FOUNDATIONS

workplace. Cam-

oral research. New

bration and coher-
cience 39:176–190.

ices: Understanding

11

Structuring Accountability Systems in Organizations: Key Trade-Offs and Critical Unknowns[1]

Philip E. Tetlock and Barbara A. Mellers

When things go wrong, one rarely needs to wait long to hear angry cries to "hold the rascals accountable." Intelligence agencies are no exception to this blame game, as we can see by surveying the past decade of recriminations over the failures to predict the 9/11 attacks and the lack of weapons of mass destruction in Iraq (Posner, 2005a). But "accountability" is no panacea, as we can see by surveying the sprawling experimental and field research literatures on the impact of various types of accountability on various forms of human judgment (Lerner and Tetlock, 1999). If "accountability cures" exist for what ails intelligence analysis, those cures will need to be far more complex and carefully calibrated than cries for "greater accountability" imply—and will need to be implemented in carefully controlled and phased field research trials to ensure that the desired effects outweigh the undesired.

The research literature on accountability spans work in social psychology, organization theory, political science, accounting, finance, and microeconomics (agency theory)—and offers us an initially confusing patchwork quilt of findings guaranteed to frustrate those looking for quick fixes. Sometimes "accountability" helps. Researchers have documented conditions under which accountability improves the calibration of probability estimates (Siegel-Jacobs and Yates, 1996; Tetlock and Kim, 1987), checks

[1]We are grateful to Cherie Chauvin, Jeffrey Cooper, Tom Fingar, Daniel Kahneman, Robert Jervis, John Morgan, Paul Tetlock, and Amy Zegart for helpful comments on earlier versions of this chapter. We especially thank Baruch Fischhoff for his assistance in sharpening our argument.

self-enhancement biases (Sedikides et al., 2002), makes people attend more seriously to hypothesis-disconfirming evidence (Kruglanski and Freund, 1983), and makes people more self-aware and accurate judges of covariation in their environment (Hagafors and Brehmer, 1983; Murphy, 1994). Other times, "accountability" has perverse effects, increasing defensiveness and escalating commitment to sunk costs (Simonson and Staw, 1992), increasing susceptibility to being distracted by low-diagnosticity cues (Tetlock and Boettger, 1989) and superficially plausible but specious reasoning (Barber et al., 2003), and inducing rather indiscriminant ambiguity aversion (Curley et al., 1986) and risk aversion (Tetlock and Boettger, 1994).

At other times, achieving consensus on whether "accountability" is helping, hurting, or having no effect is impossible because so much hinges on observers' sympathies and perceptions of whose ox is about to be gored. Examples of such accountability controversies include debates over how to structure relations between auditors and their corporate clients (Moore et al., 2006) or between legislators and citizens (Kono, 2006), how to make teachers accountable for improving student performance (Chubb and Moe, 1990), and how to hold managers accountable for making decisions on race-and-gender neutral grounds (Kalev et al., 2006; Tetlock and Mitchell, 2009). Setting up rules and incentives to encourage desired—and discourage undesired—behavior is a much-discussed problem that is far from fully solved (Kerr, 1975).

Understanding the effects of accountability requires clear definitions for both the independent and dependent variables, as well as a conceptual scheme for characterizing the myriad organizational processes for implementing accountability. On the independent-variable side, accountability is an omnibus term for a complex bundle of variables captured by the multipronged question: Who must answer to whom for what—under what normative ground rules and with what consequences for passing or failing performance standards (Schlenker, 1985; Scott and Lyman, 1968; Tetlock, 1985, 1992)? This definition could apply to virtually any level of analysis, from the societal to the interpersonal. One could hold governments accountable for policy miscalculations; governments could hold intelligence agencies accountable for flawed guidance; agency heads could hold their managers accountable for failure to check errors; and managers could hold individual analysts accountable for making the initial errors. To make this chapter manageable, we focus on the accountability pressures operating on individual analysts in their immediate working environment.

Even within this restricted focus, our definition still allows for enormous parametric complexity. One could hold analysts accountable to colleagues of the same or higher status whose own views are either well known or unknown before analysts submit their work, whose interpersonal style is more collaborative or adversarial, who are focused more on the process

by which analysts reach conclusions or on the bottom-line accuracy of those conclusions, who are known to be tolerant or intolerant of dissent, or who are known to be moderate or extreme in their reactions to success or failure. Each of these variations has the potential to influence either what analysts say they think (e.g., public attitude shifting) or how they actually think (e.g., private thought processes shift toward self-criticism or self-justification) (Tetlock, 1992). For these reasons, accountability is often viewed as a crucial construct for bridging more micro, behavioral science approaches (that focus on inside-the-head processes of decision makers) and social science approaches (that focus on the organizational and political structures within which individuals are embedded). Any reasonably complete characterization of the environment within which analysts work cannot ignore the accountability relationships that bind them to key constituencies and that can influence both what and how they think (March and Olsen, 1989).

On the dependent-variable side, policy debates over accountability systems invariably pivot on implicit or explicit conceptions of what would constitute enhanced performance. The official answer for intelligence agencies in the early 21st century—providing timely and accurate information that enables policy makers to advance our national interest—is too open-ended to be of much practical value. We need specific guidance on the types of "criterion variables" that proposed accountability systems are supposed to be maximizing or minimizing. Do we want analysts to focus on "policy-maker (customer) satisfaction," even if that means subtly signaling them to engage in sycophantic attitude shifting to support politicians' flawed world views (Prendergast, 1993; Tetlock, 1992)? Do we want to insulate analysts from political pressures, even if that means they become less responsive to policy makers' legitimate and often urgent needs? Do we want to incentivize only accuracy, with no regard for the inherent unpredictability of the environment, and risk rewarding analysts who play fast and loose with evidence, but just happen to be lucky on some recent big calls? Do we want to put the spotlight on measures of how analysts think—and reward the rigorous, punish the sloppy, and attach virtually no weight to "who gets what right?" How should we balance timeliness and accuracy as analysts move from cases that require rapid action to those that allow for more leisurely contemplation of ambiguities and trade-offs?

This chapter wrestles with the foundational questions: What insights can we glean from the vast interdisciplinary literature on accountability on how to balance the relevant trade-offs—and how the organizations dedicated to analyzing national security information should structure the work lives of those charged with offering useful guidance to decision makers? Furthermore, if one were to design—from scratch—the accountability

ground rules for monitoring the "performance" of intelligence analysts and incentivizing "improvement," what should one do?

Initially, the complexity of the choice set looks overwhelming. Accountability systems can be classified in so many ways—indeed, in principle, there are as many forms of accountability as there are distinct relationships among human beings. That said, two key issues loom especially large in political debates over how to structure responsibility for intelligence analysis processes and products (Posner, 2005a, b). The first is beyond the control of intelligence agencies—the degree of autonomy they should possess vis-à-vis their political overseers—and we touch on this briefly. The second is very much under the control of intelligence agencies—how they should structure their internal accountability systems for defining and facilitating excellence—and we devote most attention to this topic.

#1 Balancing Clashing Needs for Professional Autonomy and Political Responsiveness

How dependent should those who preside over intelligence analysis be on the approval or disapproval of their democratically elected political masters in Congress and the Executive Branch of government? Should they enjoy as much autonomy as, say, the Federal Reserve or should they—like most other political appointees—serve strictly at the pleasure of the President?

At one end of the continuum are those who insist on complete subordination of intelligence analysis to the political priorities of policy makers. Some might even argue that democratic political theory requires nothing else: analysts' overriding concern should be satisfying their "customers" in congressional committees and Executive Branch offices. But this argument may be too extreme in light of the substantial independence enjoyed by the Federal Reserve in analyzing economic trends and setting monetary policy, and in light of the evidence that the more independent central banks are, the better they manage complex trade-offs between unemployment and inflation (Alesina and Summers, 1993; Fischer, 1995). Complete subordination of intelligence analysis to political masters raises the risk of incentivizing analysts to be sycophants who fear telling their political bosses anything that could incur their wrath, thereby facilitating groupthink-like insulation from dissonant insights.

At the other end of the continuum are those who argue for the independence that the Federal Reserve exercises in analyzing monetary policy or that the National Science Foundation exercises in choosing which grants to fund. Some might insist that the power to punish intelligence agencies for offering ideologically unwelcome reports is incompatible with the goal of getting ruthlessly objective analysis. But this argument also may go too far. In an open and pluralistic society, politicians may often pay a steep penalty

for politicizing intelligence assessments—and consumers of intelligence need some leverage over the suppliers to ensure responsiveness to legitimate needs. Few would want analysts to be as free to ignore policy demands as they would be if they enjoyed the job security of tenured professors. Analysts who feel accountable only to each other in this scenario might start writing like academics: only for each other.

Whether the current system has found a sound compromise between these clashing values is beyond the purview of this chapter. But there should be little doubt about the need to factor these trade-offs into organizational design. Furthermore, there should be little doubt that where one comes down on this continuum of accountability design options will be influenced by one's assumptions about relative risks of excessively pushy politicians and of an excessively insulated analytical community that exploits lax oversight to pursue its own agenda.

#2 Balancing Clashing Needs for Rigor in Processes and for Creativity in Coping with Rapidly Changing Events

Whatever degree of institutional independence one prefers for intelligence agencies, how should we gauge the performance of analysts working in these agencies? Should we embrace pure process accountability and focus solely on the rigor of the underlying analytic process? Or should we embrace outcome accountability and focus solely on who gets what right across issue domains and time frames (pure outcome accountability)? Or should we embrace some form of hybrid process-outcome system that assigns adjustable weights to rigor and accuracy and that, depending on the task requirements of the moment, allows for the possibility of letting process trump outcome in some settings, of letting outcome trump process in other settings, and of assigning equal weights in yet other settings?

Advocates of process accountability maintain that the best way to reach the optimal forecasting frontier—and stay there—is to hold analysts responsible for respecting certain logical and empirical guidelines. As we can see from the list of standards for analytic tradecraft contained in Intelligence Community Directive No. 203, June 21, 2007 (Director of National Intelligence, 2007, pp. 3-4), the current emphasis appears to be on process accountability within the Office of the Director of National Intelligence:

- Properly describes quality and reliability of underlying sources (how do you know what you claim to know?);
- Properly caveats and expresses uncertainties or confidence in analytic judgments;
- Properly distinguishes between underlying intelligence and analysts' assumptions and judgments;

- Incorporates alternative analysis where appropriate;
- Demonstrates relevance to U.S. national security;
- Uses logical argumentation;
- Exhibits consistency of analysis over time, or highlights changes and explains rationale; and
- Makes accurate judgments and assessments—although this comes with the understandable caveat: make the most accurate judgments and assessments possible given the information available . . . and known information gaps. . . . Accuracy is sometimes difficult to establish and can only be evaluated retrospectively if necessary information is collected and available.

Indeed, strong arguments can be made for stressing process over outcome accountability. Proponents of process accountability warn that it is unfair and demoralizing to hold analysts responsible for outcomes palpably outside their control—and doing so may stimulate either risk-averse consensus forecasts (herding of the sort documented among managers of mutual funds—whereby individuals believe, they can't fire all of us—Bikhchandani et al., 1998; Scharfstein and Stein, 1990) or, when retreat is impossible, escalating commitment to initially off-base forecasts (defending those positions as fundamentally right but just off on timing—Simonson and Staw, 1992; Tetlock, 2005).

A large body of work shows how easily outcome-accountability systems can be corrupted—even in competitive private-sector firms (Bertrand and Mullainathan, 2001; Sappington, 1991). For example, when H. J. Heinz division managers received bonuses only if earnings increased from the prior year, managers found ways to deliver consistent earnings growth by manipulating the timing of shipments to customers and by prepaying for services not yet received, both at a cost to the firm as a whole (Baker et al., 1994). One can easily imagine analogs in which managers of analysts find ways of inflating their "accuracy scores" by increasing the number of "easy things" they are responsible for predicting or by introducing more generous scoring rules that permit reclassifying errors as "almost right" or "just off on timing" or derailed by inherently unpredictable exogenous shocks. Furthermore, one also can easily imagine how outcome-accountability pressures operating on individual analysts could lead to information hoarding and even sabotage of each other's efforts (although this classic problem can be mitigated by emphasizing outcome accountability for team performance—and by designing disincentives for hoarding).

Process-accountability proponents also warn of how impractical—as well as unfair—outcome-accountability systems can be. They are certainly right that assessing the accuracy of real-world political forecasts is a nontrivial undertaking. Below we list seven often-offered objections to the

feasibility of factoring accuracy metrics into the standards to which analysts are held accountable:

1. Self-negating prophecies in which initially sound predictions (that would have been right if policy makers had not acted on them) now appear incorrect because policy makers did act on them (one still much-debated example: Y2K).
2. Self-fulfilling prophecies in which initially unsound predictions (that would have been wrong if policy makers had not acted on them) become correct because policy makers did act on them (one possible example: aggressive counterinsurgency measures against a subpopulation that was not pro-Taliban beforehand, but becomes pro-Taliban as a result of the measures).
3. Exogenous shocks or missing information on key variables that cause lower probability outcomes to occur—and cast into false doubt fundamentally sound analyses of causal dynamics.
4. Exogenous shocks that cause credit to be assigned to far-fetched theories.
5. Arbitrary time frames for assessing the accuracy of many predictions (should we call a Sovietologist wrong if he or she thought in 1988 that the Union of Soviet Socialist Republics would disintegrate within a 10-year frame, but not the 5-year frame—or should we praise him for being so much closer to right than the expert consensus in 1988?).
6. When forecasts are premised on conditional adoption of policy x, how can we know whether the forecasts were right when non-x was adopted.
7. The "I-made-the-right-mistake" defense (the consequences of making a Type 2 error of underestimating the enemy are vastly greater than those of making a Type 1 error of overestimating the enemy). Inflating the probabilities of the more serious error is reasonable if that it is the only way to achieve the desired policy outcome.

Proponents of outcome accountability reply that, although accuracy is indeed an elusive construct, there are ways to address these objections—and that rough measures with known limitations are vastly better than no measures (Tetlock, 2005). Properly implemented, outcome accountability empowers people to seek ingenious analytic strategies not formally embodied in process guidelines (Wilson, 1989). Proponents of outcome accountability also worry that: (1) process accountability can readily ossify into bureaucratic rituals and mutual backscratching—Potemkin-village facades of process accountability and rigor designed to deflect annoying questions from external critics (Edelman, 1992; Meyer and Rowan, 1977); and (2)

process accountability can distract analysts from the central task of understanding the external world by squandering cognitive resources on impression management aimed at convincing superiors of how rigorous their analytical processes are (Lazear, 1989). A healthy dose of outcome accountability alerts us to the possibility that even the best on-paper, process-accountability mechanisms can be corrupted in a multitude of ways—whether they are opinionated managers who suppress information or peer reviewers who fail to catch errors in draft reports because the reviewers are too homogeneous in outlook or because they have been intimidated by dogmatists higher up in the bureaucratic food chain. Outcome accountability sends a much-needed signal to beleaguered dissenters that, although they may suffer the slings and arrows of short-term career damage by taking unpopular positions (e.g., being labeled as "process deficient" by groupthink mindguards [Janis, 1972]), formal mechanisms will be in place to compensate them for the losses—and then some. The logic here is akin to that in whistleblower protection legislation—namely, to encourage the sort of dissent that is often suppressed even in relatively well-functioning organizational systems. Inasmuch as process-accountability systems can fail in a host of unintended ways—many of which can be offset by carefully calibrated doses of outcome accountability—it becomes harder to defend a categorical rejection of all efforts to explore the potential value-added of outcome accountability.

Again, whether the current system has found a sound compromise between competing accountability design templates is beyond the purview of this chapter. But again there should be little doubt about the need to factor these trade-offs into organizational design, and little doubt that where one comes down on this continuum of accountability design options will be influenced by one's implicit or explicit assumptions about the relative risks of process accountability being corrupted and degenerating into a bureaucratic formality versus the relative risks of holding people unfairly accountable for the inherently unforeseeable, and prompting them to engage in ever more elaborate forms of trickery designed to inflate their accuracy scores.

Of course, the choice between process and outcome accountability is not dichotomous. One can easily imagine an enormous range of blends of process and outcome accountability, many of which can be found in the private sector. One well-known process-outcome hybrid in the world of finance is RAROC (risk-adjusted return on capital) guidelines, which place constraints on the risks that decision makers are allowed to take with firm money, but still incentivize decision makers to maximize returns within those guidelines.[2] Within a RAROC world, one can profit handsomely

[2]RAROC was developed and first implemented by Charles Sanford at Banker's Trust Company in the mid-1970s.

from being right if one works within the firm's process guidelines on acceptable risk taking—but one can also lose one's job if one makes "too much money" by violating those guidelines and exposing the firm to unacceptable potential risk.

In judging the wisdom of infusing more outcome accountability into intelligence analysis, much hinges on one's answers to the following question: Are current systems already functioning so close to the optimal forecasting frontier that additional outcome accountability is unlikely to improve aggregate performance—and likely instead simply to shift error-tolerance thresholds?

THE INFORMATION ENVIRONMENT

The performance of any analytical system, and the contribution of any attempt to improve it, depends on the difficulty of the task. That depends, in turn, on the difficulty of extracting a signal from the world being analyzed and the error tolerances of those who depend on the system for useful insights.

Figure 11-1 formalizes these demands, drawing on signal detection theory, a mainstay of behavioral science (Green and Swets, 1966; as well as this volume's McClelland, Chapter 4, and Arkes and Kajdasz,

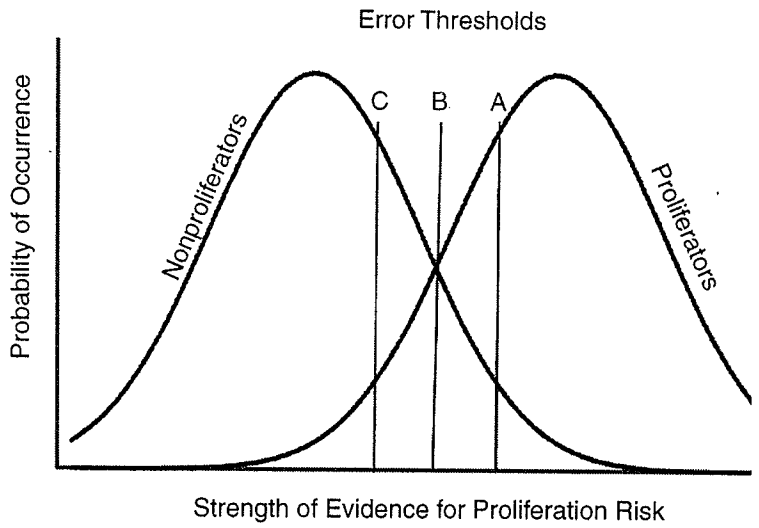

FIGURE 11-1 A world that permits modest predictability.
SOURCE: Generalized from Green and Swets (1966).

Chapter 7). In Figure 11-1, the analytical task is to assess the risk of nuclear proliferation, in a world where the two distributions represent the behavior of nations that truly are and are not proliferators. Imagine that each observation is drawn at random from the appropriate distribution. In most cases, analysts can tell whether an observation comes from a proliferator or a nonproliferator. However, in some cases in the middle range, the evidence is either hard to assign to one kind of country or even misleading (such that proliferators look like nonproliferators and vice versa). Analysts cannot avoid making errors when the evidence falls in the overlapping zone, no matter how strong or what types of accountability pressures are on them.

Which errors analysts make will, however, depend on how they interpret those accountability pressures. Imagine three political forecasters who are equally skilled at discriminating proliferation signals from noise, but differ in the error thresholds they believe their task requires. These thresholds are indicated by vertical lines A, B, and C. They represent the balance of evidence that each forecaster uses to distinguish "nonproliferators" and "proliferators." Forecaster C will tolerate many false alarms to avoid one miss, and so is more likely to overconnect the dots. An extreme example of this value orientation was the position allegedly taken by Vice President Dick Cheney with respect to the invasion of Iraq in 2003: Even a 1 percent probability of a nuclear weapon falling into the hands of terrorists was "too much" (Suskind, 2006). Forecaster B assigns equal weight to the two types of errors. Forecaster A will tolerate many misses to avoid a single false alarm, so will not make the proliferation "call" on a nation that sends only weak indicators, either because intelligence gathering is weak or the target nation is skilled at deception.

Figure 11-2 displays the same dilemma in a different way, plotting hit rates against false alarm rates. Perfect forecasters achieved a 100 percent hit rate at 0 percent cost in false alarms, falling at the point in the top left corner. Forecasters with no ability, who simply guessed, would fall along the main diagonal, at a point reflecting their understanding of the system's tolerance for hits versus false alarms. For example, forecasters with a strong aversion to misses would consider nearly every observation to be evidence of proliferation and end up near the upper right corner (in Figure 11-2). The three forecasters from Figure 11-1 appear on the same curve, representing their (identical) ability to extract signals about proliferation, but their differing understanding of the appropriate error thresholds. Note that A and C, who see strong aversion to false alarms or misses respectively, produce forecasts not that different from forecasters who just guess. Of course, that divergence could be vitally important.

The area between the curve and the diagonal in Figure 11-2 represents the value that forecasters add, above chance. To move to a higher

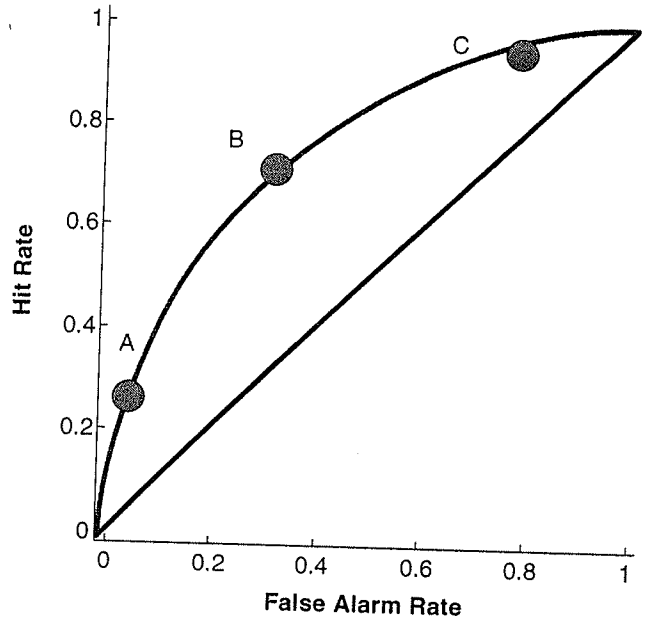

FIGURE 11-2 Possible hit-versus-false-alarm trade-offs in a world of modest predictability (see Figure 11-1).
SOURCE: Generalized from Green and Swets (1966).

ability curve, forecasters need to be better analysts or to have better information—or live in a world where the two kinds of nations behave more distinctly. Unless an accountability system changes forecasters' analytical ability—how they process information or the types of information they process—all it can do is change the relative mix of errors. Getting analysts' thresholds attuned to those that the organization desires can have value. However, that is a different enterprise from attempting to improve their aggregate forecasting performance. Indeed, if the threshold shifts are seen as serving political fashions, they may even discourage thoughtful analysis (why bother thinking when the best way to get ahead is simple conformity?).

Figure 11-3 depicts a more highly predictable world, with less overlap between distributions, in which analysts can more readily distinguish proliferators from non-proliferators. However, because the thresholds remain the same, in absolute terms, they reflect much different (and much higher) levels of knowledge. The solid curve in Figure 11-4 shows that the corresponding forecasting frontier has been pushed toward the top left corner. As a result, analysts can now achieve the same hit rate with a much lower

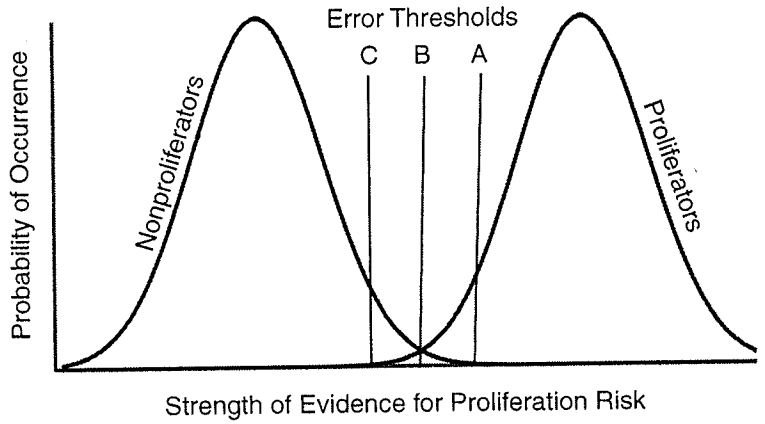

FIGURE 11-3 A world that permits a very high degree of predictability.
SOURCE: Generalized from Green and Swets (1966).

false alarm rate, or the same false alarm rate with a much higher hit rate. However, exploiting this new forecasting ability requires understanding the quality of the judgments that it allows. Unfortunately, although people are more confident when they are more knowledgeable, the correlation is weaker than it should be.

Observers tend to be overconfident in domains where they know little—and underconfident in domains where they know a lot (Erev et al., 1994; Lichtenstein and Fischhoff, 1977; Moore and Healy, 2008). As a result, analysts and their clients might not take full advantage of improved ability unless they knew how good it was. This is likely to happen only if the analytical organization is committed to evaluating its accuracy systematically. That requires comparing analyses with actual events, while maintaining a consistent threshold for calling events (in this case, whether nations are proliferators or not). Maintaining that threshold requires a clear and effectively communicated organizational philosophy, implemented with appropriate incentives, the topic of the next section.

Does anyone, however, know how close we are to the optimal forecasting frontier? How can we determine whether we are living in a Figure 11-1 or 11-3 world? Here awareness of and candor about ignorance are essential: No one has a strong scientific claim to know. Assessing where the prediction ceiling might be in political–military–economic domains of highest priority to the intelligence community is an inherently open-ended assignment (short of the absolute and extremely improbable ceiling defined by R-squared values of 1.0)—and the peer-reviewed literature, notwithstanding some heroic efforts (Armstrong, 2005; Bueno de Mesquita, 2009),

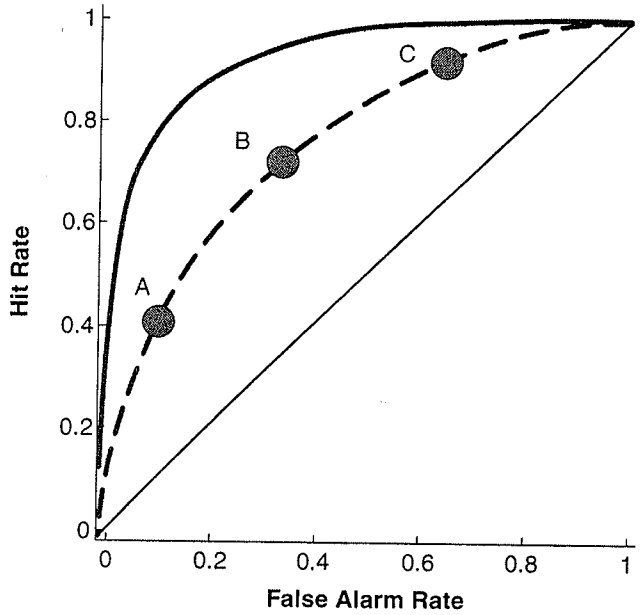

FIGURE 11-4 Possible hit-versus-false-alarm trade-offs in a world with a very high degree of predictability (see Figure 11-3).
SOURCE: Generalized from Green and Swets (1966).

has barely scratched the surface of this massive undertaking. This means that, although many may have opinions about the wisdom of exploring the feasibility and desirability of closer scrutiny of the accuracy of analytical judgment, no one knows how much—or little—we stand to learn from such studies. This sort of ignorance can be extraordinarily expensive even if we assume the possibility of only modest increments in performance—a reasonable assumption in light of evidence that just a single round of feedback can substantially improve calibration (in the long run, modest increments could save lives and money on a massive scale for military and economic decisions that hinge on accurate subjective-probability estimates [Lichtenstein and Fischhoff, 1980]).

This mix of deep ignorance with high stakes makes a strong case for conducting low-cost studies designed to explore the likely yield from developing sophisticated accuracy metrics and then institutionalizing level playing-field competitions that pit different analytical mindsets/methods against each other repeatedly across domains and time. Setting the scope for such competitions is beyond that of this chapter, but to ensure reasonably comprehensive coverage, these competitions should vary along at least five

dimensions, each with the potential to change the rank ordering in predictive power of mindsets/methods:

1. **Skewed versus evenly balanced base rates.** Events can range from the extremely rare (genuine black swans) to the quite routine (things that happen between 30 and 70 percent of the time). Being right is easy when predicting events with extremely skewed base rates, such as whether Phoenix, Arizona, will have rain on a summer day: Just predict no rain all of the time.
2. **Stable versus unstable environments.** Incrementalist analytical approaches, which update their predictions in light of experience, should perform best in environments with well-defined, slowly changing base rates. Such approaches fail dramatically, however, in unstable environments when "base rates," whether defined in cross-sectional or longitudinal terms, lose meaning: In 1991, what was the base-rate probability of a multiethnic empire, such as the Union of Soviet Socialist Republics, disintegrating? In 2001, what was the probability of a fundamentalist Islamic terrorist group pulling off an unprecedented mass-casualty attack on American soil? Available evidence suggests that expert judgment and statistical models alike do a poor job at identifying punctuated-equilibrium points, at transitions between stable and unstable environments—and then back again (Armstrong, 2005; Taleb, 2007; Tetlock, 2005).
3. **Relative severity of asymmetry of Type 1 versus Type 2 errors.** Rare events—such as the 9/11 strikes or a terrorism-sponsoring state acquiring nuclear capabilities—are often those for which we have great (although not infinite) tolerance for false alarms—and little patience for misses, even if accompanied by the excuse—"well, 99 percent of our reports were right." If proficiency can be acquired in predicting such rare events, such proficiency quite possibly has been purchased at the expense of accuracy in analyzing more mundane events. To check this possibility—and the acceptability of the price—researchers need to conduct comparisons of forecasting accuracy for both low and high base-rate outcomes, and managers need to communicate to analysts the value-weighted accuracy functions that they want analysts to maximize (e.g., I am willing to tolerate dozens of false alarms to avoid a single miss for these rare events, but I attach equal importance to avoiding false alarms and misses for these more common events).
4. **Temporal distance.** Although imagining exceptions is possible—such as processes with lots of short-term volatility that settle down over time—analyses should be more accurate for events closer in time. Indeed, as Kahneman and Klein (2009) suggest, beyond a certain point, long-range political forecasting may become impossible because of the potential for trivially small causes to have enormous and inherently unforeseeable effects. (They pose the following counterfactual thought experiment: How much different would 20th century history be if the three fertilized eggs—Stalin, Hitler, and Mao—had been female rather than male?).
5. **Different levels of analysis.** Performance may differ for micro and macro levels of analysis. Analysts are unlikely to be equally proficient in assessing the chances, say, of "the leadership of country X will make decision Y," "the European Commission passing resolution Z," "the voters of country Y electing candidate A," "financial markets dropping by XX percent," and "the alliance between countries C and D holding firm."

Such distinctions matter. Each of these factors may affect analysts' performance—as well as the power of accountability norms and incentives to improve performance. There is no guarantee that those methods or mindsets that have an edge in predicting lower base-rate, longer run shifts in macro processes, such as nation-state disintegration and the rise or fall of religious fundamentalism, will be the same as those that have an edge in predicting short-term tactical shifts in the behavior of individual leaders, such as shifts in trade negotiation stances. Indeed, considerable evidence shows there will be no all-round winner (Armstrong, 2005)—and it is prudent to think of prediction competitions as complex, Olympics-like tournaments, with numerous qualitative and quantitative subdivisions of events.

THE EVALUATIVE STANDARD

The signal detection theory framework, embodied in Figures 11-1 through 11-4, assumes that what matters, when holding analysts accountable, is their ability to reduce the risks of false negatives and false positives, weighted by the costs of each type of error. Such a long-term, large-sample perspective protects analysts who have had bad luck on an issue—and protects policy makers from analysts who get lucky. However, analysts' working conditions can impose other incentives, such as "pleasing one's immediate boss" or "policy maker (customer) satisfaction," which can translate into sycophantic attitudes shifting toward managers' and/or politicians' flawed world views (Prendergast, 1993). Conversely, efforts to insulate analysts from political pressures may make them less responsive to policy makers' legitimate and often urgent needs—in which case the value of their knowledge may be lost. Moreover, accuracy alone

may have relatively little value unless the information arrived in a timely, comprehensible fashion, so that its full meaning can be extracted.

Agency theory (Baker et al., 1994; Gibbons, 1998; Sappington, 1991) provides an account of how to align the goals of principals and agents linked by contracts and organizational ties. Here, the principals are the policy maker clients and the agents are the analysts, respectively. We start with the following four simplifying assumptions that treat principals and agents as rational egoists:

1. Assume a linear production function: $y = a + e$, where y is all intelligence products that the principal (the U.S. government) values; a is the effort that agents (analysts) must expend to produce y (and which is under their control); and e is noise that causes production to rise or fall, in unpredictable ways, outside analysts' control. Assume that outcome accuracy is the analytical product that policy makers value most.
2. Assume a linear wage contract: $w = s + by$, where w is wages, s is a fixed salary, agreed upon before knowing how productive an agent will be, and b is the bonus rate (at which the bonus rises per unit increase in y). The slope, b, is zero for process-accountability systems in which analysts' wages are all salary, with no bonus for predictive accuracy.
3. Assume a linear pay-off function for agents: $w - c(a)$, the realized wage minus the disutility of doing the work. Analysts' intrinsic motivation makes the utility term, $-c(a)$, less negative.
4. Assume a linear pay-off function for the principal, $y - w$, or the realized output net of wages.

According to agency theory, the more random the environment, the more workers will prefer the security of wage compensation, which imposes no risk on the agent $(b = 0)$. Given the randomness that intelligence analysts see in their world, they should adopt a risk-averse position and prefer a guaranteed salary that comes with process compliance over the uncertain bonuses that come with big prediction successes. One might generalize the empirical claim by positing that the larger the value of e, the more employees will be willing to trade increments in b for higher salary guarantees. The preferred compensation scheme from the government's perspective, as e increases, however, is less clear. On the one hand, the government wants to reward its workforce for skill, not luck. On the other hand, as e rises, it might become increasingly attractive to transfer responsibility for mistakes to intelligence agencies, then down the chain of command to analysts.

Applying agency theory runs into the same problem as applying accountability schemes. There is no unbiased measure of y, the output of

good analysis. The same problem arises in other domains, where employers struggle with finding performance goals and aligning accountability norms and incentives with them, even when the goals are well defined (Kerr, 1975). In principal–agent theory, this is called the "hidden action problem" of how to induce the agent to take a "correct" action that the principal wants, but cannot directly observe (Holmström and Milgrom, 1991). The technical solution is easy to prescribe, but hard to follow: Evaluate agents with metrics that most closely correlate with the observed but desired action. In the intelligence context, achieving this solution requires identifying process or outcome metrics that correlate with rigorous, open-minded analyst behavior that maximizes the chances of political assessments that are useful to analysts' clients.

Unfortunately, agency theory does not offer off-the-shelf solutions. Process metrics are closer to the behavior that employers hope to shape, but their application can rely on potentially faulty supervisor judgments (see this volume's Hastie, Chapter 8, and Kozlowski, Chapter 12), including cases where supervisors' preconceived theories of the outcome affect their judgments of the process. Process metrics can also lead to mechanical adherence to process over substance. Outcome-accountability metrics allow analysts the freedom to find the best ways to work through their problems, keeping them focused on the ultimate goal of their labors. They allow, even require, formal recognition of past difficulty and reporting thresholds, as conceptualized in signal detection theory. Of course, as the fine print in promotions for financial products reminds us, past performance is no guarantee of future performance. Moreover, cross-context consistency in the accuracy of political forecasts has been found to be low, although significantly above zero (e.g., Tetlock, 2005).

In brief, agency theory cannot conclusively answer the process–outcome question. But it does suggest the value of experimenting with process–outcome hybrids, and it does identify the institutional incentives that must be considered when conducting assessments of analysts' performance. Those incentives are expressed in organizations' formal rules, as in incentive schemes and in supervisors' rating procedures. If properly set, they express the signal-detection-theory formalisms in practical terms, recognizing both the limits to analysis and its goals. Agency theory provides guidance on how this can be done, as well as cautionary tales on how it can be corrupted, by those hoping to dodge accountability or to skew analyses.

CLOSING THOUGHTS

Accountability norms and incentives are essential to coping with the core challenges that all organizations confront: overcoming the constraints

of bounded rationality (Kahneman, 2003; March and Simon, 1993) and parochial interests (Pfeffer and Sutton, 2006), and enabling more effective use of information and expertise than would have been possible if we had relied on randomly selected individual or small-group components of the organization. We can easily see why so many reach reflexively for accountability solutions when things go wrong. The "woulda-coulda-shoulda" counterfactuals are too tempting: "surely we could have avoided that stupid error if we had just tweaked these accountability guidelines in this direction and those other guidelines in this other direction."

But such reflexive fixes are as likely to make things worse as they are to make things better. As should now be obvious, we cannot deduce from first principles an optimal accountability system for monitoring intelligence analysts and incentivizing them to add "more value"—even if we possessed clear-cut consensus metrics of "value." But we can take the following three constructive steps, as described in the paragraphs below.

Step #1: Be More Explicit About Strengths and Weaknesses of Competing Models

First, we can be more explicit about the strengths and weaknesses of competing models for how to organize accountability systems, such as process versus outcome versus process-outcome hybrid forms of accountability. Here it is also worth keeping in mind the larger context of this debate, which recurs across diverse policy domains. Observers' preferences for process versus outcome accountability tend to be correlated with how much or how little observers trust the organization and the human beings staffing it. The greater the distrust, the greater the likelihood that observers will worry about how readily process-accountability systems can be corrupted or diluted and thus favor "more loophole-resistant" outcome-accountability systems (Tetlock, 2000). For instance, critics of corporate America tend not to trust corporate personnel managers to implement affirmative action programs rigorously and tend to suspect that companies' process-accountability systems for ensuring equal employment opportunity are mere Potemkin village facades of compliance. They demand numerical goals and statistical-outcome monitoring of treatment of minority groups (Tetlock and Vieider, 2011). Conversely, critics of public schools tend not to trust public school administrators and teachers to run a rigorous curriculum, and suspect that process-accountability systems for monitoring school performance are mere Potemkin village facades. They demand objective outcome testing data (Tetlock and Vieider, 2011). Organizations can sense when they are not trusted—and it is crucial to avoid the emergence of perceived correlations between recommendations of process accountability and "we trust you" and recommendations of outcome accountability and

"we do not trust you." The intended message is: We just want to find out what works best.

Step #2: Be More Open About Our Knowledge of Optimal Forecasting

Second, we can be more open about our limited knowledge of where the optimal forecasting frontier lies for various categories of intelligence problems—and of what value we believe could be added by commitment to developing better accuracy measures and to conducting experiments on the power of various interventions to move analytic performance closer to the forecasting frontier. Here again, the context to the debate is larger. Some observers explicitly argue that much talk about intelligence reform is ill conceived and rests on occasionally ridiculous and unrealistic expectations about what reform can deliver. In this view, the major political challenge is not improving analyst performance; it is reducing expectations for that performance (Betts, 2009). Organizations do, of course, often welcome lowering of performance expectations—and harm can be done by demanding the impossible—but one would not be doing intelligence agencies a long-term favor by incorrectly concluding there is no room for improvement if subsequent events reveal improvement to have been possible. Imagine that 10 years from now, various private-sector and prediction-market initiatives start reliably outperforming intelligence agencies in certain domains—and, during his or her daily intelligence briefing, the President turns to the Director of National Intelligence and says: "I can get a clearer sense of the odds of this policy working by averaging public sources of probability estimates."

Step #3: Preempt Politicization and Clarify Arguments

Third, we can do a better job of preempting politicization and clarifying where the factual-scientific arguments over enhancing intelligence analysis should end and the value-driven political ones should begin. Once the scientific community has enumerated the organizational design trade-offs and key uncertainties, and has compared the opportunity costs of inaction with the tangible costs of the needed research, policy makers must set value priorities, asking: Do the net potential benefits of undertaking the embedded-organizational experiments and validity research sketched here outweigh the net observed benefits of not rocking the bureaucratic boat and continuing to insulate current policies and procedures from scientific scrutiny and challenge?

REFERENCES

Alesina, A., and L. H. Summers. 1993. Central bank independence and macroeconomic performance: Some comparative evidence. *Journal of Money, Credit, and Banking* 25(2):151–162.

Armstrong, J. S. 2005. *Principles of forecasting*. Boston, MA: Kluwer Academic.

Baker, G., R. Gibbons, and K. J. Murphy. 1994. Subjective performance measures in optimal incentives contracts. *Quarterly Journal of Economics* 109(4):1,125–1,156.

Barber, B. M., C. Heath, and T. Odean. 2003. Good reasons sell: Reason-based choice among group and individual investors in the stock market. *Management Science* 49(12):1,636–1,652.

Bertrand, M., and S. Mullainathan. 2001. Are CEOs rewarded for luck? The ones without principals are. *Quarterly Journal of Economics* 116(3):901–932.

Betts, R. 2009. *Enemies of intelligence: Knowledge and power in American national security*. New York: Columbia University Press.

Bikhchandani, S., D. Hirshleifer, and I. Welch. 1998. Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of Economic Perspectives* 12(3):151–170.

Bueno de Mesquita, B. 2009. *The predictioneer's game: Using the logic of brazen self-interest to see and shape the future*. New York: Random House.

Chubb, J., and T. Moe. 1990. *Politics, markets and schools*. Washington, DC: Brookings Institution Press.

Curley, S. P., J. F. Yates, and R. A. Abrams. 1986. Psychological sources of ambiguity avoidance. *Organizational Behavior and Human Decision Processes* 38:230–256.

Director of National Intelligence. 2007. Intelligence Community Directive (ICD) 203: Analytic standards. June 21. Available: http://www.dni.gov/electronic_reading_room/ICD_203.pdf [accessed May 2010].

Edelman, L. B. 1992. Legal ambiguity and symbolic structures: Organizational mediation of civil rights law. *American Journal of Sociology* 97:1531–1576.

Erev, I., T. Wallsten, and D. Budescu. 1994. Simultaneous over-and-under confidence: The role of error in judgment processes. *Psychological Review* 10:519–527.

Fischer, S. 1995. Central-Bank independence revisited. *American Economic Review* 85(2): 201–206.

Gibbons, R. 1998. Incentives in organizations. *Journal of Economic Perspectives* 12(4): 115–132.

Green, D. M., and J. Swets. 1966. *Signal detection theory and psychophysics*. New York: Wiley.

Hagafors, R., and B. Bremer. 1983. Does having to justify one's judgments change the nature of the judgment process? *Organizational Behavior and Human Performance* 31(2):223–232.

Holmström, B., and P. Milgrom. 1991. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7:24–52.

Janis, I. L. 1972. *Victims of groupthink: A psychological study of foreign-policy decisions and fiascos*. Boston, MA: Houghton Mifflin.

Kahneman, D. 2003. A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist* 58:697–720.

Kahneman, D., and G. Klein. 2009. Conditions for intuitive expertise: A failure to disagree. *American Psychologist* 64(6):515–526.

Kalev, A., F. Dobbin, and E. Kelly. 2006. Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity practices. *American Sociological Review* 71:589–617.

Kerr, S. 1975. On the folly of rewarding A, while hoping for B. *Academy of Management Journal* 18(4):769–783. Republished in 1995 in *Academy of Management Executive* 9(1):7–14.

Kono, D. Y. 2006. Optimal obfuscation: Democracy and trade policy transparency. *American Political Science Review* 100(3):369–384.

Kruglanski, A. W., and T. Freund. 1983. The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology* 19(5):448–468.

Lazear, E. P. 1989. Pay equality and industrial politics. *Journal of Political Economy* 97(3):561–580.

Lerner, J., and P. E. Tetlock. 1999. Accounting for the effects of accountability. *Psychological Bulletin* 125:255–275.

Lichtenstein, S., and B. Fischhoff. 1977. Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance* 20:159–183.

Lichtenstein, S., and B. Fischhoff. 1980. Training for calibration. *Organizational Behavior and Human Performance* 26:149–171.

March, J. G., and J. P. Olsen. 1989. *Rediscovering institutions: The organizational basis of politics*. Stanford, CA: Stanford University Press.

March, J. G., and H. A. Simon. 1993. *Organizations*. Cambridge, MA: Blackwell.

Meyer, J. W., and B. Rowan. 1977. Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology* 83:340–363.

Moore, D. A., and P. Healy. 2008. The trouble with overconfidence. *Psychological Review* 11:502–517.

Moore, D., P. E. Tetlock, L. Tanlu, and M. Bazerman. 2006. Conflicts of interest and the case of auditor independence: Moral seduction and strategic issue cycling. *Academy of Management Review* 31:10–29.

Murphy, R. 1994. The effects of task characteristics on covariation assessment: The impact of accountability and judgment frame. *Organizational Behavior and Human Decision Processes* 60(1):139–155.

Pfeffer, J., and R. Sutton. 2006. *Hard facts, dangerous half-truths and total nonsense*. Cambridge, MA: Harvard Business School Press.

Posner, R. A. 2005a. *Preventing surprise attacks: Intelligence reform in the wake of 9/11*. Lanham, MD: Rowman and Littlefield.

Posner, R. A. 2005b. *Remaking domestic intelligence*. Stanford, CA: Hoover Institution Press.

Prendergast, C. 1993. A theory of "yes men." *American Economic Review* 83(4):757–770.

Sappington, D. E. M. 1991. Incentives in principal–agent relationships. *Journal of Economic Perspectives* 5(2):45–66.

Scharfstein, D. S., and J. C. Stein. 1990. Herd behavior and investment. *American Economic Review* 80(3):465–479.

Schlenker, B. R. 1982. Translating actions into attitude: An identity-analytic approach to the explanation of social conduct. In L. Berkowitz, ed., *Advances in Experimental Social Psychology*, vol. 15 (pp. 194–208). New York: Academic Press.

Scott, M., and S. Lyman. 1968. Accounts. *American Sociological Review* 33:46–62.

Sedikides, C., K. C. Herbst, D. P. Hardin, and G. J. Dardis. 2002. Accountability as a deterrent to self-enhancement: The search for mechanisms. *Journal of Personality and Social Psychology* 83(3):592–605.

Siegel-Jacobs, K., and J. F. Yates. 1996. Effect of procedural accountability and outcome accountability on judgment quality. *Organizational Behavior and Human Decision Processes* 65(1):1–17.

Simonson, I., and B. M. Staw. 1992. Deescalation strategies: A comparison of techniques for reducing commitment to losing courses of action. *Journal of Applied Psychology* 77(4):419–426.

Suskind, R. 2006. *The one-percent doctrine: Deep inside America's pursuit of its enemies since 9/11*. New York: Simon and Schuster.

Taleb, N. N. 2007. *The black swan: The impact of the highly improbable*. New York: Random House.

Tetlock, P. E. 1985. Accountability: The neglected social context of judgment and choice. In B. Staw and L. Cummings, eds., *Research in organizational behavior*, vol. 7 (pp. 297–332). Greenwich, CT: JAI Press. Reprinted in L. Cummings and B. Staw, eds., *Research in organizational behavior: Judgment processes*. Greenwich, CT: JAI Press.

Tetlock, P. E. 1992. The impact of accountability on judgment and choice: Toward a social contingency model. In M. Zanna, ed., *Advances in experimental social psychology*, vol. 25 (pp. 331-376). New York: Academic Press.

Tetlock, P. E. 2000. Cognitive biases and organizational correctives: Do both disease and cure depend on the ideological beholder? *Administrative Science Quarterly* 45:293–326.

Tetlock, P. E. 2005. *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Tetlock, P. E., and R. Boettger. 1989. Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology* 57(3):388–398.

Tetlock, P. E., and R. Boettger. 1994. Accountability amplifies the status quo effect when change creates victims. *Journal of Behavioral Decision Making* 7:1–23.

Tetlock, P. E., and L. I. Kim. 1987. Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology* 52(4):700–709.

Tetlock, P. E., and G. Mitchell. 2009. Implicit bias and accountability systems: What must organizations do to prevent discrimination? In B. M. Staw and A. Brief, eds., *Research in organizational behavior*, vol. 29 (pp. 3–38). New York: Elsevier.

Tetlock, P. E., and F. Vieider, 2011. *Accountability, agency, and ideology: Exploring managerial preferences for process versus outcome accountability*. Unpublished manuscript, Wharton School of Business, University of Pennsylvania.

Wilson, J. Q. 1989. *Bureaucracy: What government agencies do and why*. New York: Basic Books.