clear, positive impact that other decision-makers might not have captured. Yet, foreign policy impact is not the same thing as foreign policy competence. By analogy, if you bet half of your life savings on a coin flip and you win, then that would have a large, positive impact on your life, but it would not indicate that you are a skilled decision-maker. Objectively evaluating decision-making competence requires identifying verifiable and unbiased components of that attribute. Section 2 showed that our concept of relative foresight meets these criteria; reporting on President Obama's deliberations for the bin Laden raid suggests that his assessment of uncertainty surrounding bin Laden's location was less accurate than the views that most other national security officials expressed based on similar information.

The Afghan Surge

In December 2009, President Obama announced that he would send 30,000 additional U.S. soldiers to Afghanistan as part of a policy that became known as the "Afghan Surge." The Afghan Surge helped the United States and its allies to recapture a substantial volume of key terrain in Afghanistan, but it failed to establish conditions for containing the Taliban insurgency. Today, the Afghan Surge is widely considered to have been a strategic failure that prolonged costly U.S. involvement in an unsuccessful war (Malkasian 2021, 218-314).

Yet, the fact that the Afghan Surge failed to obtain a better outcome does not provide an objective basis for questioning Obama's competence. Obama and his advisers had several

reasons to believe the stakes involved with fighting the Taliban were relatively high. Those stakes involved preventing Afghanistan from relapsing into a safe haven for terrorists to plot attacks on the United States, preserving Afghanistan's substantial progress with respect to economic development, democratization, and human rights, and avoiding a strategic defeat that could cause allies in other parts of the world to question U.S. military capabilities and resolve. Given these stakes, Obama approved a plan that carried a significant risk of failure. To conclude that this risk was excessive, one would need to estimate at least four parameters: the Surge's *ex ante* probability of success, the Surge's prospective costs, the degree to which a successful Surge would advance U.S. interests, and how these factors would compare to the expected utility of other options Obama could have pursued at the time. As Section 1 described, none of these factors can be estimated objectively.

Reporting on this case indicates that Obama was relatively pessimistic about the chances that the Afghan Surge would succeed as compared to the views of his military advisers. The Surge was originally requested by Obama's top general in Afghanistan, Stanley McChrystal, in August 2009. Most accounts of this episode report that Obama was skeptical of McChrystal's request, and therefore demanded a protracted series of debates questioning McChrystal's assumptions (Hastings 2012, 149-150; Malkasian 2021, 235-236). Surge advocates then leaked McChrystal's proposal to the *Washington Post* in order to make it politically difficult for Obama to override military advice. McChrystal and other top military leaders, such as Central Command (CENTCOM) chief David Petraeus and Chairman of the Joint Chiefs of Staff Michael Mullen, also gave public remarks endorsing the Afghan Surge and warned of dire consequences if Obama did not accept that recommendation. Secretary of Defense Robert Gates later wrote that these

actions were widely seen as “blatant lobbying” designed to force President Obama to approve the Afghan Surge despite his doubts that the policy offered significant prospects for success (Gates 2014, 367).

Unlike deliberations over the bin Laden raid, where key officials quantified their estimated chances that bin Laden was living in Abbottabad, there are no reports of Obama and his advisers explicitly debating the chances that the Afghan Surge would succeed. However, secondary sources and participant memoirs generally agree that Obama was relatively pessimistic about the Surge’s chances of stabilizing Afghanistan. Most key members of Obama’s national security team supported the Surge, including Secretary of State Hillary Clinton, Secretary of Defense Gates, Joint Chiefs Chairman Mullen, CENTCOM commander Petraeus, and General McChrystal. The principal opponent of the Surge among Obama’s national security team was Vice President Joe Biden, who argued that a more limited counterterrorism mission would be more politically sustainable, and thus more effective in the long-term (Marsh 2014, 270-275; McHugh 2015).

Obama’s objection to the Afghan Surge specifically revolved around his doubts that the strategy would work. Obama voiced those doubts continuously through several months of meetings. He later wrote that did not believe his military advisers had convincingly explained why augmenting U.S. forces in Afghanistan would significantly change the prospects for stabilizing the country. Instead, Obama characterized the Surge request as being primarily driven by “ideological and institutional concerns rather than by the objectives we’d set” (Obama 2020, 442). Even though Obama wrote this statement after the fact, it is credible because it involves

Obama admitting that he made a questionable decision. By contrast, Obama's incentives to posture for posterity would encourage exaggerating the degree to which the Surge plan seemed justifiable on its merits at the time.

Since relative foresight captures just one component of foreign policy competence, Obama's relatively pessimistic assessment of the Afghan Surge's prospects hardly absolves him from plausible criticism regarding other aspects of how he handled that debate. For example, a critic could argue that Obama's decision to approve the Afghan Surge was inappropriately dictated by political pressure. Yet, there are two reasons why such a line of reasoning cannot ground objective assessments of Obama's foreign policy competence. First, in order to conclude that political pressures caused Obama to make a suboptimal decision, it is necessary to prove that the Surge did not maximize national interests, in expectation. Otherwise, a proponent of the Surge could argue that yielding to professional advice on the matter is exactly what a competent decision-maker should have done. Refuting that argument requires demonstrating that the chances of the Surge succeeding were too low to justify pursuing the goal of defeating the Taliban. We showed in Section 2 that reaching such conclusions requires making value judgments and estimating subjective probabilities.

Moreover, even if scholars could somehow prove that the Surge did not advance expected national interests in Afghanistan, that would still be insufficient to prove incompetent decision-making given other factors that could have plausibly shaped Obama's reasoning. Presidents need to preserve their political capital in order to maintain their effectiveness at managing foreign policy (Larson 2003; George 2006, 74). Sustaining political criticism for overruling his

military advisers on the Afghan Surge could thus have damaged Obama's public standing in a manner that degraded his ability to advance U.S. interests in the long run. Evaluating this tradeoff once again requires estimating subjective probabilities (the chances that political criticism for rejecting the Surge would have reduced Obama's flexibility when handling other issues) and making value judgments (whether avoiding the costs of the Afghan Surge would justify sacrificing future foreign policy goals). By contrast, Obama's pessimism about the Afghan Surge provides a concrete signal that his relative foresight was superior to that of his military staff.

Section 4. Methodological Challenges to Measuring Relative Foresight

This section identifies what we view as the five principal methodological challenges involved with measuring relative foresight. Each of these challenges is analytically tractable, in the sense that none undermines the paper's central claim that relative foresight offers an objective metric for assessing foreign policy competence. Articulating these challenges nevertheless reveals that this framework has important limitations. The paper's concluding section builds on this point when connecting the paper's argument to broader questions about the nature and limits of foreign policy analysis.

One obvious methodological drawback to using relative foresight as a measure of foreign policy competence is that individual assessments of uncertainty provide limited evidence of decision-makers' predictive capabilities. Just as the world's best poker players lose many hands over the course of a session, a competent foreign policy decision-maker will sometimes appear to lack

relative foresight as a result of bad luck rather than incompetence.¹⁶ Yet, this problem can be solved by gathering more information. The more data we gather, the more reliably measures of relative foresight will track decision-makers' true talent, just as we would expect that strong poker players are more likely to win sessions that involve a larger number of hands. The Appendix explains how researchers can aggregate information across decisions in order to draw more confident conclusions about foreign policy decision-makers' competence.

Aggregating assessments of foreign policy competence across cases raises a second methodological question, which is how researchers should treat the fact that foreign policy choices vary across many dimensions.¹⁷ For example, some foreign policy decisions are particularly consequential. One might prefer to give the most consequential decisions the greatest weight when assessing foreign policy competence in order to grade leaders on the choices that matter most. Additionally, some foreign policy problems are less predictable than others. Perhaps these data points should receive less weight when assessing foreign policy competence, on the grounds that no one could be expected to perform well in those contexts. The idea of weighting foreign policy decisions based on importance and predictability follows the intuition that we would ideally evaluate leaders based on the "value added" they bring to foreign policy decision-making. That potential value undoubtedly varies from case to case, and an ideal measure of relative foresight would capture that variation.

¹⁶ For example, consider a decision-maker who correctly perceives that a policy has a 75 percent chance of success, whereas the median advisor believes the policy has a 50 percent chance of success. In one-quarter of these cases, the policy will fail and the decision-maker will appear to lack relative foresight, despite making a judgment that was more accurate than the norm.

¹⁷ We thank an anonymous reviewer for directing attention to this point.

Yet, as Section 2 explained, scholars lack clear criteria for quantifying the relative importance of different foreign policy issues (Baldwin 2000). Any attempt to weight relative foresight scores by issue importance introduces subjectivity by relying on researchers' value judgments. Moreover, even if researchers could quantify the degree-of-difficulty associated with predicting different outcomes of foreign policy decisions, it is not obvious how to incorporate that information into assessments of relative foresight. On the one hand, we might think that tougher challenges should receive *less* weight in assessing foreign policy competence, on the grounds that no one could be expected to perform well in such settings. On the other hand, we might think that tougher challenges deserve *more* weight when assessing foreign policy competence, on the grounds that the most capable leaders should be expected to handle the toughest calls. Researchers who are committed to objective analysis thus have no option but to treat all data points equally when assessing relative foresight.¹⁸

A third challenge in assessing relative foresight is that foreign policy decision-makers rarely quantify their assessments of uncertainty in a manner that resembles President Obama's discussion of the chances that bin Laden was living at Abbottabad (Dhami 2018; Friedman 2019). However, the paper's discussion of Sadat's handling of the October War and Obama's handling of the Afghan Surge shows that this lack of quantification does not preclude drawing inferences about relative foresight. Decision scientists have also developed empirically-

¹⁸ Some readers may object that treating all decisions equally is itself a value judgment, implicitly assuming that all foreign policy decisions yield equal analytic insight. Yet, this choice is a logical one to make in the absence of clear evidence to suggest otherwise. By analogy, sports commentators frequently argue over which athletes perform best "in the clutch," and there is clear value in excelling under pressure, but it is generally challenging to separate variations in "clutch" performance from random noise. As a result, more sports statistics are calculated by averaging performance across all conditions, without attempting to weight data unequally, just as we propose with respect to assessing relative foresight.

grounded techniques for translating verbal assessments of uncertainty into numeric percentages (e.g., Beyth-Marom 1982; Mandel 2015).¹⁹ These translations impart measurement error into any analysis, because decision-makers often use verbal probability terms in idiosyncratic ways. Yet, we noted earlier that such measurement error does not undermine the validity of using relative foresight as a tool for evaluating foreign policy competence. In principle, these inferred probabilities can be made more reliable and more precise by gathering larger quantities of data. Efforts to interpret verbal probability assessments thus satisfy the criterion of verifiability that we laid out in Section 1.

A fourth and related methodological issue is that it is always impossible to know what decision-makers “really” think. Archival records and participant accounts can capture what decision-makers write and say, but those public expressions may not reflect private beliefs. Prior to authorizing the raid on Abbottabad, for example, President Obama might have expressed unusually pessimistic views to shield himself from criticism in the event that the mission failed, and not because he actually thought there was merely a fifty-fifty chance that bin Laden was living in the suspected compound. The challenge of discerning private beliefs from public expressions surrounds a wide range of scholarship in history and political psychology and it does not invalidate our framework. Rather, we argue that this challenge amounts to another form of measurement error. As scholars collect more and better sources, they should be able to develop more precise inferences about decision-makers’ true beliefs.

¹⁹ For example, Mandel (2015) asked subjects to quantify their interpretation of verbal probability estimates in intelligence reports. He found that the median subject interpreted the phrase “very unlikely” as indicating a probability of 10 percent, “likely” as 75 percent, and so forth.

Finally, our discussion has thus far focused on evaluating the relative foresight conveyed in a single assessment of uncertainty. Competent decision-makers should also be able to update their beliefs effectively as new information becomes available. Thus, some decision-makers who initially hold inaccurate beliefs about the consequences of a prospective decision might rapidly improve their perceptions, whereas others cling more stubbornly to mistaken views (Bar-Joseph and McDermott 2016). Our framework can incorporate this dynamic by combining relative foresight scores across multiple assessments of a single decision problem. Thus, if President Obama had updated his assessment of the Afghan Surge's prospects from an initial assessment (which we can denote p_0) to a revised judgment (p_1), we could estimate relative foresight scores for those judgments separately and then average those scores together to develop a more comprehensive picture of Obama's relative foresight. This method can scale to cover multiple belief revisions.

Researchers could use similar data to address a separate question of whether decision-makers tend to update beliefs in the correct direction.²⁰ Such analyses could gauge decision-makers' capacity for learning, which is surely another important component of foreign policy competence. However, these measures would suffer an important drawback: decision-makers whose initial judgments were less accurate would then have greater opportunities to demonstrate effective learning. Thus, higher learning scores would not, by themselves, indicate foreign policy competence. By contrast, one advantage of our concept of relative foresight is

²⁰ We thank an anonymous reviewer for suggesting this extension.

that, when holding other factors constant, higher scores always indicate more desirable performance.

Section 5. Implications for Foreign Policy Analysis

This paper contributes to long-standing debates about the degree to which scholars can resolve questions about leaders' skill in making foreign policy decisions. By demonstrating that relative foresight can be measured objectively, our argument shows that good strategy is not entirely an illusion. Yet we also explained that relative foresight captures just one part of foreign policy decision-making. We also demonstrated that individual foreign policy decisions provide limited signals about leaders' predictive abilities, and described how aggregating those signals across cases requires tackling a series of methodological challenges. It may thus be fair to say that strategy is *mostly* an illusion, at least from the standpoint of researchers who seek to advance arguments that do not inherently rely on value judgments or subjective probabilities. Questions such as "Was Napoleon a strategic genius?" or "did Lyndon Johnson make the right decision to escalate the war in Vietnam?" will almost always leave room for reasonable people to disagree.

This observation suggests that researchers should be careful about asserting normative judgments about foreign policy decision-making. By extension, our analysis raises caveats for descriptive studies that advance nonrational explanations for foreign policy decision-making. In order to prove that foreign policy decisions were warped by miscalculation, domestic politics, or psychological factors, researchers must demonstrate that competent decision-makers should have known that a different course of action would have done a better job of advancing

national interests, in expectation (Glaser 2010, 2-3; Stein 2013, 369). Though careful scholarship can clarify the assumptions that researchers need to make in order to conclude that a foreign policy decision was nonrational, these holistic judgments are generally impossible to prove. By contrast, our case studies showed that narrower analyses of relative foresight can generate meaningful conclusions that do not depend on subjective judgments.

Our argument has at least three additional implications for foreign policy analysis. First, we have shown that drawing objective conclusions about foreign policy competence requires placing leaders' beliefs in the context of other views that were available at the time. Thus, while empirical scholarship on foreign policy decision-making often relies on scrutinizing leaders' assumptions in depth, we advocate for adopting a wider empirical lens that specifies how leaders and their advisers assess the chances of achieving their strategic goals. Our argument thus suggests broadening empirical evaluations of foreign policy leadership in some ways (by focusing on the distribution of informed views) while narrowing them in others (by focusing on assessments of uncertainty even though they comprise just one element of strategic decision-making).

A second implication of our analysis is that successful foreign policy decisions can reflect poorly on the judgment of leaders who are relatively pessimistic, while failed foreign policies can reflect well on the judgment of leaders who are relatively skeptical of a decision's prospects for success. This logic suggests that praise or criticism for foreign policy competence should be apportioned in a manner that is only partially related to decision outcomes. Yet, when scholars seek to identify the flaws in foreign policy decision-making, they overwhelmingly focus their

attention on policies that did not turn out well (Bar-Joseph and McDermott 2017). Our argument suggests that this reflects a potentially severe bias in how scholars conduct empirical research. Our concept of relative foresight provides a tool for avoiding that bias.

Finally, our analysis indicates that foreign policy competence can only be judged in relation to other views that were available at the time. This suggests that it will generally be impossible to objectively rank the competence of leaders who served in different contexts. For example, if we examined a large number of decisions, we might conclude that Henry Kissinger possessed more or less relative foresight than Madeleine Albright. Yet, this does not provide clear grounds for determining which of those U.S. Secretaries of State had better judgment about international affairs. Since Kissinger and Albright dealt with different problems and managed different teams, it is difficult to attribute any observed differences in their relative foresight to individual competence, as opposed to having faced problems with different degrees of difficulty, or working with advisers who possessed different capacities. All we can objectively infer when assessing foreign policy competence is how much insight decision-makers added to the specific contexts in which they served.²¹

²¹ A relevant analogy would be how sports statisticians estimate metrics of an athlete's "value over replacement." It is generally challenging to compare the quality of athletes who competed in different eras, given that it is hard to know the extent to which "replacement value" has shifted over time. Value-over-replacement statistics have nevertheless become ubiquitous in debates about athletic performance. This paper explains what a similar metric would entail when evaluating foreign policy competence.

Appendix: Quantifying Relative Foresight

An intuitive way to quantify relative foresight would be to compare a decision-maker's Brier Score to the average Brier Score across the distribution of viewpoints held by other observers who possessed similar information. Yet, this method will not produce an unbiased measure of relative foresight because different foreign policy problems generate different distributions of opinion. For example, if every individual who estimated the chances a policy will succeed gives an answer between 49-51 percent, then it would be impossible for anyone to receive a Brier Score that is more than 4 percent worse than the average. By contrast, if the range of estimates is evenly distributed from 10-90 percent, then the worst judgment would end up being 324 percent worse than average. Evaluating relative foresight by comparing leaders' Brier Scores to average Brier Scores will thus treat leaders differently depending on the kinds of topics they confront.

To solve this problem, we can standardize the difference between a leader's Brier Score and the average performance across the distribution of available views. Thus, when President Obama said there was a "fifty-fifty" chance that Osama bin Laden was living in Abbottabad, his judgment would receive a Brier Score of 0.25. Section 3 described how Obama received at least five additional estimates of the chances that bin Laden was living at Abbottabad (40 percent, 60 percent, 70 percent, 80 percent, and 95 percent). The average Brier Score across this distribution, including Obama's 50 percent estimate, is 0.15. The standard deviation across Brier Scores assigned to these estimates is 0.14. We would therefore say that President Obama's judgment was 0.74 standard deviations worse than the mean. Standardizing Brier

Scores in this manner captures a leader's position within the distribution of views that were available at the time, which is what the concept of relative foresight aims to measure.

In order to combine information about leaders' relative foresight across cases, we can simply average these standardized Brier Scores.²² A more sophisticated way to combine this information would be to employ a Bayesian framework that treats every prediction as a signal of a decision-maker's "true" relative foresight. However, this approach requires scholars to make assumptions about how those signals were distributed. For example, one intuitive way to construct such a model would be to assume that observed signals of relative foresight (e.g., Obama's score for assessing the chances that bin Laden was at Abbottabad) are normally distributed around leaders' "true" talent levels. But since there is no way to verify that assumption, it would undermine the effort to establish an objective standard for measuring foreign policy competence. We therefore propose that averaging standardized Brier Scores, which requires no distributional assumptions, is the most defensible way to quantify the quality of a leader's relative foresight.

References

Adkin, Mark. 1996. *The Charge: Why the Light Brigade Was Lost*. London: Leo Cooper, 1996.

Badie, Dina. 2010. "Groupthink, Iraq, and the War on Terror: Explaining U.S. Policy Shift Toward Iraq." *Foreign Policy Analysis* 6 (4): 277-296.

²² Thus, if Obama's standardized Brier Scores across four separate judgments turned out to be -0.74, +0.05, +0.25, and +0.80, then his average measure of relative foresight would be +0.46.

Baldwin, David. 2000. "Success and Failure in Foreign Policy." *Annual Reviews of Political Science* 3: 167-182.

Bar-Joseph, Uri. 2005. *The Watchman Fell Asleep: The Surprise of Yom Kippur and its Sources*. Albany, N.Y.: State University of New York Press.

----- and Rose McDermott. 2017. *Intelligence Success and Failure*. New York: Oxford University Press.

Bergen, Peter. 2012. *Manhunt: The Ten-Year Search for Bin Laden from 9/11 to Abbottabad*. New York: Crown,

Betts, Richard K. 2000. "Is Strategy an Illusion?" *International Security* 25 (2): 5-50.

Beyth-Marom, Ruth. 1982. "How Probable is Probable? A Numeric Translation of Verbal Probability Expressions." *Journal of Forecasting* 1 (3): 257-269.

Bowden, Mark. 2012. *The Finish: The Killing of Osama bin Laden*. New York: Atlantic Monthly.

Brennan, John. 2020. *Undaunted*. New York: Celadon.

Burchill, Scott. 2005. *The National Interest in International Relations Theory* (New York: Palgrave Macmillan).

Connelly, Owen. 2006. *Blundering to Glory: Napoleon's Military Campaigns*, 3rd ed. Lanham, Md.: Rowman and Littlefield.

Dhami, Mandeep K. 2018. "Towards an Evidence-Based Approach to Communicating Uncertainty in Intelligence Analysis." *Intelligence and National Security* 33 (2): 257-272.

-----, Ian K. Belton, and David R. Mandel. 2019. "The 'Analysis of Competing Hypotheses' in Intelligence Analysis." *Applied Cognitive Psychology* 33 (6): 1080-1090.

Downs, George W. and David M. Rocke. 1994. "Conflict, Agency, and Gambling for Resurrection: The Principal-Agent Problem Goes to War." *American Journal of Political Science* 38 (2): 362-380.

El-Gamasy, Mohamed Abdel Ghani. 1993. *The October War*. Cairo, Egypt: American University in Cairo Press.

Finnemore, Martha. 1992. *National Interests in International Society*. Ithaca, N.Y.: Cornell University Press.

Friedman, Jeffrey A. 2019. *War and Chance: Assessing Uncertainty in International Politics*. New York: Oxford University Press.

-----, Jennifer Lerner, and Richard Zeckhauser. 2017. "Behavioral Consequences of Probabilistic Precision." *International Organization* 71 (4): 803-826.

----- and Richard Zeckhauser. 2015. "Handling and Mishandling Estimative Probability: Likelihood, Confidence, and the Search for Bin Laden." *Intelligence and National Security* 30 (1): 77-99.

Gat, Azar. 1989. *The Origins of Military Thought: From the Enlightenment to Clausewitz*. Oxford, U.K.: Clarendon Press.

Gates, Robert. *Duty*. New York: Knopf.

Gelb, Leslie and Richard Betts. 1979. *The Irony of Vietnam: The System Worked*. Washington, D.C.: Brookings.

George, Alexander L. 2006. *On Foreign Policy: Unfinished Business*. Boulder, Col.: Paradigm.

Geva, Nehemia and Alex Mintz eds. 1997. *Decisionmaking on War and Peace: The Cognitive-Rational Debate*. Boulder, Col.: Lynne Rienner.

Glaser, Charles. 2010. *Rational Theory of International Politics*. Princeton, N.J.: Princeton University Press.

Hastings, Michael. 2012. *The Operators*. New York: Plume.

Herek, Gregory, Irving Janis, and Paul Huth. 1987. "Decision Making during International Crises: Is Quality of Process Related to Outcome?" *Journal of Conflict Resolution* 31 (2): 203-226.

Hermann, Margaret, Thomas Preston, Baghat Korany, and Timothy Shaw. 2001 "Who Leads Matters: The Effects of Powerful Individuals." *International Studies Review* 2 (2): 83-131.

Heuer, Richards and Randolph Pherson. 2020. *Structured Analytic Techniques for Intelligence Analysis*, 3rd ed. Washington, D.C.: CQ Press.

Hoffmann, Stanley, Samuel P. Huntington, Ernest R. May, Richard N. Neustadt, and Thomas C. Schelling. 1981. "Vietnam Reappraised." *International Security* 6 (1): 3-26.

Horowitz, Michael C., Allan C. Stam, and Cali M. Ellis. 2015. *Why Leaders Fight*. New York: Cambridge University Press.

Israeli, Rafael. 1985. *Man of Defiance: A Political Biography of Anwar Sadat*. Totowa, N.J.: Barnes & Noble Books.

Jervis, Robert. 1997. *Systems Effects: Complexity in Political and Social Life*. Princeton, N.J.: Princeton University Press.

Johnson, Dominic D. P. 2020. *Strategic Instincts: The Adaptive Advantages of Cognitive Biases in International Politics*. Princeton, N.J.: Princeton University Press.

Khong, Yuen Foong. 1992. *Analogies at War*. Princeton, N.J.: Princeton University Press.

Larson, Deborah Welch. 2003. "Politics, Uncertainty, and Values" in Stanley A. Renshon and Deborah Welch Larson eds., *Good Judgment in Foreign Policy* (Lanham, Md.: Rowman and Littlefield, 2003), pp. 309-319.

Leggiere, Michael. 2023. "Napoleon and the Strategy of the Single Point" in Hal Brands ed., *The New Makers of Modern Strategy: From the Ancient World to the Digital Age* (Princeton, N.J.: Princeton University Press), pp. 344-368.

Malkasian, Carter. 2021. *The American War in Afghanistan*. New York: Oxford University Press.

Mandel, David R. 2015. "Accuracy of Intelligence Forecasts from the Intelligence Consumer's Perspective." *Policy Insights from the Behavioral and Brain Sciences* 2 (1): 111-120.

Marrin, Stephen. 2012. "Is Intelligence Analysis an Art or a Science?" *International Journal of Intelligence and CounterIntelligence* 25 (3): 529-545.

Marsh, Kevin. 2014. "Obama's Surge: A Bureaucratic Politics Analysis of the Decision to Order a Troop Surge in Afghanistan." *Foreign Policy Analysis* 10 (3): 265-288.

McHugh, Kelly. 2015. "A Tale of Two Surges: Comparing the Politics of the 2007 Iraq Surge and the 2009 Afghanistan Surge." *SAGE Open* 5 (4): 2158244015621957.

McManus, Roseanne W. 2021. "Crazy Like a Fox? Are Leaders with Reputations for Madness More Successful at International Coercion?" *British Journal of Political Science* 51 (1): 275-293.

Mearsheimer, John J. and Sebastian Rosato. 2023. *How States Think: The Rationality of Foreign Policy*. New Haven, Conn.: Yale University Press.

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip Tetlock. 2015. "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions." *Perspectives on Psychological Science* 10 (3): 267-281.

Mintz, Alex and Carly Wayne. 2016. *The Polythink Syndrome: U.S. Foreign Policy Decisions on 9/11, Afghanistan, Iraq, Iran, Syria, and ISIS*. Stanford, Calif.: Stanford University Press.

Morell, Michael. 2015. *The Great War of Our Time*. New York: Twelve.

Mueller, John E. 1980. "The Search for the 'Breaking Point' in Vietnam: The Statistics of a Deadly Quarrel." *International Studies Quarterly* 24 (4): 497-519.

Neustadt, Richard and Ernest May. 1986. *Thinking in Time: The Uses of History for Decision Makers*. New York: Free Press

Obama, Barack. 2020. *A Promised Land*. New York: Crown.

Rathbun, Brian C. 2019. *Reasoning of State: Realists, Romantics, and Rationality in International Relations*. New York: Cambridge University Press.

Reiter, Dan and Allan C. Stam. 2002. *Democracies at War*. Princeton, N.J.: Princeton University Press.

Renshon, Stanley A. and Deborah A. Larson eds. 2003. *Good Judgment in Foreign Policy*. Lanham, Md.: Rowman and Littlefield.

Sanger, David. 2012. *Confront and Conceal: Obama's Secret Wars and Surprising Use of American Power*. New York: Crown.

Schafer, Mark and Scott Crichlow. 2010. *Groupthink Versus High-Quality Decision Making in International Relations*. New York: Columbia University Press.

Shazly, Saad. 1980. *The Crossing of the Suez*. San Francisco, Calif.: American Mideast Research.

Smith, Alastair. 1998. "International Crises and Domestic Politics." *American Political Science Review* 92 (3): 623-638.

Stein, Janice Gross. 2013. "Threat Perception in International Relations" in *The Oxford Handbook of Political Psychology*, 2nd ed., ed. Leonie Huddy, David O. Sears, and Jack S. Levy. New York: Oxford University Press.

Tetlock, Philip E. 2005. *Expert Political Judgment: What Is It? How Can We Know?* Princeton, N.J.: Princeton University Press

Vickers, Geoffrey. 1965. *The Art of Judgment: A Study of Policy Making*. New York: Basic Books.

Yarhi-Milo, Keren. 2018. *Who Fights for Reputation: The Psychology of Leaders in International Conflict*. Princeton, N.J.: Princeton University Press.