UNITED STATES INSTITUTE OF PEACE www.usip.org

SPECIAL REPORT

1200 17th Street NW • Washington, DC 20036 • 202.457.1700 • fax 202.429.6063

Jack A. Goldstone

Using Quantitative and Qualitative Models to Forecast Instability

Summary

- Preventing violent conflict requires early warning of likely crises so that preventive actions can be planned and taken before the onset of mass violence.
- For most of the post–World War II period, policymakers and intelligence agencies have relied on experts to make qualitative judgments regarding the risk of instability or violent changes in their areas of study. Yet the inability of such experts to adequately predict major events has led to efforts to use social and analytical tools to create more "scientific" forecasts of political crises.
- The advent of quantitative forecasting models that give early warning of the onset of political instability offers the prospect of major advances in the accuracy of forecasting over traditional qualitative methods.
- Because certain models have a demonstrated accuracy of over 80 percent in early identification of political crises, some have questioned whether such models should replace traditional qualitative analysis.
- While these quantitative forecasting methods should move to the foreground and play a key role in developing early warning tools, this does not mean that traditional qualitative analysis is dispensable.
- The best results for early warning are most likely obtained by the judicious combination of quantitative analysis based on forecasting models with qualitative analysis that rests on explicit causal relationships and precise forecasts of its own.
- Policymakers and analysts should insist on a multiple-method approach, which has greater forecasting power than either the quantitative or qualitative method alone. In this way, political instability forecasting is likely to make its largest advance over earlier practices.

About the Report

The United States Institute of Peace's Center for Conflict Analysis and Prevention commissioned this report for its ongoing project examining new approaches to early warning for political instability and mass violence. Analysts generally agree that the policy process benefits from both results of statistical models and qualitative expert judgment. But where judgments from qualitative and quantitative models diverge, decision makers are frequently left without a sound strategy for preferring one result over the other or resolving differences between them. Drawing on his experience in developing and using qualitative structural analogy models and quantitative statistical models (including for the Political Instability Task Force), Jack Goldstone provides practical guidance on how different models can be used together to generate more accurate forecasts.

Jack A. Goldstone is the Virginia E. and John T. Hazel Jr. Professor at the George Mason School of Public Policy and an Eminent Scholar. His work has focused on social movements, revolutions, forecasting instability, and issues in state-building and reconstruction. The author or coauthor of nine books, Professor Goldstone is a leading authority on regional conflicts, has served on a U.S. Vice-Presidential Task Force on State Failure, and is a consultant to the Department of State, the Federal Bureau of Investigation, and the U.S. Agency for International Development. Professor Goldstone received his PhD from Harvard University.

Special Report 204 March 2008

Contents

Introduction

There is a strong tendency in both government and private enterprise to seek out the "best" method of doing a task and then rely on that method. In manufacturing or policy design, this approach makes a great deal of sense—why use a mix of best and second-best methods if one approach has been demonstrated to be clearly superior?

In forecasting violent conflict, the traditional method has been to rely on expert analysis of individual countries or regions, drawing on the knowledge of policy professionals and academics who have worked on the country or region in question for a substantial period of time. Such qualitative analysis generally rests on the expert's subjective analysis of a mix of sources, including news reportage and other media sources, other open-source data, and embassy and intelligence reports.

Recently, however, a number of quantitative models, based on objective analysis of open-source data, have been offered to analysts as a supplement to their traditional analysis. Notable examples are the military's Analyzing Complex Threats for Operations and Readiness (ACTOR) model and the Political Instability Task Force (PITF) model.[1] These models are developed by training algorithms on historical data, usually examining several decades in the post–World War II era, to arrive at factors that are effective at separating countries at high risk for near-term instability from those likely to remain stable. They use variables that indicate weakness in a state's ability to manage conflict or conditions that are likely to precipitate violence. Such variables include details of regime institutions, the nature of regime elites or ideologies, levels of income per capita or infant mortality, membership in international organizations, the level of conflict in surrounding states, the incidence of economic or political discrimination or exclusion of population groups, economic and demographic trends, and histories of prior conflict.

Many policy and intelligence professionals have been understandably reluctant to embrace these models. The quantitative models seem to dispense with the key problem addressed in traditional analysis, namely, identifying the motivations and likely actions of key actors, in favor of juggling a number of more abstract systemic indicators. The models also seem to set aside areas deemed crucial in traditional analysis, such as the history, culture, and contingent events of specific countries or communities.

Moreover, experts defend their judgments as not merely "gut-level" analysis. Rather, expert analysis usually draws on causal mechanisms, based on qualitative models or analytical principles, drawn from the latest social science research on revolution or regime change. It is thus not surprising that many analysts feel that their range of sources, their solid social science foundations, and their own experience and judgment will usually provide results as good as or better than those obtained from quantitative forecasting models.

This divergence of methods creates a dilemma: Should experts continue to rely on their own traditional methods of analysis and judgment? Or should they instead shift to relying on quantitative models for forecasting coming crises?

In fact, forecasting is a very different enterprise than manufacturing a product or designing a policy to achieve a specific end. In both of these cases, the task is to reach a known end with maximum efficiency. In forecasting, the goal is to identify an unknown (in this case, the precise risk of near-term instability). In identifying an unknown, better results can often be gained by triangulating with several different approaches. Therefore, even if one approach has seemed to yield better results than another, one can often get even better results by using a combination of best and second-best approaches to jointly triangulate on the unknown outcome.

The International Institute of Forecasters (IIF) has done extensive work on choosing among many methods of forecasting. In the case of forecasting large or sudden changes, where one cannot simply extrapolate from current trends, they note that it is imperative to have an explicit causal model that identifies those factors that affect the system's stability. They note that there are two general kinds of such models: quantitative models that

generate predictions of change from input variables and structural analogies that identify qualitative changes that have led to instability in a variety of exemplar cases. The IIF also notes that where one has data for both approaches, it is often useful to combine the approaches to produce an overall forecast.[2]

This report discusses the principles according to which one should or should not combine quantitative models and structural analogies in forecasting political instability. The goal is not to promote one method or the other, but to describe how using both methods in the most appropriate manner can yield superior forecasts.

Quantitative Forecasting Models and Structural Analogy Approaches

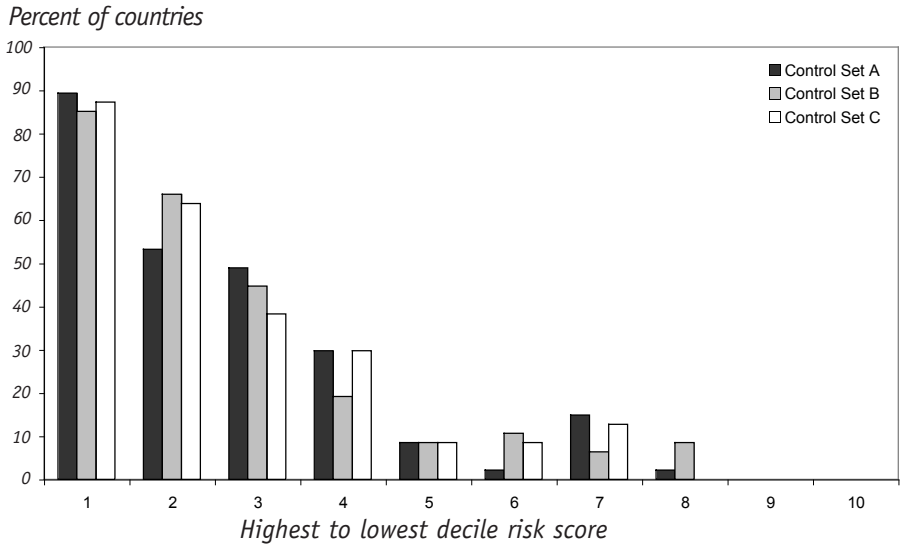
Quantitative forecasting models assume that the risks of near-term crises and violent conflict in a given county at any moment in time can be largely (but not entirely) presented as a function of a discrete number of variables, and the level of those variables.[3] These models generate a "risk score" for each country, ranging them along a scale from low to high levels of risk of near-term conflict. To test these models for accuracy, one can create a binary division of countries into "at-risk" and "not-at-risk" categories by setting a cutoff point on the risk scale, placing countries above a certain level in the "at-risk" category, and those below that level in the "not-at-risk" category. The set of countries designated as "at risk" is then compared with the observed list of countries that did experience a political crisis in a given time period. Similarly, the set of countries designated as "not at risk" is compared with the observed list of countries that remained stable. The accuracy of the model can then be given as the average of these two comparisons. Thus if the model has 65 percent of the observed crises in its "at-risk" category and 75 percent of the stable countries in its "not-at-risk" category, we would say its overall accuracy is 70 percent.

In an ideal world for forecasting, all countries would either have a 100 percent certainty of experiencing violent conflict in the near term, or a 0 percent chance of developing such conflict, and a quantitative forecasting model would accurately place all, or nearly all, cases in the proper category. In reality, of course, the world is not that sharply divided. Some countries are very unstable and at high risk of violent conflict if any precipitating event—such as a mass demonstration, an assassination, a rigged election, violence in a neighboring country, or an economic crisis—should occur. Other countries are very stable and will avoid conflict even if such events do occur. Yet other countries are somewhat unstable and are likely to fall into violent conflict if a large-scale precipitating event, or a combination of such events, occurs, but are likely to remain stable if such events are few or minor in scale. Thus, if the actual occurrence of violent conflict depends on a combination of a country's inherent instability level, plus the presence of a precipitating event of sufficient magnitude whose occurrence is unpredictable, errors in forecasting are inevitable. It is possible to be highly accurate regarding stable countries (where precipitating events are unlikely to matter, as stability will dominate), and fairly accurate regarding very unstable countries (where almost any kind and size of precipitating event will likely provoke a conflict), but difficult to be accurate on moderately unstable countries, where actual conflict will depend most on whether precipitating events of a certain magnitude occur, which is inherently unpredictable.

If the actual occurrence of violent conflict depends on a combination of a country's inherent instability level, plus the presence of a precipitating event of sufficient magnitude whose occurrence is unpredictable, errors in forecasting are inevitable.

Figure 1 shows the observed incidence of political crises (either civil or revolutionary war, genocide, or violent or sudden decline of democracy) according to the deciles of risk scores in the latest PITF model. The model is extremely accurate in predicting stability for the lowest risk countries (note that there are zero crises observed in model deciles 9 and 10). It is also fairly accurate for predicting near-term crises for the very highest risk countries (roughly 90 percent of countries in the first decile of risk scores were observed to experience political crises within two years of observation). Nonetheless, there remain a number of countries at middling risk, those with risk scores in the second through fourth deciles, where the observed incidence of crises is roughly 30 to 60 percent. Hence, where

FIGURE 1: **Countries experiencing political crises two years after observation, 1955–2003** *(by decile of risk score in PITF global forecasting model, for all onsets of political crises and three randomly drawn control sets of stable countries.)*



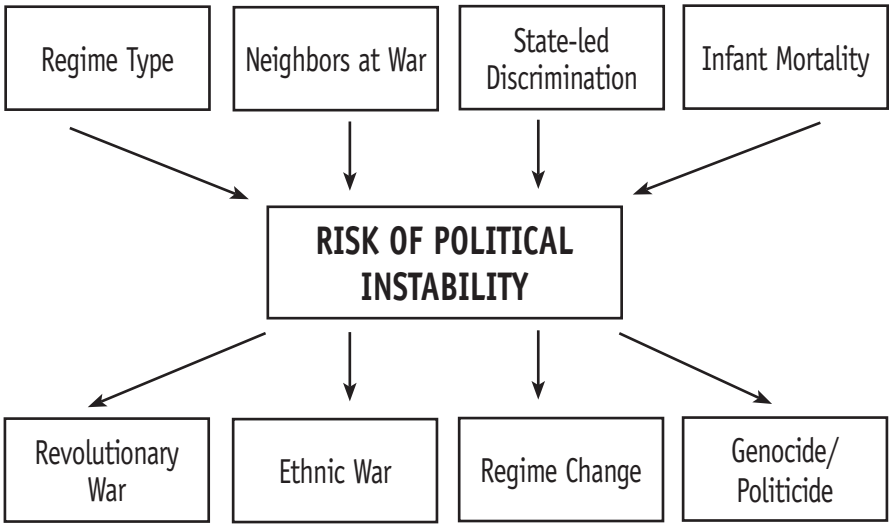
any binary designation of countries as either "at risk for near-term conflict" or "not at risk" will necessarily produce missed predictions. That is, even if the model is completely accurate in predicting that 60 percent of those countries in the second risk decile will experience political crisis, the model does not and cannot tell us which of those countries will in fact have conflicts and which will remain stable. Thus, even a model that is highly accurate in terms of ranking countries will have considerable uncertainty for some groups of countries regarding what will actually happen in that group.

The quantitative forecasting models are also somewhat constrained by the need to have fairly complete data—a model is no good for discriminating and identifying likely crises around the world if data are only available for half or fewer of the countries, or only for the wealthiest and data-rich countries. The timeliness of the models is also limited by the frequency with which data are updated and the accuracy of the data. Thus, if quantitative models with useful predictive values depended on data that had to be extremely complete, timely, and accurate for numerous variables for each country of the world, they would be of little value in practice. Fortunately, that is not the case. The PITF model, for example, is able to attain accuracy of over 80 percent in identifying countries that will have, or will not have, major political crises two years after the data of the observation period, using just four variables—the regime type (derived from the annually updated Polity data set on regime characteristics)[4]; infant mortality (estimated annually from UN data)[5]; the presence or absence of high levels of discrimination (derived from the Minorities at Risk data set, which gives annual data on groups facing discrimination)[6]; and the number of neighboring countries with violent conflicts (obtained from the annually updated Armed Conflict and Intervention data set).[7]

The PITF model is able to attain accuracy of over 80 percent using just four variables—the regime type, infant mortality, the presence or absence of high levels of discrimination, and the number of neighboring countries with violent conflicts.

The very simplicity of this model, however, raises questions among those trained to think in terms of complex interrelationships among key actors as the major factors governing risks of near-term instability. For this reason, among others, structural analogy models have great appeal.

FIGURE 2: **Diagrammatic depiction of PITF global model**



Structural analogies are a mode of analysis based on identifying key similarities across countries. If detailed case studies of several countries on the edge of revolution find a set of repeated relationships—such as conflict between states and elites, states under pressure from external conflict or internal economic crises, and mobilization of worker or peasant communities (e.g., Theda Skocpol's study of revolutions[8])—then one way to ask whether a given country is on the edge of revolution is to examine that case and see if similar conditions can be identified. What is important here is how well an analyst can identify meaningful similarities, as opposed to coincidental or superficial similarities. For example, if conflict between states and elites is a critical forerunner of revolutionary conflicts, then the analyst needs to be able to determine whether the conflicts observed in a given state are deep and likely to drive elites into revolt or superficial and likely to be easily resolved. Knowing how to "read" elites and what signs to look for to interpret group interests and likely actions are the crucial skills in using this particular structural analogy to guide forecasting.

Analysis based on structural analogies relies heavily on expert knowledge regarding the key relationships, interests, and capabilities of various actors and groups in the society being studied. Where the analogy reveals multiple overlaps and similar dynamics to well-studied cases of conflict onset, analysts can have high confidence that their projections regarding future conflict are accurate. But one must guard against misuse of analogies based on factors that seem similar, but are in fact unfolding in different ways or can be thrown off by different contingencies in different societies.[9]

It is much harder to estimate the accuracy of predictions using analogies, since the level of detailed analysis required means that analysts rarely venture predictions for dozens or hundreds of countries (like the quantitative forecasting models), against which success rates for prediction can be measured. Instead, analysts using structural analogies are likely to give an estimated forecast accuracy based on how close they think their country of study is to the model cases used to ground the analogy. This method of analysis thus tends to give a projection for risks of conflict that is specific to individual countries (e.g., how closely does Iran in 2008 resemble Russia in 1916?) rather than assigning relative risks to clusters of countries (e.g., does Iran in 2008 belong in one of the higher or lower deciles of risk compared to other countries in 2008?).

Analysts using structural analogies are likely to give an estimated forecast accuracy based on how close they think their country of study is to the model cases used to ground the analogy.

Combining Models for Better Forecasts

Forecasting unknowns is fraught with difficulty. In forecasting instability, no approach is likely to yield near-perfect results. This is in part due to problems with incomplete or inaccurate data, but is also in part due to the nature of forecasting and the character of crisis onset itself.

Regarding data problems, the data needed to make forecasts for any given country are often incomplete, missing, or biased. This is true of both quantitative and qualitative data. Thus quantitative data on income per capita, infant mortality, regime type, total GDP, trade, and other factors are likely to have errors, especially for countries with less than highly sophisticated statistical offices. Even for the United States, GDP figures are routinely revised by substantial amounts in the year after the initial report. A 5 percent error range on most quantitative data seems a minimal allowance for errors in reporting and estimating the data; yet models that combine several variables that have this range of error are likely to have some compounding and result in errors of 10 percent or more in their final output.

Regarding qualitative data, the exact degree to which elites are in harmony or conflict, to which opposition movements have popular support, or to which the ruler is supported by neighboring or foreign states, is not always easy to pin down. Actors' public and private statements may differ, and they may intentionally conceal their preferences pending events. Journalistic accounts may be biased because of their sources or reporters' own views. Events and elite and popular motivations at the local level may be even harder to ascertain on a countrywide basis.

It is also likely that even if perfect data were available, forecasting models would be less than perfect. Quantitative forecasting models necessarily are based on *average outcomes* across countries and cannot take into account every individual factor in each particular country that might affect its stability. To the extent that such individual factors increase or decrease the risks of instability in a given country at a given time, even well-tested and fairly accurate quantitative models will exhibit specification errors. Even for qualitative models, the analyst's ability to identify and assess the key factors affecting a given country may be limited by everything from overreliance on a certain data source (newspapers, perceptions of embassy staffers) to a focus on certain actors or kinds of events that played key roles in the prior history of that country. Country experts using qualitative models are particularly prone to errors from overlooking factors that may affect the model if such factors have not previously been significant in their studies of that country. In short, even models that approach a perfect command of the *common factors that affect instability across different countries* cannot reflect all possible interactions or added effects with factors that are specific to individual countries at a certain time.

Even models that approach a perfect command of the common factors that affect instability across different countries cannot reflect all possible interactions or added effects with factors that are specific to individual countries at a certain time.

The above considerations are not meant to justify acceptance of poor results. A model with accuracy of 50 percent is not useful, period—a coin flip is a more economical and equally accurate approach. However, it does suggest that most forecasting models, whether qualitative or quantitative, will not exceed 80-85 percent accuracy; with perhaps 90 percent for some models that focus on countries that are unusually similar in their characteristics and behavior. Much also depends on the degree to which analysts are open to seeing new data, or their ability to judge how unexpected or novel factors interact with the factors specified in the analytic models.

Given that both qualitative and quantitative approaches will most likely be limited by problems with the accuracy and completeness of both their data and their analytic models, what is the best way to proceed to maximize accuracy of prediction?

The greatest benefits come from using approaches that are as *independent as possible*. Models that differ slightly in specification but rely on basically the same data will have some benefits in triangulation, but models that differ substantially in specification and rely on different data will have the most benefits. There are two considerations that lead to this conclusion. The first, dealt with in the next section, has to do with

the arithmetic of independent predictions. The second, dealt with in the following section, has to do with the substantive way in which independent models can help analysts to probe their forecasts and identify potential errors or problems in their analysis. It may be difficult to accept that using two models, one of which is less accurate, will give results superior to that of the better model alone. Yet if the models are independent, meaning that they draw on such different sources that their conclusions are able to vary widely, that is in fact the case. This can be demonstrated in a simple example using varied dice to predict outcomes.

An Illustrative Example of Combining Forecasts

Imagine two analysts rolling dice in separate rooms. The first analyst has a magic die such that when she names a country and rolls it, a roll of 1 is always an accurate forecast of instability, a roll of 2 is accurate half the time in forecasting stability, and rolling other numbers (3 to 6) gives an always-accurate forecast of stability. This is a very good but still imperfect forecasting tool, as it is accurate only 5.5/6 of the time and is wrong 0.5/6 of the time. Its accuracy is thus 91.7 percent.

Now imagine the second analyst has a slightly inferior die: A roll of 1, 2, or 3 is only 2/3 accurate when taken as a prediction of instability, while a roll of 4 to 6 is always accurate in predicting stability. This "model" is accurate 5/6 of the time, or 83 percent.

Here is the question: If you want the best results, should you always rely on only the first analyst and her magic die? Or should you check with both?

Using only the first analyst is not too bad; you will have some accurate predictions of instability (roll of 1) and many accurate predictions of stability (rolls of 3 to 6). But what about the 17 percent of cases where the analyst rolls a 2? You then will be unsure of what to predict.

At those times, it is enormously valuable to have the second analyst. If the second analyst rolls a 4 to 6, you know for certain that the ambiguous 2 should be taken to indicate stability. If the second analyst rolls a one or a two or a three, you know with 2/3 certainty that the ambiguous 2 should be taken to indicate instability.

The chance of an error using both analysts is thus substantially reduced. The chances of a faulty prediction using both analysts has been reduced to 1/6 (the chance of analyst one rolling a 2) times 1/2 (the chance of analyst two rolling a 1 or 2 or 3) times 1/3 (the chances of the 1 or 2 or 3 rolled by analyst two being wrong in its prediction). That is, the chance of a wrong prediction when using both analysts is only 2.8 percent.

That is to say, by starting with a model that is accurate 91.6 percent of the time and adding a model that is less accurate (correct only 83 percent of the time), judicious combination of their results produces a joint analytic system that is going to be correct 97.2 percent of the time.

This result may seem counterintuitive—shouldn't adding a less accurate model reduce the overall accuracy of the prediction? The answer is no, *provided that the models are independent* (or orthogonal, in formal terms—that is, the results of one model do not influence the results of the other). The reason is that in identifying an unknown, the additional triangulating power of a second model adds to the identification, even if the second model is less powerful than the first. As long as the second model has some substantial power in its own right, its use in combination with a stronger, independent model improves the overall accuracy of prediction.

More generally, having at least two independent approaches to assessing instability, if they point in the same direction, greatly increases the confidence of predictions (how to handle the case in which they point in opposite directions is examined in the next section). If we consider the possibility of using a qualitative forecasting model based on structural analogies in conjunction with the PITF quantitative model, in neither case do we have a "magic die" in which we know exactly which predictions are correct and which

Having at least two independent approaches to assessing instability, if they point in the same direction, greatly increases the confidence of predictions.

are not (although the PITF predictions of very low risk of instability are virtually always correct). What we have is a quantitative model that is known to be about 80 percent accurate, and a qualitative model that has unknown accuracy, but which can also be fairly good, although not perfect because of issues with data and specification noted above. For the sake of argument, let us say that the qualitative model is just as good—at least 80 percent accurate.

If we have two models that identify cases as "stable" or "unstable," and both are 80 percent accurate, then by using one model alone, we can never get more than 80 percent of predictions of instability correct. However, if we use both models, and they both predict a country to be unstable, our confidence in that prediction is much greater, as the odds that both models could be wrong is only 4 percent. That is, predictions of instability made jointly by both models would be correct 96 percent of the time, not just 80 percent. Similarly, any one model would only accurately identify stable cases correctly 80 percent of the time, while countries that appear as stable in both models would be stable 96 percent of the time. Thus combining both models, where they agree, would cut the error rate from 20 percent down to 4 percent; this is a substantial gain in accuracy!

Where the two models disagree—that is, one method identifies a case as "stable" and the other identifies the case as "unstable"—should you just use the model that you believe to be more accurate? Again, no—there is no sound basis for doing this. Let us say one model is believed to be slightly better. If one model is wrong 20 percent of the time, and the other model is wrong only 15 percent of the time, then by picking one model you at best improve your assessment accuracy from 20 percent to 15 percent (assuming you know for sure which model is most accurate!). However, if you use the results of each model to interrogate the other to check your confidence in those results, you can increase your accuracy much further. This is essentially the process used above in the dice example of consulting both dice. We will discuss the procedures for doing this in the next section. The main point here is that if you interrogate both models to determine which model is more likely correct, this procedure can reduce your error rate to 10 percent or less.

If you interrogate both models to determine which model is more likely correct, this procedure can reduce your error rate to 10 percent or less.

The bottom line is that in forecasting, it is always better to have two distinct models than one, even if one method is believed to be superior, as long as *both* methods have some predictive power and the two methods are independent.

Of course, if you had a single model that was known to be 98 percent accurate or better, the gains from using a second model in conjunction would be smaller, and probably not vital. However, since all known political forecasting models have errors, and the models are not more than 80 to 85 percent accurate, quite substantial increases in predictive accuracy and confidence would result from using two models together, rather than relying on just one.

Using Two Models When They Predict Different Outcomes

We have noted above that the crucial factor in increasing triangulation power of multiple forecasting models is that they rely on different principles, so that one cannot simply deduce from the outcome of one analysis what the outcome of the alternative analysis will be.

The quantitative and structural analogy models noted above fit this condition quite well. The quantitative model is based on classing an individual country according to relative risk, compared to other countries, depending on the level of a risk function derived from the values of a set of open-source variables, and how that level compares with the average level derived from a large number of countries observed to have experienced near-term crises. The qualitative or structural analogy model is based on an expert assessment of how closely conditions in a given country resemble conditions in a small group of "model" countries known to have experienced ensuing conflict. Even if the quantitative and qualitative analysis use some of the same key factors (for example, infant mortality

as a variable to indicate poverty or grievances), it is unlikely that one could tell from one analysis how the other analysis will unfold. This is because the structural analogy metric is more a matter of conforming to a "model" involving a large number of simultaneous relationships, while the quantitative forecasting model uses metrics that produce a continuous scale across hundreds of cases.

Given these differences, by using both models together, and using each to interrogate the other, one can often identify which model is at fault. After all, if one model identifies a case as stable, and the other model suggests it is unstable, one model is wrong. If you take each model in isolation, you do not know which is wrong; you only have the known error rate as a guide. But if you compare the two models by examining their data and assumptions, and judging how well each model fits its assumptions, you have a much better chance to determine which model is wrong. Once you determine which model is most likely wrong, you can much more confidently take the other model as correct. Indeed, the more certain you are which model is giving a false prediction, the more certain you can be of making the right forecast.

It is generally much easier to determine the relative soundness of two models dealing with one case than to compare the soundness of two models across a wide array of cases, or to put an absolute number on the fidelity of one model to one case. Given a careful examination of the data and assumptions that underlie two distinct models, in light of detailed study of that case, one can often gain very high confidence—90 percent or better—that one model is more accurately grounded than the other. That 90 percent confidence then becomes the confidence level in choosing which model's predictions apply to that case, rather than the 80 percent or so confidence that is a measure of the accuracy of one model against all cases.

Since I have worked on the development of both structural analogy models and quantitative models, I will use these as examples in the rest of this paper.[10] The qualitative model of state breakdown that I have developed is similar in many ways to the PITF quantitative model. Both emphasize state capacity, elite divisions, and popular mobilization, but they come at these in different ways. The qualitative model asks the analyst to determine the level of strain on state financial resources, while the PITF model simply looks at infant mortality as a measure of the level of economic development of the country as a whole—poorer countries usually have poorer and less capable states. The qualitative model asks the analyst to determine the level of divisions among the elites and cleavages between the ruler and elites; the PITF model uses the Polity IV factionalism and regime-type scores to make a similar assessment. The qualitative model asks the analyst to determine the potential for mass mobilization: Are there common economic or political grievances capable of motivating significant groups? Do those groups have the capacity—through neighborhoods, villages, urban districts, and local leaders—to organize for collective action? The PITF model focuses on concrete evidence of conditions that would usually lead to shared grievances: high levels of political or economic discrimination and conflicts in neighboring states that could motivate or inflame disorder within a country.

The PITF model focuses on concrete evidence of conditions that would usually lead to shared grievances.

While both models seek to get at the same general issues behind state vulnerability, the fact that they do so using different methods and focusing on different specific variables and data suggests that they are in fact largely independent. Thus, if their results differ, it is most likely because in one model the variables observed are not reflecting the true conditions behind the model or theory, and thus the forecast of that model should be set aside. In that case, the analysts need to look beyond the factors in the model to try to determine what is "throwing off" one analysis or the other.

For example, a country may have a relatively high level of infant mortality, indicating a weak or ineffective state and meager resources for the government. Yet a qualitative case analysis of that country may find that the government's revenues are solid and that patronage and military provision by the ruler looks sustainable. How could that be? The answer is that the government is either enjoying privileged access to resources that are not indicative of broader economic prosperity (e.g., the rulers are exploiting resources

such as oil or diamonds and controlling that wealth for themselves) or the government is getting foreign aid that is stabilizing and extending its finances. But it is then incumbent on the qualitative analyst to show that such factors are in play and causing their results to differ from the PITF model. One can then be confident that the PITF model is being misled by not catching those other factors.

For example, one of the countries that the PITF model consistently misread was The Gambia, which was an anomalously stable democracy in West Africa from 1970 to 1994. The Gambia should have been unstable, given its poverty and regime type. Yet The Gambia, as a former British colony, gained substantial reexport and smuggling revenues from being the only country in the region outside the French-supported West African (CFA) franc currency zone, where France contributed to keeping currencies stable but overvalued by pegging the CFA franc to the French franc. Most Gambians, and especially the government, gained from the revenue streams that flowed from cross-border trade. However, in 1993, Senegal closed its borders with The Gambia, drying up the stream of cross-border goods, and in January 1994, France devalued the CFA franc by 50 percent, drastically cutting The Gambia's smuggling revenues and opportunities. Within a year of these events, the democratic government was overthrown by the military, which seized power in July 1994.

Another example is Thailand, where the model predicted stability under the Thaksin premiership. What the model did not pick up were the allegations of corruption against Thaksin and the maneuvering of the military to obtain the king's support for a military coup. In this case, if a qualitative analyst had pointed to these factors and forced the quantitative modeler to track Thailand more closely on a month-to-month basis, the PITF model would most likely have noticed that the 2005 designation of the regime as a "full democracy" no longer held after the boycotted April 2006 elections. In this case, the qualitative analysis would have shown clear evidence of factionalism and increasing isolation of the Thaksin regime from major institutional and urban supporters that was not incorporated into the quantitative model based on annual (hence 2005) data. This should have led analysts to discount or revise the assessment of the quantitative model and favor a higher risk assessment than would be obtained from the quantitative model alone. The lesson here is that qualitative analysis may uncover factors outside the PITF model that lead to a contrary result, but the sustainability of those factors then must become an important part of the analysis. If they are removed, the predictions of the PITF model then become more likely to be realized.

A country with high infant mortality and hence limited state effectiveness and resources may truly have fiscal problems that a qualitative analyst might overlook.

The reverse may also be true. That is, a country with high infant mortality and hence, according to the PITF, limited state effectiveness and resources, may truly have fiscal problems that a qualitative analyst might overlook. That is, the government's finances might be more fragile than they appear—dependent on foreign assistance that might be withdrawn or built on debt that is not sustainable in the long term. Such a state might be more susceptible to fiscal crises in the event of an economic downturn.

With regard to elite divisions, where the models give different predictions it is important for the qualitative analyst and the PITF analysis to confirm their data. The most powerful factor in the PITF model in raising the risks of instability is a particular set of regime characteristics, namely the combination of partial democracy and factionalism. This combination apparently injects elite divisions into politics in a highly combustible way. If a qualitative analysis argues that elite divisions are not severe in a country that PITF scores as partially democratic and factional, the onus is really on the qualitative analyst to show why the PITF scores of partial democracy or factionalism are wrong. Indeed, a PITF score of partial democracy and factionalism should prompt a qualitative analysis that downplays elite divisions to look again and see why those Polity codes were given; there may in fact be greater elite divisions than had first seemed to be the case, and the qualitative analysis might be in error.

By similar logic, where the PITF codes for a country show high discrimination or conflict in neighboring states, a qualitative analysis that argues for stability needs to show

that such discrimination does not in fact exist or that conflicts in neighboring states are isolated and not likely to spill over. Otherwise, the qualitative analysis is probably the one in error.

Let us now go through a set of specific procedures to follow when the predictions of the qualitative analysis and the quantitative analysis differ.

The qualitative model predicts stability but the PITF model predicts instability

In this case, one should first use the PITF model to interrogate the qualitative analysis. If the PITF score is in the highest decile, where the PITF model has been shown to be extremely accurate, one should begin by accepting the PITF forecast as most likely accurate. The qualitative analyst would then have to show in convincing fashion sound reasons to think the data on which the PITF model score for that country is faulty.

However, if the PITF score is in the second or third decile (predicting a 30 to 60 percent incidence of instability), one should be more open to letting the results of the qualitative analysis indicate whether that country is more or less likely to have a near-term crisis. The qualitative model would predict stability if two or more of the following conditions hold: (1) state finances are adequate and secure; (2) elite unity and support for the regime are strong or at least opposition elites are divided and weak; and (3) popular discontent is muted or ineffective, either because the regime remains popular or because the discontented are dispersed, leaderless, and lack any organization or framework for mobilization.

Thus, the PITF results would force the qualitative analyst to show evidence that, even for a relatively poor country, finances are sound. This means that sources of revenue are stable and adequate to meet expenses, that no excessive debts are being run up, and that expenditures are under control. Corruption is a major potential problem in this regard, and qualitative analysis should make sure that corruption is not at levels that are undermining government capacity.

It is also crucial for the qualitative analysis to identify the factors that are either exacerbating or overcoming potential elite cleavages and determining the likely future degree of cooperation with the regime. Is it foreign support or patronage that is providing elite support for the regime? How stable is that support? If opposition elites are weak because they are divided, what could lead them to unite against the regime?

Finally, if the PITF results are driven by high levels of discrimination or of conflict in neighboring states, it is vital for the qualitative analysis to show whether or not these conditions are leading to mobilization of regime opponents. If not, are the obstacles to mobilization temporary or likely to last?

If the qualitative analysis shows that the regime is financially strong, has elite unity and support, and there is little potential for mass mobilization, then one can conclude that the country is likely to remain stable. It is then worthwhile to interrogate the PITF variables and ask if any of them seem to be mismeasured or changing. For example, is regime type correctly identified? Some regimes change within the two-year window of PITF prediction, becoming more or less democratic. Has factionalism been resolved by some agreement? Have discriminatory policies been changed? It is important to try to determine whether the PITF variables are mismeasured or if there are other factors revealed by the qualitative analysis that are countering the effects of poverty, factionalism, discrimination, etc. Once those factors are identified, the sustainability of those factors then becomes an important part of the analysis, alerting analysts to possible changes that could precipitate instability.

If the qualitative analysis shows that the regime is financially strong, has elite unity and support, and there is little potential for mass mobilization, then one can conclude that the country is likely to remain stable.

In recent years, a significant number of partial democracies have proven more stable than would have been expected by the PITF model. This seems most likely due to foreign interventions, which have suppressed the effects of factional competition and stabilized conflict situations. One can point to the role of EU negotiations in Turkey as leading to

measures that temporarily reduced conflict with the Kurds, although that has reappeared. In Côte d'Ivoire, Liberia, and other African countries, recent interventions have helped forge (temporary?) elite agreements to bridge factional divides.

However, a qualitative analysis that predicts stability when the PITF model predicts high or even moderate instability should always be questioned before it is accepted. All too often, qualitative analysis tends to project from the present, rather than find discontinuities that lurk beneath the surface. It was precisely because qualitative analysts tended to project continuity for regimes such as the shah's Iran and the Soviet Union that the PITF model was commissioned. A PITF prediction of instability may be in error, but it may also be saying that fundamental causes of instability are present, and that while they may be countered for a period of time by the government's access to certain resources, foreign intervention, or temporary agreements, they need to be considered as active threats should the countervailing factors be removed. Together, the qualitative and PITF analyses are likely to provide a more accurate and more thorough forecast than either one alone.

The qualitative model predicts instability but the PITF model predicts stability

This situation is most likely to arise in the analysis of dictatorships and full democracies with low incomes. In both cases, the PITF model tends to overpredict stability, mainly because the quantitative model is less sensitive to short-term events and changes that indicate shifting conditions below the surface of such regimes. For countries in the lowest deciles—which tend to be autocracies with low discrimination against specific groups and moderate to high incomes, or full democracies with moderate to high incomes—the PITF model is extremely accurate in its forecast of little or no chance of near-term violent conflicts.

Yet for countries in the lower-medium deciles (5 to 8) on the PITF model, analysis based on structural analogy can be extremely valuable in providing further insights into the likely risk of conflicts. For example, as analysts since de Tocqueville have noted, dictatorships are most at risk when they are engaging in reforms designed to "loosen" or "open up" the regimes. From the French Revolution following the meetings of the Estates General that had been called by the king to help him reform state finances to the collapse of the Soviet Union following elections that had been called by Gorbachev to help him reform the Communist Party, warning bells for instability should always ring when dictatorships change course. Unfortunately, such stirrings of reform do not appear in the PITF data in real time because it would take a full year or more of successful reforms to change the classification of a regime from "autocratic" to "partial autocracy," and some efforts at reform do not show up at all because they are not fully enacted, even though the efforts are an alert to elites and popular groups that change is possible.

Even regimes that have been partial democracies without factionalism or full democracies may be vulnerable to creeping corruption or efforts of elected leaders to monopolize power.

Conversely, even regimes that have been partial democracies without factionalism or full democracies may be vulnerable to creeping corruption or efforts of elected leaders to monopolize power. This too can lead to instability, but may not show up in the PITF data on regime types until a year or two after qualitative analysis already discerns cracks in the current regime.

In these cases, the qualitative analysis should be used to interrogate the PITF results. The qualitative analysis should ask if the regime identification by PITF is faulty, perhaps overtaken by recent events. The qualitative analysis should identify the basis for elite divisions and for elite actions against the state, even where the PITF has not coded for factionalism. Finally, the qualitative analysis should lay out how popular mobilization is developing, or could arise, even if the PITF model does not display high levels of discrimination or conflicts in neighboring countries.

If the qualitative analysis convincingly demonstrates that several of the conditions for impending instability are present—the state is in financial difficulty or embarking on reforms designed to shore up its power, elites are increasingly opposed to the current

ruler or rulers, and popular groups are increasingly organizing themselves for antiregime actions and building ties to elites in opposition to the regime—then it is most likely that the PITF prediction of stability is wrong and the country is at risk.

General Considerations

When a qualitative analysis and the PITF model disagree, it is important to study both sets of predictions, and not simply disregard one or the other, *precisely because both models have different weaknesses*. Qualitative analysis is likely to overpredict stability because qualitative analysts tend to focus only on the kinds of problems that have arisen previously in their countries. Moreover, they tend to project the present—in which the state may seem strong—into the future. In fact, states often fail because new problems spill in from beyond frontiers, long-standing discrimination suddenly becomes intolerable following a ruler's overstep (e.g., assassination or assaults on popular opponents), or a shift occurs in foreign support for a ruler. Thus, the PITF model may often be a better guide to problems that can lead to instability in the event of a slight shift of events.

When a qualitative analysis and the PITF model disagree, it is important to study both sets of predictions because both models have different weaknesses.

Moreover, the PITF model is the only forecasting model known to have a proven record of accuracy on global historical data of 80 percent or more. Most forecasting models have not been extensively tested on historical and out-of-sample data. The PITF model has been so tested, and its accuracy rate is established. However, we really do not know how accurate are the predictions of traditional analysis based on qualitative analysis and structural analogies.

The structural analogy model may be more readily attuned to the particularities of a specific country, but one can never be certain how accurate it is overall. Moreover, the qualitative model is only as good as the analyst is willing to be precise in making an estimate. Qualitative analysis by experts needs to be pushed to make a clear statement: "Yes or no, will country X have a crisis in the next 12 to 24 months?" Or perhaps a risk percentage: "We estimate there is a 75 percent likelihood that country X will have a crisis in the next 12 to 24 months." A forecast that simply says "we think the risks of country X having a crisis in the next year are small" needs to be recorded as a "no." A forecast that says "we think the risks of country X having a crisis in the next year are substantial" needs to be pushed further, if it is to be usefully combined with other models. It needs to be made more explicit and precise: Is "substantial" a 50 percent likelihood of crisis? Or is it closer to 80 percent? One needs to commit to a number before one can compare predictions of different models. All too often, this is not done. This allows analysts to perhaps overestimate the accuracy of their subjective judgments because they are not sufficiently specified to allow true post-hoc assessments.

The PITF model, however, is prone to slightly overpredict instability for intermediate regime types and to underpredict instability for full democracies and full autocracies with low incomes. That is because *on average*, full democracies and full autocracies are the most stable regime types. The PITF model is less able to incorporate those factors unique to individual states that cause full autocracies and full democracies to depart from the normal stability of such regime types. These factors, however, may be picked up by qualitative analysis. They include cases in which foreign intervention stabilizes a factional division and helps to overcome it or where changes in the regime or regime/elite relationships have occurred to which the PITF model is slow to respond.

Using both models together thus provides a far superior method of screening for instability. Where countries are identified by both approaches as unstable, or by both as stable, analysts can have a much higher confidence level in those assessments than would be obtained by using either model alone. And where countries are identified by only one approach as unstable, those countries should be subjected to more detailed analysis of both the qualitative and PITF data and results to determine which of the models is most

likely misrepresenting or overlooking the true underlying conditions that lead to stability or instability.

It is also important to combine qualitative and PITF models in helping determine the *type* of instability likely to arise in a given case. For example, the PITF model in 2002 predicted instability for Iran—the degree of factionalism between reformers and hard-line clerics in the context of the mixed electoral/clerical system, plus neighboring conflicts and economic stagnation, made instability highly likely. However, what occurred in 2004 was not the kind of instability that the Bush administration had hoped for, namely, a popular movement to overthrow the clerical leadership. Rather, the reverse occurred—a backslide from partial democracy to autocracy as the clerics clamped down on protest, effectively banned opposition, disqualified dozens of reform candidates from running for parliament, and managed elections that brought to power and solidified the hold of more extreme religious leaders. Only a qualitative analysis that pointed to the strong revenues of the government from oil; the willingness of most elites, even reformers, to abide by the rulings of the supreme leader and follow the Guardian Council; and the strength of the Revolutionary Guard versus possible popular protestors could have spelled out the likely direction of change with accuracy.

Similarly, the PITF model has not shown an ability to predict the magnitude of violent conflicts. Knowing whether a political crisis is likely is one thing. However, knowing how large is the minority group at risk, how extensively various groups are armed or mobilized, or how brutal the regime is likely to be in the event of protests—factors that help determine the scale of violence—are at this point far better assessed by qualitative analysis than existing quantitative forecasting models.

In sum, if I were to design the most effective forecasting system using the two tools with which I have personally been associated—the PITF model and a qualitative model of "state breakdown"—I would insist on using both. I would maintain a PITF watch list of the entire world, while performing ongoing qualitative analysis of all countries with strategic importance or in which the United States had significant vested interests. I would also insist on frequent cross-checking between the quantitative and qualitative analysis results. Where both identify countries as unstable, I would consider that a very high-confidence prediction. Where the PITF model predicted a country as unstable that did not appear in qualitative analysis, I would ask for further qualitative investigation to confirm or discount that prediction. However, given the high known accuracy of this quantitative model, and the uncertain accuracy of the qualitative model, I would ask for a very thorough and convincing causal analysis based on the qualitative model before I withdrew confidence in the PITF forecast.

Conversely, where a qualitative analysis predicted a country as unstable that was not on the PITF watch list, I would ask for the detailed PITF data for that country and interrogate the models to determine which was more likely giving the correct signal from its data. Having two different approaches to draw on would greatly enhance my confidence that I was getting the most accurate forecasts available, and overcoming the limitations of single-model analysis that could never be more than 80 to 85 percent accurate.

Finally, it should be noted that forecasting models are not the only kind of data that would be relevant to stability analysis. In cases where the United States is engaged in rebuilding a society that has collapsed or is seeking to shore up a state that is not in immediate danger of crisis but emerging from crisis and thus still fragile, another kind of indicator is needed: a current indicator that can be used to assess a state's progress in economic and political development. This is particularly important for states that are likely to remain stable for the next few years or even decade. This stability may be due to the power of an authoritarian regime, the buoyancy of a successful export economy, or the legitimacy acquired by recent replacement of an unpopular regime with a popular leader. However, these regimes nonetheless are storing up problems regarding corruption, factional or ethnic conflict, undiversified economies, or leaders who are moving to monopolize political power. Conversely, this type of gauge is also important for countries

that are currently unstable due to recent conflicts or adverse conditions but that are trying to lay the foundations for greater stability in the future. Professor Monty Marshall and I have developed such a scale, which we have called our "State Fragility Matrix."[11] It is not designed as a forecasting tool but rather as an assessment tool for gauging a country's long-term progress toward the economic and political foundations of sustainable stability. I mention this only to show that still other tools may also be useful in assessing countries, depending on the precise goals of the analysts.[12]

As a final word of caution, I would emphasize the recent work by Philip E. Tetlock on the accuracy—or more correctly, the inaccuracy—of expert judgment.[13] Tetlock surveyed numerous experts to obtain their predictions on a variety of social science trends and demonstrated that their accuracy was generally quite poor—often no better than 50 percent. These errors arose in part because experts tended to reflect the views of other experts, but this "herd consensus" almost always lagged behind events. They also arose in part because experts are much better at generalizing or extrapolating current trends or historical conditions than at spotting previously unimportant factors that are driving discontinuous change. Their forecasts about when changes or discontinuities would occur were especially weak. The U.S. government convened the PITF precisely because the accuracy of U.S. intelligence assessments regarding stability in Iran, the Philippines, and the USSR had been so poor. Interestingly, the experts also often failed because they were seduced by the knowledge of detail into thinking they knew more, even when these details were misleading. Thus, Tetlock and other researchers have found that experts often do no better—or worse—than simple formulas in a variety of predictions. One well-tested example is that admissions committees' forecasts of a student's grades in college that weigh personal attributes, letters of recommendations, and test scores and high-school grades are no more accurate than forecasts made from a formula based only on a weighted average of test scores and high-school grades.

I would also note that qualitative forecasting can rest on a variety of methods, not all of which are equally good. Simple expert surveys tend to be poor because of the problems noted by Tetlock. Forecasts relying on "privileged sources" are notorious for their bias. For example, the forecasts for postinvasion stability in Iraq that were based on the assurances of Iraqi exiles proved way off the mark. The best qualitative forecasts tend to be based on causal models drawn from structural analogies. Thus, the forecasts of instability for postinvasion Iraq based on analogies citing Iraq's ethnic and religious divisions and the historical obstacles to getting cooperation among Iraq's Shia, Kurdish, and Sunni elites proved to be superior.

The advent of quantitative forecasting models for political instability that have known and high rates of accuracy offer the prospect of major advances in accuracy of forecasting over the traditional qualitative methods. However, this does not mean that such models should entirely displace traditional qualitative analysis. Instead, the best results are likely to be obtained by the judicious combination of quantitative analysis based on forecasting models with qualitative analysis that rests on explicit causal relationships and makes unambiguous and precise forecasts of its own.

Ideally, separate teams or analysts should carry out analysis using the quantitative and qualitative forecasting models to assure the independence of the two analyses. Then the two teams or analysts should be brought together to compare their analyses in the manner noted above and compete to demonstrate the accuracy of their analysis where their predictions differ. In this way, political instability forecasting is likely to make its largest advance over earlier practices.

The best results are likely to be obtained by the judicious combination of quantitative analysis based on forecasting models with qualitative analysis that rests on explicit causal relationships.

An online edition of this and related reports can be found on our Web site (www.usip.org), together with additional information on the subject.

Notes

1. Sean O'Brien, "Analyzing Complex Threats for Operations and Readiness (ACTOR)" 2001, www.stormingmedia.us/07/0739/A073993.html (accessed February 15, 2008); Jack A. Goldstone et al., Political Instability Task Force Report, Phase IV Findings (McLean, VA: Science Applications International Corporation, 2003).

2. J. Scott Armstrong, "Evidence-based Forecasting," International Institute of Forecasters, Forecasting Principles, October 2006, www.forecastingprinciples.com/selection_tree.html (accessed February 15, 2008).

3. A typical model would have a form such as

$$R(T + t) = f[k_0 + \sum_{i=1}^{n} k_i(X_{iT})]$$

where $R(T + t)$ is the risk of a violent conflict t months or years in the future, and is a function of the values of a fixed number of variables $X_1, X_2, X_3, \ldots X_n$ observed at time T. Such models are developed by using theories of conflict to suggest candidate variables $X_1, \ldots X_n$, and then testing how well various combinations of variables, and various functional forms, serve to identify those countries that have a higher observed incidence of violent conflict at time $T+t$ from the data observed at time T.

4. Center for International Development, Polity IV Project, www.cidcm.umd.edu/polity/ (accessed February 15, 2008).

5. United Nations Statistics Division, Demographic and Social Statistics, unstats.un.org/unsd/demographic/sconcerns/mortality/mort2.htm (accessed February 15, 2008).

6. University of Maryland Center for International Development and Conflict Management, Minorities at Risk, www.cidcm.umd.edu/mar/data.asp (accessed February 15, 2008).

7. Center for Systemic Peace, Armed Conflict and Intervention Project, www.systemicpeace.org/aci.htm (accessed February 15, 2008).

8. Theda Skocpol, *States and Social Revolutions* (Cambridge: Cambridge University Press, 1979).

9. Graham Allison, *Essence of Decision* (Boston: Little, Brown, 1971).

10. Jack A. Goldstone, Ted Robert Gurr, and Farokh Moshiri, *Revolutions of the Late Twentieth Century,* (Boulder, CO: Westview, 1991); Goldstone et al., *Political Instability Task Force Report*.

11. Monty Marshall and Jack A. Goldstone, "State Fragility Matrix," *Foreign Policy Bulletin,* 2007.

12. For more on metrics for various purposes, including state reconstruction, see Military Operations Research Society workshop papers: "The Global War on Terrorism: Analytic Support, Tools, and Metrics of Assessment," November 30 – December 2, 2004, Naval War College, Newport, RI, www.mors.org/meetings/gwot_final.htm (accessed February 15, 2008).

13. Philip E. Tetlock, *Expert Political Judgment* (Princeton, NJ: Princeton University Press, 2005).

United States Institute of Peace
1200 17th Street NW
Washington, DC 20036
www.usip.org