**Restructuring Structured Analytic Techniques in Intelligence**

Welton Chang

Elissabeth Berdini

**Acknowledgements:** The authors thank their advisors David Mandel and Philip Tetlock for their invaluable comments and criticisms. The authors also thank several anonymous Intelligence Community reviewers, Steve Rieber and Jeff Friedman for helpful comments. We also thank the Intelligence Advanced Research Projects Activity for its generous support and Jesus Chavez for research assistance.

**Disclaimer**: The views presented are those of the authors and do not represent the views of the Intelligence Advanced Research Projects Activity, the Department of Defense or any of its components, or the United States Government; nor do the views presented represent the views of the Department of National Defence or any of its components or the Government of Canada.

**Abstract**

Structured analytic techniques (SATs) are intended to improve intelligence analysis by checking the two canonical sources of error: systematic biases and random noise. Although both goals are achievable, no one knows how close the current generation of SATs comes to achieving either of them. We identify two root problems: (1) SATs treat bipolar biases as unipolar. As a result, we lack metrics for gauging possible over-shooting—and have no way of knowing when SATs that focus on suppressing one bias (e.g., over-confidence) are triggering the opposing bias (e.g., under-confidence); (2) SATs tacitly assume that problem decomposition (e.g., breaking reasoning into rows and columns of matrices corresponding to hypotheses and evidence) is a sound means of reducing noise in assessments. But no one has ever actually tested whether decomposition is adding or subtracting noise from the analytic process—and there are good reasons for suspecting that decomposition will, on balance, degrade the reliability of analytic judgment. The central shortcoming is that SATs have not been subject to sustained scientific of the sort that could reveal when they are helping or harming the cause of delivering accurate assessments of the world to the policy community.

**Introduction: What Are SATs?**

Structured analytic techniques (SATs) are "mechanism[s] by which internal thought processes are externalized in a systematic and transparent manner so that they can be shared, built on, and easily critiqued by others."[1] The aim of SATs is to improve the quality of intelligence analysis by mitigating commonly observed cognitive biases in analysts.[2]

SATs are a central feature of U.S. intelligence analysis training programs. The earliest SATs date to the 1970s, when Richards Heuer introduced the idea to CIA intelligence officer, Jack Davis. Heuer and Davis began developing the first of these methods, which they called "alternative analysis."[3] This term was tweaked to "structured analytic techniques" and referred to as such in CIA's analyst training program in 2005.[4] Post-9/11 Congressional reforms known as the Intelligence Reform and Terrorism Prevention Act of 2004 (IRTPA) affirmed their doctrinal position within the IC.[5] IRTPA mandated the use of SATs as part of a broader IC re-vamp in response to 9/11 and the misjudgment of Iraq's weapons of mass destruction (WMD) programs.[6] The techniques are also taught in DIA's Professional Analyst Career Education (PACE) program and other IC training courses (focusing on techniques used by individuals or small teams as opposed to prediction markets, crowdsourcing, and war-gaming).

Warren Fishbein and Greg Treverton, the former chair of the National Intelligence Council, described the purpose of SATs as follows:

> "… to help analysts and policy-makers stretch their thinking through structured techniques that challenge underlying assumptions and broaden the range of possible outcomes considered. Properly applied, it serves as a hedge against the natural tendencies of analysts—like all human beings—to perceive information selectively through the lens of preconceptions, to search too narrowly for facts that would confirm rather than discredit existing hypotheses, and to be unduly influenced by premature consensus within analytic groups close at hand."[7]

At their core, SATs are a set of processes for externalizing, organizing, and evaluating analytic thinking. SATs range from the simple (e.g., structured brainstorming) to the complex (e.g., scenario generation); from pure text-based methods (e.g., key assumptions check) to visual ones (e.g., argument mapping). The total set of SATs presents a cornucopia of cognitive enhancers from which analysts can choose the one best suited for the problem at hand. The CIA Tradecraft Primer identifies 12 SATs, grouped into three categories, defined by its chief analytic purpose: to foster diagnostic, contrarian, or imaginative thinking.[8] Diagnostic techniques are "aimed at making analytic arguments, assumptions, or intelligence gaps more transparent."[9] They are typically employed to evaluate hypotheses, to assess how existing evidence supports or refutes hypotheses, and to help analysts size up new evidence in light of existing information.[10] The contrarian or "challenge-analysis" techniques are used to probe existing analyses and conclusions, particularly assessments of the stability of the status quo, as well as to test the robustness of the consensus view. These SATs share "the goal of challenging an established mental model or analytic consensus in order to broaden the range of possible explanations or estimates that are seriously considered."[11] Finally, Imaginative Thinking techniques are aimed at "developing new insights, different perspectives and/or developing alternative outcomes".[12]

In the following sections, we delve into the claimed benefits of SATs over unaided analysis. We then describe, in detail, the flaws in the conception and design of SATs which render the claimed benefits implausible. Finally, we offer recommendations for improving the development, testing, and evaluation of SATs by better integrating scientific findings. As some intelligence scholars have recognized, the burden of proof should fall on those who urge the intelligence community (IC) to devote resources to SATs—or indeed any other cognitive-enhancement intervention.[13]

## Why SATs Are Used in Intelligence

SATs are used in intelligence analysis to help analysts: (a) produce more objective and credible judgments by mitigating cognitive biases;[14] (b) cope with information overload; and (c) make their thought processes more rigorous, consistent, and transparent, both to themselves and to policy makers, serving as an accountability mechanism.[15] SATs are supposed to chart a feasible middle ground between methods that are rigorous but technically daunting (e.g., Bayesian networks) and unbridled intuition. The objective of SATs is beyond reproach. Below we sketch three claimed categories of benefits.

## Claim 1: SATs Improve Judgment Quality through Debiasing

Proponents claim that SATs "reduce the frequency and severity of error" of intelligence assessments and estimates.[16] Debiased reasoning can, under the right circumstances, produce more accurate judgments.[17] Some base this claim in SATs' face validity. For instance, Heuer argues, "if a structured analytic technique is specifically designed to mitigate or avoid one of the proven problems in human thought processes, and if the technique *appears* to be successful in doing so, that technique can be said to have face validity [emphasis added]."[18] Table 1 lists the 12 SATs in the CIA Tradecraft Primer and their putative targeted bias.[19] Status quo bias and confirmation bias are the two most common biases that SATs are intended to mitigate.

**Table 1. SATs Target Cognitive Biases**

| Technique | Category | Description of Technique | Targeted Cognitive Bias or Limitation |
|---|---|---|---|
| Key Assumptions Check | Diagnostic | List and review assumptions on which fundamental judgments rest | Status quo bias; attribution error; wishful thinking |
| Quality of Information Check | Diagnostic | Evaluate reliability, completeness and soundness of available information sources. | Status quo bias; confirmation bias; selective exposure; inadequate search |
| Indicators or Signposts of Change | Diagnostic | Periodically review observable trends to track events, monitor targets, and warn of change. | Anchoring and under-adjustment |
| Analysis of Competing Hypotheses (ACH) | Diagnostic | Identify alternative explanations and evaluate evidence bearing on hypotheses. | Status quo bias; confirmation bias; attribution error; selective exposure; congruence bias; anchoring and under-adjustment |
| Devil's Advocacy | Contrarian | Challenge consensus by building strong cases for alternatives. | Confirmation bias, Status quo bias |
| Team A/Team B | Contrarian | Use of separate analytic teams that contrast two (or more) views. | Confirmation bias, Status quo bias |
| High-Impact/Low-Probability Analysis | Contrarian | Highlight unlikely events with potential policy impact. | Status quo bias |
| "What If?" Analysis | Contrarian | Assume a high-impact event has occurred and explain why. | Confirmation bias, Status quo bias |
| Brainstorming | Imaginative | Use an uninhibited group process to generate new ideas. | Status quo bias |

| | | | |
|---|---|---|---|
| Outside-In Thinking | Imaginative | Identify the full range of basic forces and trends that could shape an issue. | Status quo bias; errors in syllogism, illogical arguments |
| Red Team Analysis | Imaginative | Try to replicate how an adversary would think about an issue. | Confirmation bias; attribution error; mirror-imaging |
| Alternative Futures Analysis | Imaginative | Explore multiple ways a highly uncertain situation can develop. | Status quo bias |

Status quo bias, targeted by 10 of the 12 core SATs, occurs when analysts overweight the "no-change" hypothesis in their assessments.[20] In theory, SATs can check status quo bias by encouraging analysts to consider the plausibility of abrupt-surprise scenarios and, by implication, the fragility of the current equilibrium of forces.

Other SATs are geared to check confirmation bias, which can compromise analysts' work by restricting attention to a favored hypothesis, interpreting evidence in ways that bolster existing beliefs, and unfavorably viewing evidence that undercuts beliefs.[21] In an ethnographic study of analysts, Rob Johnston noted that IC culture pushes analysts to conform to the community consensus.[22] This finding was not limited to junior analysts; quite the opposite. Experience and expertise were positively correlated with favoring the consensus.[23] According to James Bruce, the IC's inaccurate assessment of Iraq's WMD efforts might well have been different if analysts had used the right SATs: "what little observable evidence there was of [Iraq's nuclear reconstitution]… was not only over interpreted but also was not assessed relative to any available evidence to the contrary."[24] Bruce conjectures that had analysts used ACH, they would have been required to consider alternative hypotheses and more closely examine disconfirming evidence, which could at least have lowered confidence in the favored

hypothesis.[25] Karl Spielmann similarly suggests that testing many hypotheses can help analysts overcome prevailing mindsets by ensuring relevant information is not overlooked, such as cultural differences between our culture and those of our adversaries.[26] For example, Devil's Advocacy attempts to mitigate confirmation bias by encouraging divergent perspectives.

**Claim 2: SATs Organize Complex Evidence**

SATs are also supposed to help analysts: (a) organize information; (b) identify relevant and diagnostic reporting, pulling useful signals from background noise. Heuer and Pherson state that SATs "break down a specific analytic problem into its component parts and [specify] a step-by-step process for handling these parts."[27] Treverton further argues that SATs are especially suited for solving intelligence issues that deal with transnational actors (e.g., terrorist groups), which offer limited historical-context clues, are not easily detected and monitored, and are often said to act unpredictably.[28]

**Claim 3: SATs Instill Rigor and Make Analysts' Thought Processes Transparent, Thus More Logically Sound**

SAT proponents believe that by systematically structuring analysis, the IC is simultaneously improving the accuracy and objectivity of analysis while acknowledging that analysts must be accountable to high standards of reasoning.[29] SATs are intended to push analysts to think thoroughly and rigorously by decomposing problems and externalizing their thought processes so that reasoning mistakes can be corrected.[30] Heuer also states that exposing one's arguments to scrutiny encourages transparency, which "helps ensure that differences of opinion among analysts are heard and seriously considered early in the analytic process."[31]

SATs thus make it easier to see how others arrived at different judgments. If analysts have conflicting views, they can retrace the steps in their thought processes to see where and

why they diverged in their assessments. This can stimulate discussions of different interpretations of evidence or of varying views of the veracity of sources. And it can bring to light information that some had overlooked, properly pooling previously private information. Importantly, transparency enables other analysts to identify errant thinking, whether it stems from insufficient search or faulty interpretations of evidence found. Transparency might also boost policymakers' valuations of IC assessments. Policymakers bear responsibility for outcomes, so their decisions ultimately must be justifiable to the public.[32]

**Two Core Reasons Why SATs Are Unlikely to Deliver on Promised Benefits: Bias Bipolarity and Noise Neglect**

Unfortunately, little is known about whether structured analytic techniques improve, have no effect on or even degrade analysis because there is scant scientific research on their effectiveness. We propose that the root problems with SATs derive from the failure to deploy best practices in coping with the two fundamental sources of error that bedevil all efforts to improve human judgment: systematic bias and random noise in the processes by which people generate, test and update their hunches about the world. Specifically, SATs fail to address (a) the inherently bipolar nature of cognitive biases and the omnipresent risk that well-intentioned attempts to reduce one bias, say, over-confidence, will amplify its opposing bias, under-confidence; (b) the cumulative nature of error in multi-stage assessments, in which the noise in the conclusions from the one stage gets fed into the next—and the risk that well-intentioned efforts to reduce noise by decomposing complex judgments into sequences of simpler ones will make those judgments less, not more, consistent—an oversight we call noise neglect. Both problems stem from the lack of sustained efforts to subject SATs to scientific tests of efficacy and scrutinizing the processes for logical validity.[33] The net result is that it is difficult to know

when SATs are sparing us from serious mistakes, when they are creating more problems than they are solving, or when they are just ineffective.

**Figure 1. Effects of Noise Neglect and Bias Bipolarity on Judgmental Accuracy**[34]

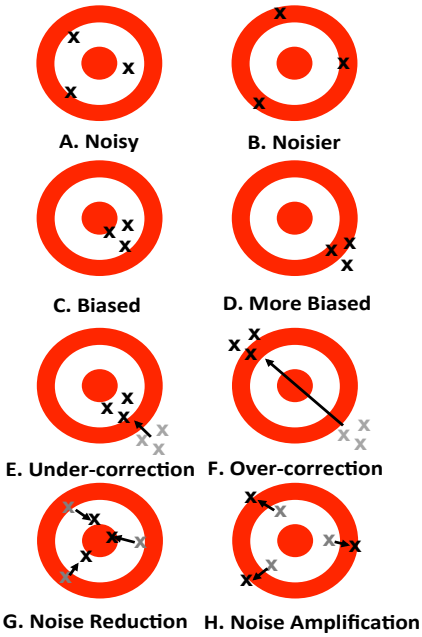


Figure 1 illustrates the core concepts of bias bipolarity and noise neglect that guide our critique. The first row of Figure 1 depicts variation in the noisiness of judgments, which is a function of how far our independent efforts to hit the same target fall from each other. Suppose analysts are all trying to estimate when the latest North Korean missile will be launched—and that the analysts are working from identical data. The predictions displayed in Target B are manifestly noisier—which is to say, inferior to—those displayed in Target A. All else equal, a rational IC would prefer analysts making the forecasts depicted in Target A. The noisier the judgments, the greater the risk of being wrong, dramatically wrong—and that risk is not offset by

the occasional lucky random hit. It follows that the IC should want SATs that, all else equal, bring down noise (i.e., random variation in analytic judgments of exactly the same data).

The second row depicts variation in bias, which is a function of how systematically our independent efforts to hit the same target deviate from the bull's eye. The Target C judgments display a directional bias and the Target D forecasts displays an even more pronounced one. The amount of noise in the Targets C and D judgments is identical—and also identical to the amount of noise in Target A. If the IC worked in a world in which biases were overwhelmingly unidirectional, falling in one quadrant, it would not have to worry about the risks raised by SATs concentrating on reducing over-confidence, status quo bias or excessive deference to the consensus. But there are good reasons for doubting we live in such a world—and for worrying that warning only about over-confidence will eventually cause under-confidence, that pushing single-mindedly against the status quo bias will eventually trigger over-predictions of change and that straying far from consensus thinking will eventually cause us to waste time on frivolous contrarian arguments. All too often, SATs focus on only one pole of the bipolar bias continuum.[35]

The third row depicts what can go wrong when SATs fail to push hard enough against the prevailing bias (Target E) or push too hard against it and trigger a mirror-image bias (Target F). SATs based on a unipolar model of bias are at continual risk of producing a bipolar backlash effect. Getting closer to the bull's eye requires SATs that avoid the mistakes of both pushing too little and pushing too hard. Helping analysts pull off this balancing act requires them to the bipolar nature of the bias, giving them feedback on their current susceptibility to both types of error, and encouraging them to experiment with evidence-weighting strategies that bring them closer to the bull's eye, without over-shooting the mark.

The fourth row focuses on what can go wrong when SATs fail to recognize that problem decomposition is a two-edged sword that can amplify or reduce unreliability, depending on how procedures are interpreted and executed. Target G shows the result of a successful effort to bring down measurement error (relative to Target A) and Target H shows the result of a badly botched effort to reduce unreliability (again, relative to Target A).

In our view, SATs rest on a deep ontological misconception about the nature of bias and a deep methodological misconception about the nature of measurement error and how best to reduce it.

Recently, Chang and Tetlock identified examples of lopsided analytic training that played up one side of the error equation—the need to be wary of overconfidence and of underweighting the potential for change—and downplayed the other side—the risks of under-confidence and of exaggerating the prospects for change.[36] Cognitive biases can cut in either direction, and SATs that focus exclusively on one pole of a bipolar bias render analysts vulnerable to the mirror-image bias, a vulnerability that becomes most obvious when the environment shifts. For example, people tend to be under-confident on easier tasks and over-confident when faced with harder ones.[37] Similarly, depending on the environment, people's tendency to neglect base rates can cause them either to over-weight or under-weight the status quo."[38]

Improving judgment via SATs requires balancing opposing biases. Debiasing is akin to growing a houseplant: under-water and watch it die; over-water and get the same result. "Add-water-and-watch-it-grow" is not a viable horticultural strategy, just as "add-SATs-and-watch-analysis-flourish" is not a viable cognitive psychological strategy. This is because the biases that SATs are designed to prevent are associated with opposing biases that can come into play when SATs overcorrect for the original errors. The problem is that SATs, as treatments of cognitive

bias, tend to be applied with the assumption that both the degree and direction of bias afflicting analysts is known when in fact this is rarely known. The worst-case scenario arises when the SAT is so effective that it causes the analyst to exhibit the *opposite* bias. Unidirectional debiasing techniques are brittle because the degree and direction of bias among analysts is often moderated by situational factors, such as time pressure, and individual differences, such as the extent to which analysts naturally seek disconfirming information and their level of expertise on a subject.[39]

Robust debiasing strategies should aim to reduce deviations from accuracy by mitigating both the bias and the opposing one. This concept is distinct from the idea of debiasing by reducing the magnitude of one pole of a bias. As an example, take calibration feedback, which targets both over- and under-confidence, and contrast that with techniques that warn against the perils of overconfidence—trying to move people in a particular direction just the right amount requires knowing which direction and by how much.[40]

Thus, SATs apply a one-size-fits-all-biases approach to judgmental correction. But analysts vary in the amount of bias they exhibit based on the specifics of the situation, including variables such as evidentiary quality, target behavior and changeability, and even the past accuracy of judgments. Indeed, analysts could bias their judgments so that they wind up not "making the last mistake"-- potentially what occurred in over-connecting the dots in the Iraq WMD case as a response to under-connecting the dots in the case of 9/11.[41] Similarly, target behavior that does not change often can result in analysts overly favoring the status quo, which appeared to be the case in the IC's missed calls on Russia's invasion of Crimea and the Arab Spring uprisings. Analysts also vary in their ability and motivation to self-identify instances of biased cognition. In the worst case, analysts may mistakenly believe they are perfectly unbiased

or that they are exhibiting one bias when they are actually exhibiting the opposing one. SATs depend on analysts and their managers correctly identifying the direction and magnitude of their own biases, probably a rarely satisfied precondition. Many arguments in favor of SATs come after major intelligence failures and which SATs to use in future analysis is thus informed by wanting to prevent another "PLA crossing the Yalu" or "Soviet missile gap" misjudgment. However, such hindsight-tainted inquiries may actually lead to the adoption of techniques that are a net negative for improving judgments on a future issue. For example, analysts concerned with how much weight they are giving to the status quo might adopt a Devil's Advocacy approach which, in turn, puts them at risk of over-weighting engaging-but-low-probability scenarios. Brainstorming and scenario generation might have made 9/11-style attacks more imaginable, but by attending to too many far-fetched scenarios, the net effect might well have been to obscure more plausible attack avenues, such as homegrown extremists conducting mass killings with assault weapons.

Without proper measurement, it is impossible to gauge the direction and magnitude of biases that limit accuracy. Thus, there is great uncertainty over which biases to prioritize for correction—and how far correction should go before it results in overcorrection. Table 2 shows how opposing biases can arise from one-sided use of SATs.

**Table 2. SATs Can Potentially Trigger Opposing Biases**

| Technique | Targeted Bias | Activated Opposing Bias |
|---|---|---|
| Key Assumptions Check | Status quo bias; fundamental attribution error; wishful thinking | Fundamental situational error |

| | | |
|---|---|---|
| Quality of Information Check | Status quo bias; confirmation bias; selective exposure; inadequate search | Under-confidence; excessive search |
| Indicators or Signposts of Change | Anchoring and under-adjustment | Over-confidence; over-adjustment; excessive volatility |
| Analysis of Competing Hypotheses (ACH) | Status quo bias; confirmation bias; attribution error; selective exposure; congruence bias; anchoring and under-adjustment | Incoherent probabilities |
| Devil's Advocacy | Status quo bias | Base-rate neglect |
| Team A/Team B | Status quo bias | Groupthink, Group Polarization |
| High-Impact/Low-Probability Analysis | Status quo bias | Base-rate neglect |
| "What If?" Analysis | Status quo bias | Base-rate neglect |
| Brainstorming | Status quo bias | Incoherent probabilities |
| Outside-In Thinking | Status quo; illogical arguments | Ignorance fallacy; overconfidence |
| Red Team Analysis | Confirmation bias; attribution error; mirror-imaging | Groupthink; overconfidence; "otherness": believing others to be fundamentally irrational |
| Alternative Futures Analysis | Status quo bias | Incoherent probabilities |

Many SATs are designed to challenge the consensus view. But focusing exclusively on ratcheting down excessive conformity can have the unintended effect of inducing excessive second-guessing. Indeed, a recent study of Canadian strategic intelligence analysts found just that—geopolitical forecasts were systematically under-confident rather than over-confident.[42] Nor was such evidence relegated to a particular experience level of analysts: junior and senior

analysts alike displayed significant under-confidence in forecasting. Experimental research corroborates these findings. In one study, military intelligence analysts who judged the probabilities of two mutually exclusive and exhaustive hypotheses assigned too little certainty to the hypotheses.[43] Since SATs are designed to challenge prevailing lines of reasoning, they could inadvertently undercut the confidence of already under-confident analysts. By couching judgments in unnecessary uncertainty, intelligence agencies water down the indicative value of intelligence. Under-confidence can foster, not reduce, uncertainty in the minds of decision makers.

Table 2 also reveals that SATs heavily focus on preventing status-quo bias, which can activate the opposing bias of base-rate neglect. Daniel Kahneman and Amos Tversky coined the term to describe "situations in which a base rate that is known to a subject, at least approximately, is ignored or significantly underweighted."[44] When analysts are pushed to focus on the possibility of change, they unsurprisingly over-predict change.[45] For instance, "What If?" Analysis and Brainstorming encourage deviant ideas that can steer analysts away from diagnostic information. Underweighting of cross-case base-rates is typically accompanied by the overweighting of case-specific indicators, causing miscalibration of confidence estimates.[46] Specifically, people unreasonably privilege observable, case-specific evidence, and are less sensitive to the predictive validity of each piece of evidence given the prior odds of each outcome occurring (as prescribed by a Bayesian approach).[47] The root problem is that most people have difficulty assessing probabilities without proper training.[48] Studies have shown that when evidence is extreme relative to a base-rate, individuals become 'over-confident' by Bayesian standards, whereas "weak" evidence yields 'under-confident' judgments.[49] Further, individuals tend to be over-confident with a small sample size (i.e., minimal collection of

indicative evidence) and under-confident with a large sample size (i.e., great breadth of indicative evidence), which demonstrates an inability to draw inferences proportional to the strength of available evidence.[50] People likely do not properly weigh new evidence in conjunction with their prior beliefs, causing irrational confidence in judgments. Overall, SATs that target status quo bias could amplify the inflation or deflation of associated probabilities, an unintended consequence that makes the improbable seem more probable than warranted.

**Neglecting Noise: Do SATs Reduce Noise in Judgments?**

In addition to slighting bias bipolarity, we believe that certain SATs unnecessarily introduce noise into an already noisy process, causing inconsistent judgments to become even more inconsistent (in scientific parlance, unreliable). Reliability of assessments is a necessary but not sufficient condition of their validity. Many intelligence analysts are highly skilled, attaining expertise through years of experience and specialized training. If *nothing* has changed, we should expect an expert analyst examining the same evidence using the same SATs to reach the same judgments at different times—that is, to be both internally consistent and temporally stable.[51] However, several types of SATs rely on decomposing and separately addressing the components of a problem (e.g., evidence, hypotheses, conclusions). Unfortunately there is a lot of subjectivity in the decomposition process, in part, due to the ambiguity of the SAT-prescribed processes. This sets the stage for an assumed strength of SATs—proceduralizing thinking processes as an alternative to "mere intuition"—to impair intelligence analysis.

Ultimately, SATs rely on subjective analytic inputs (which extends even to interpreting relatively objective scientific-signature and remote-sensing data). Without well-defined rules that reliably improve interpretation of inputs and sharpen analytic outputs, SATs may serve only as a

vehicle transporting subjectivity from one end of the process to the other. The SAT process becomes an end in itself, dressing up subjective judgments in a cloak of objectivity. The inconsistent handling of the decomposed parts—hypotheses, evidence, and hypothesis-evidence linkages—is especially worrisome.

SATs allow inconsistent handling of raw intelligence, particularly judgments of how evidence bears on alternative hypotheses. For example, in exercises involving over 3,000 students from multiple agencies, Heuer and Pherson found that "students disagree on how to rate the evidence in about 20-30 percent of the cells in [an ACH] matrix."[52] This is unsurprising given that the ACH technique does not provide detailed guidance on how "consistency" should be defined, despite the centrality of the concept of "consistency" in all ACH rating exercises.

Analysts may also actively expand the original conceptual framework of a favored hypothesis to accommodate new evidence (whereas confirmation bias shoehorns evidence into the hypothesis). Thus, it is not the *most accurate* hypothesis that emerges as the *most likely* from an SAT framework, but the hypothesis that is *most ideologically consistent* with an analysts' preconceived notions—a problem known as "conceptual stretching" or "elastic redefinition."[53] Analysts may not recognize their favoritism toward a hypothesis, which can affect how leniently or harshly they judge the consistency of evidence for favored or disfavored hypotheses.[54]

SATs also have difficulty handling interdependencies among strands of evidence. Techniques such as ACH and Indicators and Signposts of Change untangle such messes by instructing analysts to break problems into their components. This reductionist treatment ignores the first-order interactions between variables and the feedback loops and subsequent acceleration or deceleration of interactions between variables. The outbreak of World War I emerged from a

complex intertwining of relationships, which is difficult to decompose without distortion.[55] If interdependencies are not accurately captured in simulations, a small variation can cause drastic perturbations.[56] More importantly, the SATs rely on the analyst to properly decompose the problem, offering little guidance on how far to take the reductionist exercise and leading again to the institutionally-sanctioned-subjectivity problem. Beebe and Beebe note that although SATs can help analysts evaluate binary relationships, they do not address the feedback loops and chained interactions involving many variables, including elements outside those captured in the existing analytic space.[57] The difficulty in representing the interactions between variables is what makes complex situations so difficult to understand and predict.[58]

How evidence is weighed (determining the relative importance of sources) is another key aspect of analysis largely neglected by structured methods. Without more explicit rules, analysts are left to their own subjective devices. ACH is one of the few methods that do capture the value of different evidence via its 'weighted inconsistency score.' However, ACH needs to provide more explicit guidelines for weighting evidence consistently. Different analysts assigning different weights (i.e., inter-rater disagreement) can lead to incoherent and inaccurate weight assignments.[59] ACH's description for ranking evidence by diagnosticity is subject to interpretation and can thus lead to inconsistent rankings across individuals (and even within individuals depending on when they looked at the evidence (the phenomenon of latent self-contradiction)).

Noise neglect can also lead to biased reasoning. Methods such as ACH do not check on whether analysts have a shared interpretation of "consistency" and "inconsistency" of evidence with hypotheses—leaving the meaning of consistency subjective. The instructions for ACH (e.g., the doctrinaire version in the CIA's primer) are underspecified, and as currently designed, will

often lead to inter-subjective disagreement over what evidence "fits" or "doesn't fit" with a given hypothesis. The CIA's primer on ACH lists "attacks on journalists" as being inconsistent with the Japanese cult Aum Shinrikyo being a "kooky cult" and neutral with relation to it being a "terrorist organization," but reasonable analysts could view attacking the media as a common terrorist tactic.[60] If analysts are thinking this way about evidence within ACH, then ACH is merely a structured approach to facilitating the analyst's use of the representativeness heuristic—in effect, ACH requires the analyst to answer the question, "To what extent does this evidence seem to fit or match this hypothesis?"[61] The representativeness heuristic has been implicated in several judgment biases, including the conjunction fallacy, insensitivity to sample size, and overestimating the probative value of individuating information.[62]

SAT trainers and critics alike point to the need to dispel the illusion of analytic soundness afforded by SATs. If an analyst goes through the technique's motions without significant challenge to their externalized thinking, SATs may provide only a veneer of analytic legitimacy without true improvements to analytic quality. As currently designed, SATs enable imprecision (i.e., unsystematic deviations from accuracy) which can degrade the reliability of judgments – even though the declared goal is to force analysts to generate more precise ones. In the absence of new information, the same analyst using the same SAT should generate the same judgment over time (be reliable). Reliability is a necessary but not sufficient factor for establishing validity—analysts should at least not contradict their own prior judgments of exactly the same data.[63] Despite these flaws, the intelligence literature tends to assume that SATs work largely as intended. SAT advocates (and even official documents) often take an uncritical stance on their net utility.[64] These flaws remain unquestioned because SATs have only rarely been subjected to

scientific scrutiny. We therefore see a need for testing the net effectiveness of SATs as well as eliminating known flaws.

**SATs Remain Mostly Untested**

The neglect of bias bipolarity and noise has persisted so long because SATs have never been subjected to sustained scientific validation,[65] a fact that even the most ardent SAT advocates acknowledge.[66] The lack of testing has other downstream impacts, including mistrust in the techniques by rank-and-file analysts. Analysts under time pressure are understandably reluctant to put their faith in time-consuming techniques of unknown value.[67]

To our knowledge, no official report has closely examined whether structuring analysis improves reasoning as measured against the standards in ICD 203, such as proper sourcing of claims, exhibiting clear and logical argumentation, and being more accurate in assessments.[68] In 2007, the IC began formally evaluating intelligence products based on the metrics of objectivity, political independence, breadth of information, timeliness, and proper tradecraft.[69] The IC has only recently begun to systematically evaluate the *accuracy* of its estimates and to connect their accuracy to the tradecraft used to produce them.[70] Heuer and Pherson themselves concede, "the conventional criterion for validating an analytic technique is the accuracy of the answers it provides."[71] They acknowledge that there is currently "no systematic program for evaluating or validating the effectiveness of these techniques."[72]

In an independent review, Coulthart examined the efficacy of SATs based on their impact on rigor and accuracy. Brainstorming proved effective in improving analysis in 40% of cases; however, face-to-face collaborative Brainstorming (the form of Brainstorming endorsed by the CIA Tradecraft Primer) had a consistently detrimental effect on the quality of judgments.[73]

Devil's Advocacy outperformed analyses derived from consensus methods of decision-making in 70% of cases.[74] In a separate study, ACH reduced confirmation bias but only for people without an intelligence background. Those with an intelligence background showed no reduction in confirmation bias.[75] While Devil's Advocacy fared well, these preliminary results suggest that SATs as a whole warrant greater skeptical examination.

Consider also Rob Johnston's reflection on his experience with another form of competitive analysis Red Team Analysis: "[the demographics of the group were] not representative of the adversary we were intended to simulate," underscoring how methods emanating from the best of intentions can nevertheless go awry.[76] According to Johnston, there was only one person in the group who had a cultural background related to the target; that person's contributions were severely undervalued by other group members, who instead favored theories consistent with their own backgrounds.[77] This example demonstrates how mirror-imaging bias still pervades Red Team analysis, contrary to its design intention. In such cases, SATs produced a false sense of opinion diversity.

Although some have objected to evaluating the effectiveness of SATs empirically, Heuer and Pherson state that the concerns "could be largely resolved if experiments were conducted with intelligence analysts using techniques as they are used within the Intelligence Community to analyze typical intelligence issues."[78] There have been promising efforts to satisfy each requirement but they have yet to be integrated into a comprehensive evaluative framework such as the ICD 203 standards.[79] Without a clear performance standard, analysts might only seek to employ SATs because they are formalized within ICD 203, not because they are efficacious.

**Addressing Bias Bipolarity and Noise Neglect to Improve SATs**

SATs should promote analytic assessments that are reliable (i.e., consistent within analyst across time), valid (i.e., based on sound reasoning), and accurate (i.e., corresponding to truth). Achieving this goal starts by fleshing out sub-processes within techniques (how each SAT categorizes evidence, whether the process is collaborative, the situations in which that SAT should be used, etc.). This includes ensuring that SATs start from as strong a theoretical framework as possible. It is also necessary to better integrate logical and probabilistic reasoning to make SATs internally coherent. Most importantly, SATs should be evaluated on whether they lead to accurate judgments and higher quality explanations, so their potential to aid judgment can be better understood.

**Establish More Explicit Rules for Handling Evidence**

We propose establishing explicit rules to weight and categorize evidence to promote consistency in the application of SATs and, more importantly, in the assessments they support. We suggest keeping records to identify the role various types of intelligence information played in assessing previous situations.[80] The relationships between evidence (e.g., raw reporting, open source media, background information) and outcomes would help develop base-rates of event occurrence that, over time, could provide "outside view" checks on portfolios of "inside view" case-based assessments.[81] Such information can be used to more properly weight evidence within techniques such as ACH; meanwhile, analysts' source track records can be used to inform assessments of source credibility when conducting Quality of Information Checks.

Clarifying SAT rules for evaluating evidence can help analysts think critically about how they should interpret new or unfamiliar information, rather than relying on their intuition.[82] Implementing such rules would decrease ambiguity in how new observations are interpreted–

which would, in turn, lead to analysts and analytic groups dealing with evidence more consistently. For example, explicit categorization rules would clarify objectively what constitutes a 'moderate concern' signpost versus a 'substantial concern' signpost within the Indicators and Signposts of Change process. How much should analysts update their beliefs about a country possessing WMD if they observe attempts to acquire aluminum tubes that may or may not be suitable for enrichment? These clarifications should minimize the potential to interpret evidence to 'fit' pet theories.

**Incorporate Probability Theory into SATs**

SATs should include mechanisms to ensure analysts reason probabilistically and can express their probabilistic assessment precisely. Numeric probabilities are preferred but it may sometimes be possible to use verbal probabilities consistently and carefully to achieve adequate precision. The aim of probabilistic assessments is not to misrepresent analysts' judgments as scientific facts, but rather to promote conditions that support clear verification of internally consistent thinking and accuracy over the long run. Such verification, in turn, would dramatically improve intelligence accountability by supplementing a virtually exclusive focus on analytic process with important indicators of accuracy.[83] By evaluating a conclusion such as North Korea is "likely" to collapse by the end of the year, analysts can calibrate their future judgments accordingly.

With precise feedback, analysts would be less likely to over-estimate the likelihood of rare outcomes and more likely to make calibrated assessments that properly identify the relative strengths of the arrayed hypotheses.[84] For improving inter-subjective consistency and agreement, assigning numeric probabilities to hypotheses would enable analysts to communicate more

granular and meaningful estimates to one another and to policymakers.[85] Alan Barnes, when he was director of the Middle East and Africa Division of the Intelligence Assessment Secretariat in the Government of Canada, introduced a nine-point probability scale and found it created a "common understanding of the degree of certainty attached to a judgment," thereby reducing problems of misinterpretation that arise when analysts communicate their judgments to others.[86] Barnes also found that increased experience with using numeric probabilities made analysts more comfortable using them.[87]

**Verify the Efficacy of SATs for Debiasing and Noise Reduction**

We recommend that SATs be tested using the scientific gold standard: control-randomized experiments, designed in such a way that they are sensitive to the specific and unique circumstances that analysts face: time-pressure, evidential uncertainty, and content bearing on national security.[88]

Furthermore, Heuer has noted that two different analysts using the same SAT, evaluating the same analytic problem, can reach different judgments due to different "mental models."[89] However, proponents of SATs should agree that if an SAT is to produce consistent results, the *same* analyst using a given SAT should interpret a given situation the same way each time he encounters it, if in fact nothing has changed. In other words, beliefs should not randomly oscillate in the absence of new information. Thus, to measure SATs for reliability, we propose various tests such as examining test-retest reliability, susceptibility to framing effects, refocusing effects, and whether using methods result in violations of logical constraints on reasoning.[90] For example, assessments of the probability of a terrorist attack in a specified location and timeframe can differ substantially from assessments of the probability of *no* terrorist attack in the same location and timeframe when subtracted from 1. If such assessments were coherent, they would

be the same. However, it appears that querying people about the occurrence or non-occurrence of an event, such as a terrorist attack, often triggers different information search, memory retrieval, and assessment processes.[91] Likewise, when the probability of success on a given type of instance is high (say 90% likely), people judge event descriptions like "exactly 1 success in 4 tries" as more probable than "exactly 3 failures in 4 tries", despite the fact that they refer to the same conjunction (i.e., "1 success and three failures out of 4 tries") and are therefore equiprobable.[92] SATs should help analysts avoid these predictable forms of logical inconsistency, but there is currently no credible evidence that they do.

Specifically, we suggest the following test-retest evaluation of temporal stability: the same analyst is given identical intelligence problems to analyze (the same evidence and objective) a month or two apart. If the technique is reliable, analysts will categorize the components the same way each time, and more importantly, arrive at the same judgment. This is an especially important test for diagnostic techniques. Requiring analysts to break information into components (e.g. evidence, indicators, sources) in the absence of well-specified criteria might lead to an unnecessarily messy process. For example, if an analyst is using Indicators or Signposts of Change and identifies a specific indicator (e.g., military discontent with a civilian government) as a 'Moderate Concern' in Trial 1, then—in the absence of new information—the analyst also should identify it as a 'Moderate Concern' in Trial 2. If we find that analysts' classifications of indicators in Trial 1 and Trial 2 are unreliable, the technique should be improved by creating more explicit rules for categorizing evidence, prior to validity testing. Test-retest reliability experiments have been conducted in professions such as radiology, tax accounting, and auditing.[93] Our proposal would provide insight into the question of whether SATs reduce, increase, or leave unchanged the degree of judgmental imprecision.

Note that the proposed method would not test the value of analytic collaboration, but it could easily by adding another design element. We could devise similar experiments for groups to test collaboration techniques, like Brainstorming and Red Teaming. The group aspect of these SATs is intended to promote a diverse range of opinions, enabling all hypotheses to be considered and the optimal judgment chosen. Thus, it may be advisable to give multiple groups the same analytic project and access to the same evidence, then measure consistency across group judgments rendered with or without the SAT. Does SAT use improve inter-group reliability? We won't know until we test.

**Establishing a Feedback Loop to Continuously Improve SATs**

Intelligence organizations should periodically update and improve SATs in response to data from accuracy assessments. Part of this feedback loop requires scoring assessments for accuracy, which can sometimes be done even when key judgments are expressed as verbal probabilities.[94] The historical accuracy of intelligence forecasts is mostly unknown.[95] By quantifying the likelihoods that analysts assign to hypotheses using a logically coherent process, we can record the frequency and degree to which each analytic structure produces accurate judgments (as measured by skill metrics such as Brier scores), to create reliable performance records. Then, we could continuously catalog all estimative judgments and conduct follow-up post-mortems to cross-check whether the hypothesis that corresponded to the correct outcome was 'on the radar' throughout, and examine why it was undervalued or dismissed.[96] These records could be used to assess current threats by researching which similar past threats analysts assessed accurately, and which factors differentiated failures from success.[97] Johnston refers to this type of system as an "institutional memory"—a library of "lessons learned" that analysts can utilize and add to.[98] A database eventually could develop into a system that allows for automatic

extraction of information relevant to a specific person, date, location, facility, organization, etc.[99] In the long run, this would save analysts time, because they could retrieve information more quickly from a centralized, coded system, and efficiently discover associations throughout the history of the topic.[100]

In addition, making feedback available to individual analysts can help them improve their judgments over time, including measures of calibration and discrimination.[101] Most people don't have an accurate sense of their strengths and weaknesses when assessing uncertainty; performance feedback has been shown to help individuals improve their performance, by keeping records to dispel self-serving illusions of skill.[102] Johnston proposed a "Performance Improvement Infrastructure" to measure individual analytic performance and the impact of different "interventions" (e.g. SATs) on their performance.[103] The performance of intelligence analysts could become a little more like weather forecasters, whose success Griffin and Tversky attributed in part to their getting "immediate frequentist feedback."[104] While intelligence analysis and weather forecasting are different activities (clouds don't have agency, for one thing) and analysts will likely never approach the calibration of meteorologists (who have data rich computer models), feedback may lead to some judgmental improvements. How much is an empirical question that can only be answered after such a system is instituted.

Finally, we propose a new, more scientifically grounded taxonomy for organizing, testing, and generating new SATs. One way of organizing SATs is to determine when best to use them by aligning them with analytic production phases (e.g., during initial drafting, during the community coordination process) and cognitive sub-processes (e.g., during broad information search, during the generation of hypotheses).[105] A more parsimonious way is to align SATs with the deeper cognitive trade-offs that analysts routinely encounter and feature prominently in the

psychological literature. For example, the current official taxonomy of diagnostic, contrarian, and imaginative techniques could be simplified to two categories: helping analysts engage in critical thinking and helping analysts engage in creative thinking. This distinction has been observed by psychologist Tom Gilovich and analogizes SATs to tests that ask "must-I-believe-it?" and tests that ask "can-I-believe-it?"[106] As analysts encounter new evidence, reconsider old evidence in light of new evidence, and generate and test hypotheses, they must balance between minimizing errors of believing things they are not logically obliged to believe (i.e., checks on excessive gullibility). They must also minimize errors of failing to give credence to possibilities they should have given weight (i.e., checks on excessive rigidity). Existing techniques such as foresight analysis help analysts on "can I believe it" type questions and devil's advocacy stress tests analytic conclusions of the "must I believe it" variety.

Put another way, some SATs will enhance foveal vision (i.e., our capacity to rapidly and accurately diagnose threats and opportunities in front of us), which is especially helpful in current and crisis intelligence functions. Others will enhance peripheral vision (i.e., our capacity to anticipate threats that are outside plausibility range of conventional wisdom—thus blending contrarian and imaginative techniques), which is helpful for longer-term assessments and horizon scanning. Aligning SATs with scientifically grounded functions will make it easier to manage the tough cognitive trade-offs analysts regularly encounter.

**Conclusion**

The great 20th century sociologist Robert Merton distinguished between the manifest and latent functions of collective practices and rituals. Rain dances are supposed to serve the manifest function of causing clouds to form and produce rain, but they also serve the latent functions of bringing the tribe together and fostering social cohesion. Merton argued that the

latent functions of a surprisingly wide range of collective activities were actually much more important to people than the manifest or officially declared functions.

Pointing out the possible latent functions of a practice can be risky. Insiders often take umbrage when outsiders suggest that insiders' public reasons for doing *X* are not their real reasons. But a serious scientific inquiry requires taking the risk and posing the impertinent question: does the IC value SATs more for their manifest or latent functions?

Thus far, we have taken the IC at its word and assumed that the manifest functions of SATs— improving analysis—are its real reason for embracing SATs. But the central claims of this article raise questions. If SATs have never been rigorously tested for efficacy, rest on an untenable unipolar conception of cognitive biases, and foster inconsistent judgments: why has this suboptimal state of affairs been allowed to persist for decades?

We see three broad sets of possibilities. First, we have under-estimated the IC and SAT proponents. The IC knows more about the effectiveness of SATs than we claim—and SATs are more effective than we have suppose. Second, the IC knows as little about efficacy as we claim—and it is not particularly interested in learning more. SATs serve valuable bureaucratic-political signaling functions. They send the message that the IC is committed to playing a pure epistemic game and trains its analysts in accordance with quasi-scientific ground rules. Success is measured not along an accuracy metric but along an impression management metric that follows from an intuitive politician mindset: are we creating the desired impression on key constituencies?[107] Third, the IC knows about as much as we worry it does—and would like to learn more—but is skeptical that it is possible to move much beyond face validity.

We are unlikely to convince readers who fall in the first and second camps but we do seek to engage readers in the third. This is because SATs hold much promise. If constructed properly and continuously tested and refined, they probably can help analysts produce more accurate and better-reasoned reports. That SATs remain largely untested is a major problem for their adoption by IC analysts because of the need for demonstrated efficacy to be a part of the SAT "convincing case." Further training on which SATs are best suited for which issues (regional, functional) can only be developed with data from rigorous applied scientific research showing that a specific SAT is well-aligned for a problem sub-type. Sometimes the best SAT may be no-SAT-at-all.

In a nutshell, this paper is an appeal for greater scientific rigor. Science, like intelligence, can be messy. Progress is slow and only comes from efforts to push boundaries. SAT proponents and opponents alike should welcome the fresh attention. The problems with SATs are serious but there are potential fixes. Although there is no guarantee that SATs will prevent the next WMD misjudgment or predict the next major terrorist attack, they could still improve analysis in meaningful and measurable ways. Figuring out how best to design SATs so that they help, not hinder, intelligence analysts should be a top priority.

## Bibliography

Artner, Stephen, Girven Richard S. and Bruce James B. *Assessing the Value of Structured Analytic Techniques in the U.S. Intelligence Community*. Rand Corporation (2016).

Bar-Hillel, Maya. "The base-rate fallacy in probability judgments." *Acta Psychologica* 44, no. 3 (1980): 211-233.

Barnes, Alan. "Making Intelligence Analysis More Intelligent: Using Numeric Probabilities." *Intelligence and National Security* 31, no. 3 (2016): 327-344.

Beebe, Sarah M. and George S. Beebe. "Understanding the Non-Linear Event: A Framework for Complex Systems Analysis." *International Journal of Intelligence and Counter Intelligence* 25, no. 3 (2012): 508-528.

Bruce, James B. "Making Analysis More Reliable: Why Epistemology matters to intelligence." *Analyzing Intelligence: Origins, Obstacles, and Innovations* (2008): 171-190.

Chang, Otto H., and Thomas M. McCarty. "Evidence on Judgment Involving the Determination of Substantial Authority: Tax Practitioners versus Students." *The Journal of the American Taxation Association* 10, no. 1 (1988): 26-39.

Chang, Welton, and Philip E. Tetlock. "Rethinking the training of intelligence analysts." *Intelligence and National Security* 31, no. 6 (2016): 903-920.

Chang, Welton, Eva Chen, Barbara Mellers, and Philip E. Tetlock, "Developing expert political judgment: the impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments." *Judgment and Decision Making* 11, no. 5 (2016): 509-526.

Collier, David, and James E. Mahon. "Conceptual 'Stretching' Revisited: Adapting Categories in Comparative Analysis." *The American Political Science Review* 87, no. 4 (1993): 845-855.

Cooper, Jeffrey R. *Curing* Analytic Pathologies: Pathways to Improved Intelligence Analysis. Central Intelligence Agency, Center for Study of Intelligence, 2005.

Coulthart, Stephen. "Why do analysts use structured analytic techniques? An in-depth study of an American intelligence agency" Intelligence and National Security 31, no. 7 (2016): 933-948.

Dhami, Mandeep K., Ian K. Belton, and Kathryn E. Careless. "Critical Review of Analytic Techniques." *Intelligence and Security Informatics Conference (EISIC)*, 2016 *European*: 152-155.

Fishbein, Warren, and Gregory Treverton. *Rethinking "Alternative Analysis" to Address Transnational Threats*. Central Intelligence Agency (2004).

Friedman, Jeffrey A., and Richard Zeckhauser. "Why Assessing Estimative Accuracy is Feasible and Desirable." *Intelligence and National Security* 31, no. 2 (2016): 178-200.

Greenwood, John D. "Two Dogmas of Neo-Empiricism: The 'Theory-Informity' of Observation and the Quine-Duhem Thesis." *Philosophy of Science* 57, no. 4 (1990): 553-574.

Griffin, Dale, and Amos Tversky. "The weighing of evidence and the determinants of confidence." *Cognitive Psychology* 24, no. 3 (1992): 411-435.

Harris, Douglas H., and V. Alan Spiker. "Critical Thinking Skills for Intelligence Analysis." in *Ergonomics – A Systems Approach*, edited by Isabel L. Nunes. InTech Open Access Publisher, 2012.

Heuer Jr, Richards J. "The evolution of structured analytic techniques." *Presentation to the National Academy of Science, National Research Council Committee on Behavioral and Social Science Research to Improve Intelligence Analysis for National Security* (2009): 529-545.

Hoffman, Paul J., Paul Slovic, and Leonard G. Rorer. "An analysis-of-variance model for assessment of configural cue utilization in clinical judgment." *Psychological Bulletin* 69, no. 5 (1968): 338-349.

Jervis, Robert. *System Effects: Complexity in Political and Social Life*. Princeton University Press, 1998.

Johnston, Rob. *Analytic Culture in the U.S. Intelligence Community: An Ethnographic Study*. Washington D.C.: U.S. Central Intelligence Agency, Center for the Study of Intelligence, 2005.

Kahneman, Daniel, and Amos Tversky. "On the Reality of Cognitive Illusions." *Psychological Review* 103, no. 3 (1996): 582-591.

Mandel, David R. "Accuracy of Intelligence Forecasts From the Intelligence Consumer's Perspective." Policy Insights from the Behavioral and Brain Sciences 2, no. 1 (2009).

Mandel, David R. "Applied behavioural science in support of intelligence: Experiences in building a Canadian capability." (2009).

Mandel, David R. "Are risk assessments of a terrorist attack coherent?" *Journal of Experimental Psychology: Applied* 11, no. 4 (2005): 277.

Mandel, David R. "Instruction in information structuring improves Bayesian judgment in intelligence analysts." *Frontiers in Psychology* 6, no. 387 (2015).

Mandel, David R. "Violations of coherence in subjective probability: A representational and assessment processes account." *Cognition* 106, no. 1 (2008): 130-156.

Mandel, David R., and Alan Barnes. "Accuracy of forecasts in strategic intelligence." *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 30 (2014): 10984-10989.

Mandel, David R., and Philip E. Tetlock. "Debunking the Myth of Value-Neutral Virginity: Toward Truth in Scientific Advertising." *Frontiers in Psychology* 7, no. 451 (2016).

Mandel, David R., Alan Barnes, and Karen Richards. *A quantitative assessment of the quality of strategic intelligence forecasts*. No. 2013-036. Technical Report, 2014.

Marchio, James. "How good is your batting average?" Early IC Efforts to Assess the Accuracy of Estimates. *Studies in Intelligence* 60, no. 4 (2016): 3-13.

Marrin, Stephen. "Evaluating the Quality of Intelligence Analysis: By What (Mis) Measure?" *Intelligence and National Security* 27, no. 6 (2012): 896-912.

Marrin, Stephen. *Improving intelligence analysis: Bridging the gap between scholarship and practice*. Routledge, 2012.

Marrin, Stephen. "Training and educating US intelligence analysts." *International Journal of Intelligence and CounterIntelligence* 22, no. 1 (2009): 131-146.

Meixner, Wilda F., and Robert B. Welker. "Judgment consensus and auditor experience: An examination of organizational relations." *Accounting Review* (1988): 505-513.

Miller, C. Chet, and R. Duane Ireland. "Intuition in Strategic Decision Making: Friend or Foe in the Fast-Paced 21st Century?" *The Academy of Management Executive* (1993-2005) 19, no. 1 (2005): 19-30.

Moore, Don A., and Paul J. Healy. "The trouble with overconfidence." *Psychological Review* 115, no. 2 (2008): 502.

Nickerson, Raymond S. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2, no. 2 (1998): 175.

Gilovich, Thomas. How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life. Simon and Schuster, 1993.

Hammond, Kenneth R. "How convergence of research paradigms can improve research on diagnostic judgment." *Medical Decision Makin*g 16, no. 3 (1996): 281-287.

Haran, Uriel, Ilana Ritov, and Barbara A. Mellers. "The role of actively open-minded thinking in information acquisition, accuracy, and calibration." Judgment and Decision Making 8.3 (2013): 188.

Heuer, Richards J. "Taxonomy of Structured Analytic Techniques." *International Studies Association Annual Convention*, 2008.

Heuer, Richards J. *Psychology of Intelligence Analysis*. Washington D.C.: U.S. Central Intelligence Agency, Center for the Study of Intelligence, 1999.

Ho, Emily H., David V. Budescu, Mandeep K. Dhami, and David R. Mandel. “Improving the communication of uncertainty in climate science and intelligence analysis.” *Behavioral Science & Policy* 1, no. 2 (2015): 43-55.

Kahneman, Daniel, and Amos Tversky. “On the study of statistical intuitions.” *Cognition* 11, no. 2 (1982): 123-141.

Kahneman, Daniel, Andrew M. Rosenfield, Linnea Gandhi, and Tom Blaser. “NOISE: How to overcome the high, hidden cost of inconsistent decision making.” *Harvard* Business R*eview* 94, no. 10 (2016): 38-46.

Kahneman, Daniel, and Amos Tversky. "On the psychology of prediction." Psychological review 80.4 (1973): 237.

Khalsa, Sundri. “The Intelligence Community Debate over Intuition versus Structured Technique: Implications for Improving Intelligence Warning.” *Journal of Conflict Studies* 29 (2009).

Lichtenstein, Sarah, and Baruch Fischhoff. "Training for calibration*." Organizational Behavior and Human Performance* 26, no. 2 (1980): 149-171.

Lehner, Paul Edward, Leonard Adelman, Brant A. Cheikes, and Mark J. Brown. “Confirmation bias in complex analyses.” *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38, no. 3 (2008): 584-592.

Tetlock, Philip E., and Barbara A. Mellers. "Intelligent management of intelligence agencies: beyond accountability ping-pong." *American Psychologist* 66.6 (2011): 542.

Marchio, Jim. “How good is your batting average? Early IC efforts to assess the accuracy of estimates.” *Studies in Intelligence* 60, no. 4 (2016): 3-13.

National Research Council. *Intelligence Analysis: Behavioral and Social Scientific Foundations*. National Academies Press (2011).

Ordonez, Lisa, and Lehman Benson. "Decisions under time pressure: How time constraint affects risky decision making." *Organizational Behavior and Human Decision Processes* 71.2 (1997): 121-140.

Phillips, Lawrence D., and Ward Edwards. "Conservatism in a simple probability inference task." *Journal of experimental psychology* 72.3 (1966): 346.

Piercey, M. David. “Motivated reasoning and verbal vs. numerical probability assessment: Evidence from an accounting context.” *Organizational Behavior and Human Decision Processes* 108, no. 2 (2009): 330-341.

United States Government. “Intelligence Reform and Terrorism Prevention Act of 2004.” *Public Law* 458 (2005): 108.

Rieber, Steven. “Intelligence analysis and judgmental calibration.” International Journal of Intelligence and Counter Intelligence 17, no. 1 (2004): 97-112.

Spielmann, Karl. “I Got Algorithm: Can There Be a Nate Silver in Intelligence?” *International Journal of Intelligence and Counter Intelligence* 29, no. 3 (2016): 525-544.

Tetlock, Philip E., and Barbara A. Mellers. “Intelligent Management of Intelligence Agencies: Beyond Accountability Ping-Pong.” *American Psychologist* 66, no. 6 (2011): 542.

Tetlock, Philip E., and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Random House, 2015.

Tikuisis, Peter, and David R. Mandel. “Is the World Deteriorating?.” *Global Governance: A Review of Multilateralism and International Organizations* 21, no. 1 (2015): 9-14.

Tversky, Amos, and Daniel Kahneman. “Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.” *Psychological Review* 90, no. 4 (1983): 293.

Walton, Timothy. *Challenges in Intelligence Analysis: Lessons from 1300 BCE to the Present*. Cambridge University Press, 2010.

Wegener, Duane T., and Richard E. Petty. “The flexible correction model: The role of naïve theories of bias in bias correction.” *Advances in Experimental Social Psychology* 29 (1997): 141-208.

West, Richard F., Maggie E. Toplak, and Keith E. Stanovich. "Heuristics and biases as measures of critical thinking: associations with cognitive ability and thinking dispositions." *Journal of Educational Psychology* 100.4 (2008): 930.

## Notes

1 Heuer and Pherson, *Structured Analytic Techniques for Intelligence Analysis*, 4.
2 Heuer, *Psychology of Intelligence Analysis*, 95.; Heuer and Pherson, *Structured Analytic Techniques for Intelligence Analysis*, 4.
3 Others, including Pherson, Heuer, and Mandeep Dhami and colleagues, have published more complex taxonomies, that match SATs to the cognitive processes which analysts initiate while on task; Central Intelligence Agency, *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*, 5.; Heuer and Pherson, *Structured Analytic Techniques*.; Dhami, Belton, and Careless, “Critical Review of Analytic Techniques.”; Heuer, “Taxonomy of Structured Analytic Techniques.”

4 Coulthart, “Why do analysts use structured analytic techniques? An in-depth study of an American intelligence agency.”; Heuer and Pherson, *Structured Analytic Techniques*.
5 *Intelligence Reform and Terrorism Prevention Act of 2004.*; Marrin, “Training and Educating U.S. Intelligence Analysts.”
6 Chang and Tetlock, “Rethinking the training of intelligence analysts.”; Heuer and Pherson, *Structured Analytic Techniques*, 9.; Coulthart, “Improving the Analysis of Foreign Affairs: Evaluating Structured Analytic Techniques,” 41.; Coulthart, “Improving the Analysis of Foreign Affairs: Evaluating Structured Analytic Techniques,” 4.
7 Fishbein and Treverton, “Rethinking ‘Alternative Analysis’ to Address Transnational Threats,” 1.
8 Central Intelligence Agency, *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*, 5.
9 *Ibid*.
10 Heuer and Pherson, *Structured Analytic Techniques*, 158.
11 Heuer and Pherson, *Structured Analytic Techniques*, 215.
12 Central Intelligence Agency, *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*, 5.
13 Cooper, *Curing Analytic Pathologies: Pathways to Improved Intelligence Analysis*; Mandel, *Applied behavioural science in support of intelligence: Experiences in building a Canadian capability.*; Marrin, *Improving Intelligence Analysis: Bridging the Gap Between Scholarship and Practice*.; Heuer, “The Evolution of Structured Analytic Techniques.”
14 Heuer and Pherson, *Structured Analytic Techniques*, 5.
15 Heuer and Pherson, *Structured Analytic Techniques*, 4.
16 Heuer and Pherson, *Structured Analytic Techniques*, 5.
17 Chang, Chen, Mellers and Tetlock, “Developing expert political judgment: the impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments.”
18 Heuer and Pherson, *Structured Analytic Techniques*, 309.
19 Central Intelligence Agency, *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*.
20 Chang and Tetlock, “Rethinking the training of intelligence analysts.”; Heuer, *Psychology of Intelligence Analysis*.
21 Nickerson, “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises,” 175.
22 Johnston, *Analytic culture in the US intelligence community: An ethnographic study*, 21.
23 Harris and Spiker, *Critical Thinking Skills for Intelligence Analysis*, 217.
24 George and Bruce, ed., “Making Analysis More Reliable: Why Epistemology Matters to Intelligence.”
25 *Ibid*.
26 Spielmann, “I Got Algorithm: Can There Be a Nate Silver in Intelligence?”
27 Heuer and Pherson, *Structured Analytic Techniques*, 4.
28 Central Intelligence Agency, *Rethinking “Alternative Analysis” to Address Transnational Threats*.
29 Greenwood, “Two Dogmas of Neo-Empiricism: The ‘Theory-Informity’ of Observation and the Quine-Duhem Thesis.”
30 Heuer and Pherson, *Structured Analytic Techniques*, xv, 9.
31 Heuer and Pherson, *Structured Analytic Techniques*, 4.

32 Spielmann, "I Got Algorithm: Can There Be a Nate Silver in Intelligence?," 526.
33 We should address the false dichotomy between structured and intuitive reasoning. Advocates of SATs portrayed them as an "alternative" to "traditional" (intuition-based) analysis. Advocates imply that without SATs, analysts are left with only their intuition to make sense of the world. It should be noted that unaided analysis is not the same as intuition. Unaided analysis can still be effortful, deliberate thinking. Analysts have a wide array of potential strategies, such as with deductive, inductive, or abductive forms of reasoning. SAT advocates are leaving a lot of reasoning improvement gains on the table that could be achieved by honing deliberate thinking.
34 Adapted from Figure "How Noise and Bias Affect Accuracy" in Kahneman, Rosenfield, Gandhi, and Blaser, "Noise: How to overcome the high, hidden cost of inconsistent decision making."
35 There remains significant debate regarding the efficacy of a variety of debiasing methods, see the discussion in Croskerry P, Singhal G, Mamede S, Cognitive debiasing 1: origins of bias and theory of debiasing, *BMJ Quality and Safety*: 23 July 2013.
36 Chang and Tetlock, "Rethinking the training of intelligence analysts."
37 Lichtenstein and Fischhoff, "Training for calibration."
38 Philips and Edwards, "Conservatism in a simple probability inference task."; Kahneman and Tversky, "On the Psychology of Prediction."
39 Ordonez and Benson, "Decisions under time pressure: How time constraint affects risky decision making."; Haran, Ritov and Mellers, "The role of actively open-minded thinking in information acquisition, accuracy, and calibration."; West, Toplak and Stanovich, Heuristics and biases as measures of critical thinking: associations with cognitive ability and thinking dispositions."; Wegener and Petty, "The flexible correction model: The role of naïve theories of bias in bias correction."; Wegener, Dunn, and Tokusato, "The flexible correction model: phenomenology and the use of naïve theories in avoiding or removing bias."
40 Moore and Healy, "The trouble with overconfidence," 502.
41 Mellers and Tetlock, "Intelligent management of intelligence agencies: beyond accountability ping-pong."
42 Mandel and Barnes, "Accuracy of forecasts in strategic intelligence."
43 Mandel, "Instruction in information structuring improves Bayesian judgment in intelligence analysts."
44 Kahneman and Tversky, "On the reality of cognitive illusions."
45 Chang and Tetlock, "Rethinking the training of intelligence analysts."
46 Bar-Hillel, "The base-rate fallacy in probability judgments."
47 Griffin and Tversky, "The Weighing of Evidence and the Determinants of Confidence."
48 Lehner, Adelman, Cheikes, Brown, "Confirmation Bias in Complex Analyses."
49 Griffin and Tversky, "The Weighing of Evidence and the Determinants of Confidence."
50 *Ibid*.
51 Kahneman, Rosenfield, Gandhi, and Blaser, "Noise: How to overcome the high, hidden cost of inconsistent decision making."
52 Heuer and Pherson, *Structured Analytic Techniques*, 311.

53 Chang and Tetlock, "Rethinking the training of intelligence analysts."; Collier and Mahon, "Conceptual 'Stretching' Revisited: Adapting Categories in Comparative Analysis."; Piercey, "Motivated Reasoning and Verbal vs. Numerical Probability Assessment: Evidence from an Accounting Context."
54 Chang and Tetlock, "Rethinking the training of intelligence analysts."
55 Jervis, *System Effects: Complexity in Political and Social Life*.
56 *Ibid*.
57 Beebe and Beebe, "Understanding the Non-Linear Event: A Framework for Complex Systems Analysis," 511.
58 *Ibid*.
59 Spielmann, "I Got Algorithm: Can There Be a Nate Silver in Intelligence?," 535.
60 CIA, "A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis", 15.
61 Kahneman and Tversky, "On the study of statistical intuitions."
62 Tversky and Kahneman, "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment."
63 Kahneman, Rosenfield, Gandhi, Blaser, "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making."
64 According to the CIA Tradecraft Primer, "Although application of these techniques alone is no guarantee of analytic precision or accuracy of judgments, it does improve the sophistication and credibility of intelligence assessments" 5.; Khalsa, "The Intelligence Community Debate over Intuition versus Structured Technique: Implications for Improving Intelligence Analysis and Warning."; Walton, *Challenges in Intelligence Analysis: Lessons from 1300 BCE to the Present*.
65 Artner, Girven and Bruce, Assessing the Value of Structured Analytic Techniques, 4.
66 Heuer, "The Evolution of Structured Analytic Techniques."
67 Coulthart, "Why do analysts use structured analytic techniques? An in-depth study of an American intelligence agency."; Artner, Girven and Bruce, Assessing the Value of Structured Analytic Techniques, 3.
68 Intelligence Community Directive 203, *Analytic Standards*.
69 Friedman and Zeckhauser, "Why Assessing Estimative Accuracy is Feasible and Desirable," 7.
70 Friedman and Zeckhauser, "Why Assessing Estimative Accuracy is Feasible and Desirable," 2.
71 Heuer and Pherson, *Structured Analytic Techniques,* 312.
72 Heuer and Pherson, *Structured Analytic Techniques,* 309.
73 Coulthart, "Why do analysts use structured analytic techniques? An in-depth study of an American intelligence agency."
74 *Ibid*.
75 Coulthart, "Why do analysts use structured analytic techniques? An in-depth study of an American intelligence agency."
76 Johnston, *Analytic culture in the US intelligence community: An ethnographic study*, 81.
77 Johnston, *Analytic culture in the US intelligence community: An ethnographic study*, 81-82.
78 Heuer and Pherson, *Structured Analytic Techniques*, 312.

79 Hammond, "How convergence of research paradigms can improve research on diagnostic judgment."; Marrin, "Evaluating the Quality of Intelligence Analysis: By What (Mis) Measure?," 896.; Friedman and Zeckhauser, "Why Assessing Estimative Accuracy is Feasible and Desirable," 6.
80 Spielmann, "I Got Algorithm: Can There Be a Nate Silver in Intelligence?," 535.
81 Tikuisis and Mandel, "Is the World Deteriorating?"
82 Harris and Spiker, *Critical Thinking Skills for Intelligence Analysis*, 217.
83 Tetlock and Mellers, "Intelligent Management of Intelligence Agencies: Beyond Accountability Ping-Pong."; Barnes, "Making Intelligence Analysis More Intelligent: Using Numeric Probabilities," 330.
84 Chang and Tetlock, "Rethinking the training of intelligence analysts."
85 *Ibid*.
86 Barnes, "Making Intelligence Analysis More Intelligent: Using Numeric Probabilities," 333.
87 *Ibid*.
88 We note that the IC-sponsored research project, Crowdsourcing Reasoning Evidence Analysis Thinking and Evaluation (CREATE), which three of the authors have or are currently participating in, is in the process of doing exactly this.
89 Heuer and Pherson, *Structured Analytic Techniques.*
90 Inspired by the noise audit idea contained in Kahneman, Rosenfield, Gandhi, and Blaser, "Noise: How to overcome the high, hidden cost of inconsistent decision making."
91 Mandel, "Are risk assessments of a terrorist attack coherent?"
92 Mandel, "Violations of coherence in subjective probability: A representational and assessment processes account."
93 Miller and Ireland, "Intuition in strategic decision making: friend or foe in the fast-paced 21st century?"; Hoffman, Slovic, and Rorer, "An analysis-of-variance model for assessment of configural cue utilization in clinical judgment."; Chang and McCarty, "Evidence on Judgment Involving the Determination of Substantial Authority: Tax Practitioners versus Students."; Meixner and Welker, "Judgment Consensus and Auditor Experience: An Examination of Organizational Relations."
94 Emily H. Ho et al., "Improving the communication of uncertainty in climate science and intelligence analysis," *Behavioral Science & Policy* 1, no. 2 (2015); David Mandel, "Accuracy of Intelligence Forecasts From the Intelligence Consumer's Perspective," *Policy Insights from the Behavioral and Brain Sciences* 2, no. 1 (2015); David R. Mandel and Alan Barnes, "Accuracy of forecasts in strategic intelligence," *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 30 (2014).
95 Spielmann, "I Got Algorithm: Can There Be a Nate Silver in Intelligence?"; Mandel and Barnes, "Accuracy of forecasts in strategic intelligence."; National Research Council, *Intelligence Analysis: Behavioral and Social Scientific Foundations*. These evaluations would complement those recommended by ODNI's James Marchio. Marchio, ""How good is your batting average?" Early IC Efforts To Assess the Accuracy of Estimates."
96 Friedman and Zeckhauser, "Why Assessing Estimative Accuracy is Feasible and Desirable," 31. Marchio, "How good is your batting average? Early IC efforts to evaluate the accuracy of estimates."

97 Spielmann, "I Got Algorithm: Can There Be a Nate Silver in Intelligence?," 527; Mandel, Barnes, and Richards, *A quantitative assessment of the quality of strategic intelligence forecasts*.
98 Johnston, *Analytic culture in the US intelligence community: An ethnographic study*, 112.
99 Harris and Spiker, *Critical Thinking Skills for Intelligence Analysis*, 218.
100 Harris and Spiker, *Critical Thinking Skills for Intelligence Analysis*, 219.
101 Tetlock and Gardner, *Superforecasting: The Art and Science of Prediction*; Rieber, "Intelligence Analysis and Judgmental Calibration."
102 Friedman and Zeckhauser, "Why Assessing Estimative Accuracy is Feasible and Desirable," 27.
103 Johnston, *Analytic culture in the US intelligence community: An ethnographic study*, 108.
104 Griffin and Tversky, "The Weighing of Evidence and the Determinants of Confidence."
105 Dhami, Belton, and Careless, "Critical Review of Analytic Techniques."
106 Gilovich, *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*.
107 Mandel and Tetlock, "Debunking the Myth of Value-Neutral Virginity: Toward Truth in Scientific Advertising."