# "How good is your batting average?" Early IC Efforts To Assess the Accuracy of Estimates

*Jim Marchio*

***"Few things are asked the estimator more often than 'How good is your batting average?' No question could be more legitimate—and none could be harder to answer."***

Sherman Kent was never more blunt or accurate than when he observed in his memoir-history of the Office of National Estimates (ONE): "Few things are asked the estimator more often than 'How good is your batting average?' No question could be more legitimate—and none could be harder to answer."[1] This article explores one of the Intelligence Community's (IC) earliest efforts to assess the accuracy of its estimative judgments. Led by Kent and his deputy, Abbot E. Smith, the IC systematically examined the judgments contained in more than 200 estimates between 1955 and 1962, sharing its findings with IC members in a series of "validity studies." The factors driving the effort, the challenges encountered in executing it, and the findings contained in the validity studies all are of value today as the Office of the Director of National Intelligence (ODNI) and other members of the IC attempt to answer the same question involving the accuracy of their analysis.[a]

This IC "experiment" conducted six decades ago reminds us how difficult it is to determine "batting averages" as well as the importance of doing so, especially if the IC is to learn from its errors and improve its estimative accuracy.

a. The author acknowledges the valuable comments of John Botzenhart on an early draft of this article.

## *Origins of the Initiative*

Exploring ways to improve the quality and, in turn, the accuracy of the intelligence analysis given to US leaders began with the origins of the IC. However, the perceived intelligence failure associated with the outbreak of hostilities in Korea in 1950 spurred new efforts.[2] In 1952, a production program for national intelligence estimates was initiated. This program provided "for a reexamination of existing estimates on critical areas or problems as well as the production of new estimates designed to improve the coverage of important topics." The program continued and expanded the practice of producing "postmortems," assessments designed "to reveal deficiencies in the preparation of selected estimates and to stimulate corrective action."[b] The

b. The term "postmortem" has been used in different ways in the IC's history. Initially, it was used to denote a product that identified shortcomings in collection and analytic research on an issue on which an estimate had been completed. As Sherman Kent noted: "In the early 1950s we initiated an exercise—collateral to the main task of the ONE—which, however laudable, became a major pain in the neck. This was the *ex post facto* examination of important estimates with an idea of identifying the most significant gaps in our knowledge. Almost from the start it was called a "postmortem." See "The Making of an NIE," *Sherman Kent and the Board of Estimates: Collected Essays* (Center for the Study of Intelligence), 25.

***The failure of post-mortems to address the validity of judgments in estimates likely prompted a new initiative to do so.***

National Security Council (NSC) report describing the effort noted that "the experience of past months in this procedure, particularly as applied in the case of estimates on the Far East, indicates that the results are beneficial."[3]

The number of postmortems completed by the Office of National Estimates increased each year.[4] By the end of 1954 well over 50 postmortems had been completed on National Intelligence Estimates (NIEs) and Special National Intelligence Estimates (SNIEs).[5] For each NIE or SNIE, postmortems identified "areas in which intelligence information is inadequate due either to gaps in collection or in research and analysis." A November 1954 report summarizing postmortem production on NIEs published between January and June of that year stated: "The most important intelligence deficiency in the Soviet Bloc is one of collection . . . in most other areas . . . the overall coverage is good. . . . The problem here is largely one of research and analysis rather than collection."[6] However, no attempt was made in the postmortems to "deal with the validity of substantive judgments made in the estimates."[7]

### *Validity Studies*

The failure of postmortems to address the validity of judgments in estimates likely prompted a new initiative to do so. Following a discussion of the "Postmortem of NIE Production" at a 26 April 1955 meeting of the Intelligence Advisory Committee (IAC), the chairman proposed that a new procedure be adopted to provide for two kinds of reviews subsequent to the completion of an estimate.[a] The first type would be "an immediate postmortem on each estimate to record the intelligence deficiencies encountered by the estimators in its preparation" and would be "prepared by the interagency group that wrote the estimate." The second would be a "review of each estimate after the lapse of an appropriate interval (usually within six months to a year) to study the validity of the estimate, i.e., how good the estimate was in the light of subsequent developments." The proposed initiative was approved and procedures for "validity studies" were drafted over the next few months.[8]

The new IAC postproduction review procedures for NIEs and SNIEs were advanced for review and approval in September 1955. Beyond clarifying and codifying postmortem actions, the draft document established "validation" procedures. "Whenever an estimate is revised," it noted, "the contributing agencies will be requested to submit a critique of the previous estimate together with their regular contribution. These critiques will be consolidated by the Board of National Estimates and coordinated with the IAC representatives." Validation studies also could be "undertaken at any time upon the initiative of the Board of National Estimates or at the request of any one of the IAC agencies." This clause was added to address instances in which estimates were revised only infrequently.[9]

Over the next seven years, nearly 150 validity studies were completed and submitted to the IAC. As planned, the studies examined most of the NIEs and SNIEs published during these years. Four validity studies were produced in 1955. This number jumped to 26 in 1956 and peaked at 28 in 1957. For the remaining years of the program, an average of 16 validity studies were completed annually.

The span of issues and geographic regions covered in the NIEs and SNIEs and, in turn, the validity studies was wide-ranging. Although the greatest number focused on the Soviet Union, its satellites, their military capabilities, and potential courses of action, multiple NIEs and SNIEs addressed the outlook for political stability and economic prospects in nearly every region in the world. Intelligence assessments on the key international crises of the period—Hungary, Suez, and Taiwan—also were assessed for their validity.[10]

Two special validity studies also were completed in the latter years of the program. These examined multiple estimates involving military, political, and economic issues on one country over an 8-to-10-year time period. One, identified as the first of its kind, was intended to be a "validity study of all National Intelligence Estimates [more than 80] concerning the USSR which were produced by the machinery [Office of National Estimates] as presently constitut-

a. The Intelligence Advisory Committee, later renamed the United States Intelligence Board (USIB), was the predecessor of today's National Intelligence Board. It was composed of the heads of IC agencies.

ed, from its beginning late in 1950 through 1957."[11] The other, completed in 1961, reviewed all the estimates produced on India in the preceding decade.[12]

### What Was Their Batting Average?

Most validity studies produced for the IAC contained a one-to-two-page summary of findings. These findings did not contain numbers or percentages to reflect how many judgments were assessed to be valid or inaccurate. Moreover, the methodology used to determine whether a judgment was valid or inaccurate is unclear. My archival research to date in declassified sources has yet to uncover a standardized approach or universal criteria used by IAC members or the ONE to make such assessments. However, it likely involved, per the 1955 guidance, evaluating the judgments "in the light of subsequent developments."

Validity studies generally conveyed their findings in general terms, with assessments falling into one of three categories:

- Judgments were or remain valid
- Judgments were flawed or inaccurate
- Unable to determine validity at this time.

Some variation in how evaluations were conveyed was evident in each category. For instance, a number of studies noted "judgments remain valid but are in need of updating" in light of recent developments.[13] In other cases, judgments were assessed as "partially correct," "partly in error," or caveated in some way to reflect estimative successes or shortcomings.[14] The validity study completed on NIE 13-58, *Communist China*, was indicative of such an approach when it reported to the United States Intelligence Board that "as of mid-1959, most of its judgments for the period through 1962 appear to be still valid except for the predictions of economic growth, which now seem clearly to have been too conservative."[15]

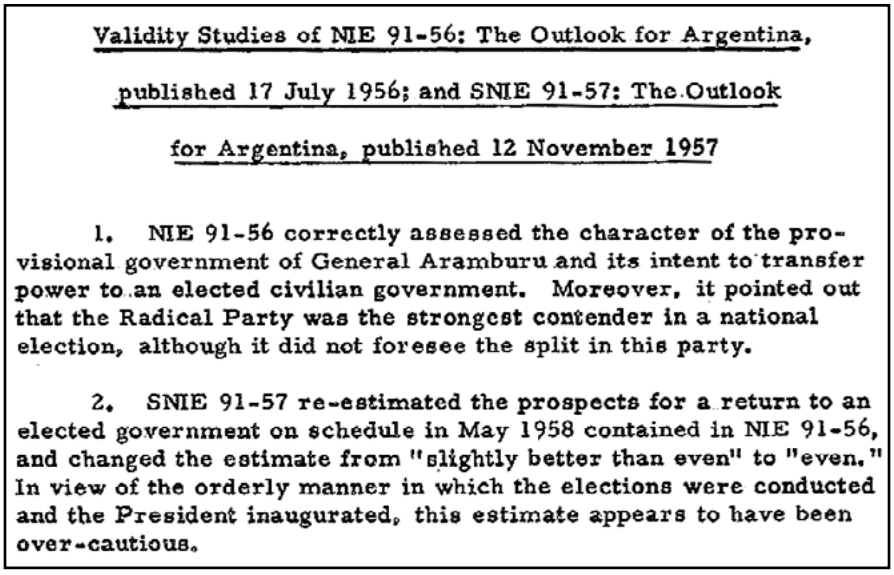
***Most validity studies produced for the IAC contained a one-to-two page summary of findings. These findings did not contain numbers or percentages to reflect how many judgments were assessed to be valid or inaccurate.***

### On Target

The majority of validity studies concluded that their primary conclusions and estimates had proved to be "valid," "generally accurate," "substantially correct," or had been borne out by developments during the period of estimate.[16] As noted in the 1958 validity study examining national estimates on the USSR, "There were hundreds of judgments in these papers, and by far the greater number of them were sound."[17]

Even in cases where individual judgments missed the mark, the key findings of the estimate were often considered valid. The validity study of NIE 11-6-56 (*Capabilities and Trends of Soviet Science and Technology*) observed, "It should be remarked, however, that these are specific developments of a kind which intelligence does not expect to predict, and failure to do so in no way affected the validity of the main estimates in this paper."[18]

Individual validity studies occasionally described factors that contributed to their accuracy. For example, the validity study on NIE 27-1-56 (*Probable Developments in*



Validity Studies of NIE 91-56: The Outlook for Argentina, published 17 July 1956; and SNIE 91-57: The Outlook for Argentina, published 12 November 1957

1. NIE 91-56 correctly assessed the character of the provisional government of General Aramburu and its intent to transfer power to an elected civilian government. Moreover, it pointed out that the Radical Party was the strongest contender in a national election, although it did not foresee the split in this party.

2. SNIE 91-57 re-estimated the prospects for a return to an elected government on schedule in May 1958 contained in NIE 91-56, and changed the estimate from "slightly better than even" to "even." In view of the orderly manner in which the elections were conducted and the President inaugurated, this estimate appears to have been over-cautious.

An illustrative summary page of an IAC Validity Study. Originally classified Secret; approved for release August 2006.

***Validity studies in some instances delved into greater detail into the sources underlying estimative errors.***

*Spain*) noted that it had proved to be "substantially correct" and that "in some particulars it anticipated trends still developing at this moment, such as the continuing of labor unrest." This forward-looking focus may also have contributed to the NIE correctly calculating that "Franco could retain power, and that oppositionist forces, although increasing in restlessness, would probably remain weak."[19] In the case of an NIE on Burma, the validity study praised the estimate for emphasizing "the dangers inherent in the situation" while avoiding going "overboard." The validity study also cited the NIE for its treatment of Burma's actions and identification of key drivers in the short as well as long run.[20]

### *Off the Mark*

Validity studies, to their credit, were just as quick to acknowledge and identify how specific judgments or assessments were off the mark as well as the reasons why. In fact, the greatest amount of time and effort in validity studies was spent in exploring where and how assessments went awry.

The estimative shortcomings identified in validity studies often involved errors of emphasis (overly cautious, underestimated, or overly emphasized) or omission (the failure to foresee or anticipate certain developments, identify key factors or drivers). The validity study on a 1954 NIE addressing probable developments in Argentina, for example, concluded, "It overestimated Peron's ability, through the policy of moderation followed after 1952, to repair army loyalty shaken by the activities of Eva Peron prior to her death in July of that year. NIE 91-54 also failed to give adequate weight to the intentions and political determinations of the Argentine armed forces, especially the navy."[21] A 1958 validity study on the estimate *Sino-Soviet Foreign Economic Policies and Their Probable Effects in Underdeveloped Areas* identified similar shortcomings: "We now believe that NIE 100-57 overestimated the extent to which competing internal demands would restrict expansion of the Bloc foreign economic program. Moreover, it did not foresee the number of opportunities which would develop in the Free World."[22]

Validity studies in some instances delved into the sources underlying estimative errors. For example, an August 1957 validity study of NIE 71.2-56, *Outlook for Algeria*, concluded the NIE "has proved incorrect in its most important estimate: that there was a somewhat better than even chance for an Algerian settlement within a 12 month period." It then went on to identify the main causes for the miscalculation:

> *a) An overestimate of France's willingness to face the realities of the Algerian situation,*
>
> *b) A failure to estimate the Mollet government's adoption of an increasingly rightist policy toward Algeria, and*
>
> *c) The unforeseen armed intervention at Suez and the subsequent intensely nationalistic French reaction.*[23]

### *Can't Tell*

The validity studies produced between 1955 and 1962 also highlighted the challenges in assessing accuracy. In some cases, it was the insufficient passage of time or the long-range nature of the issue. For example, the 1957 validity study on NIE 100-5-55, *Implications of Growing Nuclear Capabilities for the Communist Bloc and the Free World*, published in June 1955, concluded: "Many of its judgments involved long-term attitudinal trends which cannot yet be measured or checked with any preciseness and with contingent situations that have not yet arisen."[24] Similar comments were advanced regarding SNIE 12-3-56, *Probable Developments in Soviet-Satellite Relations*: "Insufficient time has passed to permit an assessment of the validity of this estimate."[25]

In other instances it was the lack of data or a rapidly changing environment that prevented an assessment. As the validity study addressing NIEs on Soviet guided missile capabilities concluded, "The amount of evidence obtained has been meager. It tends to strengthen the previous estimates, but does not permit an evaluation of their validity."[26] In a similar vein, the study *Probable Developments in Argentina* noted, "conclusion as to the validity of our estimate that any successor government to Peron would probably follow the same general internal and external policies must be reserved pending political stabilization in Argentina."[27] Thus it is not surprising that the USSR validity study reminded readers: "The words 'right, correct, accurate,' and so on, when applied to our estimates, must still be taken in a provisional sense.

Only in a comparatively small number of instances can we be perfectly sure that we were 'right.'"[28]

The possibility that US actions initiated in response to intelligence provided was another factor identified as affecting and complicating decisions on the accuracy of NIE and SNIE judgments. The validity study of NIE 93-55, for instance, pointed out that "partly as a result of army influence in the present regime and partly because of the US decision to provide substantial economic assistance to Brazil, a moderate political course, rather than further evolution to the left, as suggested in [the] NIE, has thus far prevailed."[29] Likewise, the lengthy validity study on Indian assessments observed that "predictions may have been good when they were made, but the event forecast did not occur because of a sharp change in US policy made after—or perhaps even because of—an NIE."[30]

In hindsight, those assessing the accuracy of their estimates in validity studies believed performance depended in part on the subject area. Sherman Kent observed: "We did find ourselves in a number of significant good and bad estimates, especially in those matters which involved quantifiable things like estimated growth in GNP, probable dates of initial operational capability of a new weapons system, etc. We were a lot less successful in our evaluations of our estimates of less tangible things."[31]

### *Reflections on the Record*

Several of the longer validity studies were noteworthy for their attempts to garner lessons learned by stepping back and considering the accuracy of estimates produced over extended periods. Both the 1958 USSR and 1961 India validity studies did so, identifying the most serious estimative errors, greatest successes, and factors contributing to each in order to improve future analysis.

***Several of the longer validity studies were noteworthy for their efforts to step back and consider the accuracy of estimates produced over extended periods.***

The USSR validity study was particularly forthcoming and valuable in addressing what its author considered "three truly serious" errors:

> *1. We wholly failed to foresee, and for a long time we even failed adequately to recognize and describe, the changes in the character and conduct of Soviet policy—especially foreign policy—that occurred after the death of Stalin.*
>
> *2. We failed to foresee the upheavals in the European Satellites that occurred late in 1956 or even to hint that such upheavals were possible.*
>
> *3. We failed to foresee Soviet intervention in the Middle East in late 1955.*[32]

The author then went on to explore what he considered the root cause of these errors, observing, "One phenomenon strikes me quite forcibly—it is the degree to which our most important wrong estimates, all of which were in the political field, arose out of resistance to the idea that change and development would occur in the Soviet Bloc."[33] Although the author ultimately concluded that he did not discern "persistent or recurring tendencies which have led us into error on repeated occasions and which are susceptible to correction," he reiterated the need to address "our disinclination to foresee or to recognize change."[34]

Certainly the estimative shortcomings identified in the other 156 validity studies completed during this period corroborate to a degree Abbot Smith's observations. In many cases the judgments deemed inaccurate were attributed to a failure to address and properly assess the strength of nationalism and popular unrest, two key drivers of change during the late 1950s and early 1960s.

Beyond the USSR and Indian efforts, the IAC validity studies also spurred other detailed assessments of IC analytic processes and estimates. A 1963 exchange in *Studies in Intelligence* over intelligence estimates on China's economy is one such example. The initial piece, billed as a "postmortem," presented "lessons derived from analysis of errors past" and explored how and why Western intelligence had been "so awry" in its estimates of communist China's economic strength. A rebuttal published several months later specifically noted procedures institutionalized in "validity studies" and asserted that "if the purpose of the postmortem is to learn the lessons of experience, the record should be read straight." The rebuttal went on to address at length the supposed errors and the analytic tradecraft used in estimative process.[35]

***Although the exact date the IAC ended its validity studies program is unclear, it was probably late in 1962.***

### *End of an Era*

Although the exact date the IAC ended its validity studies program is unclear, it was probably late in 1962. The last validity study I have discovered was completed in June 1962,[36] and there is no mention of such studies in an ONE activities report for first half of 1961.[37] It is likely validation studies were done on an ad hoc basis thereafter.[38]

A memo laying out the specific reasons for the discontinuation of the validity studies has yet to be declassified. Sherman Kent later wrote, "We in ONE were dismayed at our failure to do a more convincing job of the validity studies and much relieved when the IAC let the enterprise peter out."[a] However, Smith almost certainly explained the rationale for ending the effort in a 1969 article "On the Accuracy of National Estimates," which was originally classified Secret and published in 1969 in *Studies in Intelligence*.[b, 39] Smith laid out multiple reasons in the article why a "complete, objective, and statistical tally would not be worth doing."[40] He divided these reasons into two categories. The first involved the difficulty of checking accuracy; the second concerned the value of the results from these efforts.

Smith identified the sheer number of estimates contained in NIEs and SNIEs as one factor hindering an evaluation of accuracy, noting in his article that approximately 25,000 judgments would need to be assessed.[41] Beyond the number, Smith pointed to the difficulty of evaluating restricted or conditional judgments (if/then); judgments contained in less prominent locations, e.g., subordinate clauses or in the middle of an estimate; and judgments caveated with "estimative formulations" (probably, unlikely, etc.).[42]

Smith also stressed the importance, and the difficulty, of determining the impact of the context surrounding judgments. "The validity of such papers," Smith noted, "depends only partly upon the accuracy of each particular statement in them. It must also be judged by the impact and tone of the document as a whole."[43] Finally, Smith cited the lack of data in many cases to check or verify judgments as well as the challenges in ascertaining what impact US actions (action/reaction) may have had on the accuracy of an estimate.[44]

Even when possible, Smith questioned the value of the results derived from such accuracy assessments. In some cases, he asserted, the results were dubious because of changes in the environment, context, or even the methodology used in generating the estimate. Another element was that not all judgments were of the same importance. Many of them were "simply too easy" and thus a "batting average, if it were arrived at, would be worth about as much as the batting average of a major league team playing against a scrub outfit in a sandlot."[45] In sum, Smith argued "a complete, objective, statistical audit of the validity of NIE's is impossible, and even if it were possible it would provide no just verdict on how 'good' these papers have been."[46]

Although Abbot Smith's 1969 article certainly provides the most comprehensive discussion of challenges in assessing accuracy, many of same arguments are found years earlier in validity studies done on the USSR and India.[47] Foreshadowing what he would write a decade later, Smith began the 1958 Top Secret USSR validity study with this observation:

> *In theory the making of a validity study should be a simple matter—get out the old papers, read them, and note whether the estimates turned out to be true or false. In practice it is not that simple. Indeed it is so much more complicated and difficult that it has proved in many respects to be impossible, and this study has turned out quite differently from what its author had hoped it would.*[48]

Smith went on to identify in the validity study's introduction the same challenges in assessing accuracy that he surfaced in the *Studies in Intelligence* article.[49] Interestingly, while the authorship of the May 1961 validity study on India is unknown at this time, the report's discussion of the obstacles encountered in attempting to evaluate the accuracy of 10 years of NIEs on India mirrored the 1958 conclusions of the USSR validity study. Like the earlier work, the Indian validity study stressed the

a. Kent agreed with Smith's assessment of the challenges involved in evaluating accuracy. He quoted extensively from Smith's 1969 article in his memoir/history of NIEs and the ONE, concluding: "I join Mr. Smith in his regrets that we can do no better for the outsider in search of a box score." ("The Making of an NIE," 35.)

b. Abbot Smith worked with Sherman Kent as his deputy for 14 years (1953–67) in the Office of National Estimates.

multiple factors that made determining an estimate's accuracy difficult and undermined—to a degree—the value of the findings.[50]

### *Intermittent Efforts To Assess Accuracy in the Years Since*

Despite the IAC's experience with validity studies and Smith's pessimistic 1969 article, efforts to ascertain the IC's "batting average" persisted. In 1972, an in-depth study was conducted at the request of the Director of Central Intelligence Richard Helms to examine the most important NIEs and SNIEs produced in 1967, with the idea that the passage of time would aid in assessing accuracy. The study, similar in many ways to the special validity studies completed on the USSR and India over a decade earlier, looked at the estimative record for multiple topics ranging from Vietnam and Soviet military forces to China, Latin America, and Africa.[53]

Like these earlier validity studies, "1967's Estimative Record – Five Years Later" was frank in acknowledging the shortcomings and strengths of estimates produced during that year. Although the report did not include overall accuracy numbers for the estimative judgments advanced, it did provide this general assessment:

> *Broad general judgments about future capabilities and courses of action have generally held up well; such judgments are based on a broad range of considerations, not often subject to change through the appearance of specific new data. Judgments about specific capabilities existing in 1967 have also stood the test of time; they usually had hard evidence to support them, but sometimes did not. Predictions of specific future capabilities and force levels are a more chancy business; estimates in this category were sometimes right on the mark, but sometimes wide of it.*[54]

The study ended with a section that delved into broader analytic issues, including the value of the exercise itself and the challenges in evaluating accuracy:

> *If it is not fair to judge an estimate by success or failure in predictions of discrete events, it is certainly legitimate to ask whether it identified and interpreted the major forces at work in a situation. If it failed to do this, it is a poor job by any standards. A review of 1967 does not turn up any serious deficiencies on this score.*[a, 55]

Later in the decade another exercise was conducted to determine the accuracy of judgments in national-level estimates. Although the full details of the effort have yet to be declassified, a larger report addressing the overall quality of national intelligence estimates described the difficulty in determining the accuracy of judgments contained in political estimates, noting that for one year a running box score was kept on the forecasting ability of these estimates. The result proved "futile," with 50 percent of the events never resolved. Moreover, in a substantial number of the remaining 50 percent, predicted outcomes happened "but not quite in the way described in the estimates" or involved tautological judgments such as "the sun will rise tomorrow."[56] Other efforts during the 1970s to identify and assess the "track record" of national intelligence estimates took a more holistic approach, eschewing percentages of right and

#### *Postmortems: Similar Fate, Later Reincarnation*

The IC produced "postmortems" addressing collection and analytic gaps during and after the IAC's validity studies program ceased. However, like the validity studies, the postmortem program was scaled back. USIB guidelines approved in June 1964 directed that rather than being published with each estimate, postmortems should be produced

> *selectively; that is, when intelligence gaps or deficiencies are encountered which are sufficiently serious to affect the quality and completeness of national intelligence on important topics.*[51]

An IC postmortem program was reincarnated a decade later. Yet this effort was different in focus and purpose from its predecessor. In some ways the program combined elements of earlier postmortems that focused on collection shortcomings with the emphasis on analytic judgments found in validity studies. The end result was an assessment of the IC's overall performance on an issue or in response to a crisis.[52]

a. "1967's Estimative Record—Five Years Later" also may have been written by Abbot Smith. The challenges identified in assessing accuracy as well as the language used is very similar to that employed in the 1958 USSR and 1961 Indian validity studies.

wrong for general conclusions about the community's performance.[57]

Accuracy continued to be considered during the 1980s as part of a larger effort to evaluate the quality of finished intelligence. Helene Boatner, chief of CIA's Product Evaluation Staff, acknowledged in a 1984 *Studies in Intelligence* article that in judging the quality of analysis, a number of factors had to be considered:

> *Accuracy (on both facts and judgments) is one key ingredient. . . . How right or how wrong we can expect to be varies a lot by topic. . . . The accuracy of our assessments also depends on whether relationships between the facts we have and the ones we lack are fixed (physics), generally predictable within some range (economics), or highly irregular (politics). The more human decisions affect the relations between the known and unknown facts, the harder it is for an analyst to assess the present, to say nothing of predicting the future.*[58]

Boatner concluded her discussion by singling out some of the same challenges in assessing accuracy identified by Smith and previous validity studies including the "problem of action and reaction," specifically citing the issue of the accuracy of estimates of Soviet strategic weapons deployments over time. While acknowledging that mistakes had been made, she opined that the political impact of intelligence judgments "may well have had a major impact on weapons trends," with the "missile gap" controversy of the late 1950s leading to a major US defense buildup that spurred the Soviets to respond by accelerating and expanding programs already under way.[59]

## *The Post-IRTPA Environment*

The passage of the 2004 Intelligence Reform and Terrorism Prevention Act (IRTPA) and the findings of the Weapons of Mass Destruction (WMD) Commission gave renewed impetus to evaluation of the quality of intelligence analysis and to efforts to improve its accuracy. Intelligence Community Directive (ICD) 203 (Analytic Standards), signed by the Director of National Intelligence in 2007, included accuracy as its eighth tradecraft standard. ICD 203 directed analysts to "apply expertise and logic to make the most accurate judgments and assessments possible" while acknowledging "accuracy is sometimes difficult to establish and can only be evaluated retrospectively if necessary information is collected and available."[60] Other IC attempts to evaluate the accuracy of their estimates have occurred since 2004. As former CIA Acting Director Michael Morell recently noted: "One of things that most people don't know is that the Agency actually tracks how well its judgments stand up over time. And the numbers look like fielding percentages in baseball, not batting averages."[61]

## *Lessons for Today*

The need for accuracy in the intelligence assessments provided to our nation's leaders certainly has not declined in recent years. As then-CIA Director Michael Hayden remarked in 2006, "With regard to analysis, it's real simple; it's just 'getting it right' more often." The 2011 Arab Spring, the rise and success of the Islamic State of Iraq and the Levant (ISIL), Russian actions in the Ukraine and Syria as well as the terrorist attacks in Europe and in the US homeland in 2015 and 2016 all reinforce Hayden's comments.[62]

The continuing requirement for accurate intelligence has spawned new efforts from outside and within the IC to determine its "batting average." The research of multiple scholars suggests that many of the challenges associated with assessing accuracy Abbot Smith identified in 1969 can be overcome or at least mitigated, producing an outcome beneficial to IC consumers and analysts.[a] Within the IC, the ODNI has devoted more resources in the last three years to assess accuracy. This effort, unlike

a. Recent studies calling for and highlighting the feasibility of assessing accuracy include: Jeffrey A. Friedman and Richard Zeckhauser, "Why Assessing Estimative Accuracy is Feasible and Desirable," *Intelligence & National Security*, 28 November 2014, at http://dx.doi.org/10.1080/02684527.2014.980534; Welton Chang, "Getting It Right: Assessing the Intelligence Community's Analytic Performance," *American Intelligence Journal*, Vol. 30, No. 2 (December 2012): 99–108; David R. Mandel and Alan Barnes, "Accuracy of forecasts in strategic intelligence," *PNAS Early Edition*, www.pnas.org/cgi/doi/10.1073/pnas.1406138111; David R. Mandel, "How good are strategic intelligence forecasts?" 25 Sep 2014, http://policyoptions.irpp.org/2014/09/25/how-good-are-strategic-intelligence-forecasts/; Philip E. Tetlock and Barbara A. Mellers, "Structuring Accountability Systems in Organizations: Key Tradeoffs and Crucial Unknowns," in *Intelligence Analysis: Behavioral and Social Scientific Foundations* (National Research Council, The National Academies Press, 2011), 249–70.

past ad hoc examinations, represents a systematic evaluation of the accuracy of the key judgments contained in the products it evaluates as part of its annual tradecraft review mandated by the IRTPA.[63]

What lessons then does the IC's experience nearly six decades ago in attempting to assess the accuracy of its estimates offer for us today? What insights do the hundred-plus validity studies provide into determining the IC's "batting average" and the value of doing so?

These studies, much like Smith's 1969 article, remind us of the obstacles the IC will face as it tries to do more in determining the IC's and individual analyst's batting averages. Many of the same challenges that complicated or prevented the completion of these validity studies—lack of data, imprecise estimative language, conditional judgments, and action-reaction scenarios—have not disappeared with the passage of time. Moreover, some of the challenges identified in the validity studies have become more acute in recent years with the emergence of "big data" and methodological changes that have accompanied the digital revolution. These changes tend to complicate any analysis of the track record that seeks to use common yardsticks for reviewing estimates over a period of years. The same is true for changes in intelligence collection and analytic capabilities.

These validity studies also offer valuable cautions as to what renewed efforts to assess accuracy should avoid or, conversely, incorporate. One caution, just as relevant today as it was in 1958, is to concentrate accuracy assessments on key judgments. In essence, it is not the overall batting average that matters most but the IC's average—to use another baseball analogy—with "runners in scoring position." Another is to avoid focusing solely on individual judgments. It is important to keep the context of the entire assessment in mind.

***These studies highlight how critical it is to go beyond just determining and comparing batting averages and examine the reasons why judgments were off the mark or on target.***

A 1980 report on national intelligence estimates captured this issue with alacrity:

> *Postmortems of estimates whose original purpose was to undertake some kind of prediction do not help the policymaker. Such an evaluation will show only that the predicted event did or did not happen. Most policymakers already have some chosen objective in mind. What they most want to know from the estimate are the elements in the situation which would make the desired outcome more probable.*[64]

A third caution involves the issues or topics evaluated and the credibility of the results. Accuracy evaluations for some areas are more illuminating and calibrated than others. As highlighted in these historical studies, there was generally a better correlation between accuracy and more quantitative analysis than with political assessments.[65] The same was true for estimates with shorter time frames (two to three years) vice three to 10 years in the future.[66]

Finally, these studies highlight how critical it is to go beyond just determining and comparing batting averages and examine the reasons judgments were off the mark or on target. The review of judgments for validity forced analysts and their managers to reexamine assumptions and conduct essentially an analytic line review—both tradecraft best practices by today's standards. The same is true of the step-back assessments conducted in the USSR and India validity studies and the *Studies in Intelligence* exchange over Chinese economic estimates. Indeed, as noted earlier, only by doing so will it be possible to identify "persistent or recurring tendencies which have led us into error on repeated occasions and which are susceptible to correction."

There is little doubt that the IC should continue its efforts to assess accuracy, including piloting approaches now being used in prediction markets and elsewhere. At the same time, the community must remember that the most important aspect of assessing accuracy goes beyond the numbers. Abbot Smith, not surprisingly, summed it up best: "A validity study should be a vehicle of improvement, not merely of congratulation and abuse."[67] It is critical that the IC not forget this valuable lesson as it attempts again to answer the batting average question Sherman Kent and his colleagues were asked more than 60 years ago.

❖ ❖ ❖

## *ENDNOTES*

1. Donald P. Steury (ed.), "The Law and Custom of the National Intelligence Estimate" in *Sherman Kent and the Board of National Estimates: Collected Essays*, Center for the Study of Intelligence (CSI), 1994), 100; available at https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/toc.html.
2. For discussion of the impetus to IC reform efforts in the wake of the Korean War, see Michael Warner, J. Kenneth McDonald, *US Intelligence Community Reform Studies Since 1947* (CSI, April 2005), iii.
3. *Foreign Relations of the United States (FRUS), 1950–1955, The Intelligence Community* (Government Printing Office, 2007), 410–11.
4. "List of NIE Studies," undated. This document can be found via the CIA Records Search Tool (CREST). CREST is available at http://www.foia.cia.gov/collection/crest-25-year-program-archive. However, many documents are not available on-line and must be viewed at the National Archives and Records Administration (NARA) II in College Park, Maryland. Documents located in the CREST database are henceforth referenced by their Agency Action Identifier, followed by the box, folder, and document number: CIA-RDP85S00362R000600010001-3. Also see "Contents," no date, CIA-RDP85S00362R000500030001-2 for a listing of three NIEs on which postmortems were done in 1952, all of which tied into Asia and China-related topics.
5. "Contents," 173–80, and 187–98.
6. Ibid., 188.
7. Ibid., 3 May 1955, 173.
8. Intelligence Advisory Committee (IAC), Minutes of 26 April 1955 Meeting, CIA-RDP85S00362R000200060029-2.
9. "Contents," 140–43.
10. For crisis validity studies, see SNIE 12-3-56: *Probable Developments in Soviet-Satellite Relations*, SNIE 36.7-56: *Outlook for Syrian Situation*, NIE 36.1-55: *The Outlook for Egyptian Stability and Foreign Policy*, NIE 92: *Israel and Other Important Estimates on Israel in Estimates Prepared Since April 1956*; NIE 100-4-55: *Communist Capabilities and Intentions with Respect to the Offshore Islands and Taiwan* in "List of Validity Studies," 21–22, 43–44, 47–48, 81–82, & 123–24.
11. *A Study of National Intelligence Estimates on the USSR, 1950–57*, CIA-RDP79R00971A000300050001-8.
12. *Validity Study of the National Intelligence Estimates on India, 1951–60*, 31 May 1961, CIA-RDP79R00904A000700030026-6; and "Memo on Validity Study of NIEs on India, 1951–1960," 19 June 1961, CIA-RDP79R00904A000700030025-7.
13. "List of NIE Studies," 12, 92, 150.
14. Ibid., see p.18 for an example of partially correct judgment: "Moreover, it points out that the Radical Party was the strongest contender in a national election, although it did not foresee the split in this party." Also see 52, 155, and Validity Study of NIE 24-56, 6 November 1958, CIA-RDP82M00097R000600020046-4.
15. *Validity Study of NIE 13-58*, 29 July 1959, CIA-RDP82M00097R000600020033-8.
16. For an example, see *Validity Study of NIE 32-56*, 2 October 1958, CIA-DP82M00097R000600020048-2.
17. *A Study of National Intelligence Estimates on the USSR, 1950–57*, 25; *Validity Study of the National Intelligence Estimates on India, 1951–60*, 31 May 1961, 18: "Judged as objectively as is possible, it would appear that the intelligence community's record in estimative developments in India during the last decade is a good one."
18. *Validity Study of NIE 11-6-56*, 19 August 1959, CIA-RDP82M00097R000600020030-1.
19. "List of NIE Studies," 20–21.
20. *Validity Study of NIE 61-56*, 18 April 1960, CIA-RDP82M00097R000600020022-0. For other examples, see "List of NIE Studies," 30, 48: "SNIE 36.7-56 was produced during the Suez crisis last fall as a short-term paper. The estimate proved to be accurate in all major respects and most of its judgments are still valid today."
21. "List of NIE Studies,"109.
22. Ibid., 16. For other examples of overestimating and anticipation errors see 22, 31, 40, 52, 76, 88, and 103.
23. *Validity Study on NIE 71.2-56: Outlook for Algeria*, "List of NIE Studies," 57–58, 13 Aug 1957.
24. "List of NIE Studies," 55–56.
25. Ibid., 81–82.
26. Ibid., 159; also see 156–60 addressing NIE 11-6-54: *Soviet Capabilities and Probable Programs in the Guided Missile Field*, 5 Oct 1954 and its supplement NIE 11-12-55, *Soviet Guided Missile Capabilities and Probable Programs*, 20 Dec 1955.
27. Ibid., 109.
28. *A Study of National Intelligence Estimates on the USSR, 1950–57*, 23.
29. NIE 93-55: *Probable Developments in Brazil*, 15 Mar 1955, in "List of NIE Studies," 90, 8 Jan1957; see also NIE 70: *Conditions and Trends in Latin America Affecting US Security*, 12 Dec 1952, 147.
30. *Validity Study of the National Intelligence Estimates on India, 1951–60*, 1–2; also *A Study of National Intelligence Estimates on the USSR, 1950–57*, and *1967's Estimative Record—Five Years Later*, 16 August 1972, CIA-RDP79R00967A001500040010-1, 4.
31. Steury, *Sherman Kent and the Board of National Estimates*, 27.
32. *A Study of National Intelligence Estimates on the USSR, 1950–57*, 20–21.

33. Ibid., 25.
34. Ibid., 131–32.
35. "Post Mortem: The Chinese Economy," *Studies in Intelligence* 7, no.1 (1963): Approved for Public Release: 2014/07/29 C061830; and Edward L. Allen, "Chinese Growth Estimates Revisited: A Critique," *Studies in Intelligence*.7, no.2 (Spring, 1963), 1–12, CIA Historical Review Program Release as Sanitized, 18 September 1995. Also see *Validity Study of NIE 13-58*, 29 July 1959, CIA-RDP82M00097R000600020033-8 and *Validity Study of Economic Sections of NIE 13-60*, 26 April 1961, CIA-RDP-82M00097R000600020005-9.
36. *Validity Study of NIEs 51–60*, 53–59, and 36.2-60, 1 June 1962, CIA-RDP82M00097R000600020001-3.
37. "Principal Activities of the Office of National Estimates in Last Half Year," 24 October 1961, CIA-RDP80B01676R003400010001-5.
38. Memorandum from the Deputy to the Director of Central Intelligence Programs Evaluation (Bross) to Director of Central Intelligence Rayborn, 20 January 1966: "The Board of National Estimates reviews National Intelligence Estimates from time to time to ascertain retrospectively the validity of these estimates. These 'Postmortems' identify gaps which appear to have existed in the information available during the formulation of particular estimates." *FRUS, 1964–1968*, *Volume XXXIII, Organization and Management of Foreign Policy; United Nations* (GPO, 2004) available at https://history.state.gov/historicaldocuments/frus1964-68v33/d242.
39. See "Abbot Emerson Smith," *Studies in Intelligence* 27 (Winter 1983), Approved for Release 2014/07/29 C00619199.
40. Abbot E. Smith, "On the Accuracy of National Intelligence Estimates," *Studies in Intelligence* 13, no.4, (1969) CIA-RD-P79R00971A000400030001-9, 26.
41. Ibid., 25.
42. Ibid., 26.
43. Ibid., 29.
44. Ibid., 26–27.
45. Ibid., 26.
46. Ibid., 30.
47. *A Study of National Intelligence Estimates on the USSR, 1950–57*; *Validity Study of the National Intelligence Estimates on India*, 31 May 1961.
48. *A Study of National Intelligence Estimates on the USSR, 1950–57*, 1.
49. Ibid., 1–5.
50. *Validity Study of the National Intelligence Estimates on India, 1951–60*.
51. "USIB Action on Post-Mortems to National Intelligence Estimates," 24 August 1965, CIA-RDP82R00129R000100010015-3.
52. Richard W. Shryock, "The Intelligence Community Post-Mortem Program, 1973–1975," *Studies in Intelligence*.21, no.1 (Fall 1977), CIA-RDP78T03194A000400010015-5.
53. *1967's Estimative Record—Five Years Later*, 16 August 1972.
54. Ibid., 7–8.
55. Ibid., 32.
56. *Seminar on National Intelligence Estimates*, 26 February 1980, CIA-RDP81B00493R000100020010-0, p.12.
57. "The Track Record in Strategic Estimating: An Evaluation of the Strategic National Intelligence Estimates, 1966–1975," in CIA's *Analysis of the Soviet Union, 1947–1991*; see also "DCI Backup Briefing Note," 11 July 1979, NIE Track Record, CIA-RDP-86B00269R001100150001-2.
58. Helene L. Boatner, "The Evaluation of Intelligence," *Studies in Intelligence* 28, no.2 (Summer 1984), 67.
59. Ibid., 70.
60. https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards.pdf.
61. "The Cipher Brief," *Washington Post*, 14 December 2016, https://www.thecipherbrief.com/article/exclusive/fmr-cia-acting-dir-michael-morell-political-equivalent-911-1091.
62. Transcript of General Hayden's Interview with WTOP's J.J. Green, 30 November 2006. https://www.cia.gov/news-information/press-releases-statements/press-release-archive-2006/pr11302006.htm.
63. IRTPA, Section 1019.
64. *Seminar on National Intelligence Estimates*, 26 February 1980, 12.
65. Ibid., "This is not to say that postmortems of political estimates are useless. There are lessons to be learned from hindsight. One study analyzed existing estimates up to the 1960s. Predictions dealing with more quantitative analysis, such as technology or gross weapons capabilities, proved to be adequate. But the threat of political reality in the more general estimates also proved to be surprisingly good: 75–80% right for the world as a whole." (12); Boatner makes the same point in her article.
66. *1967's Estimative Record—Five Years Later*, 16 August 1972.
67. *A Study of National Intelligence Estimates on the USSR*, 1950–57, 131.

❖ ❖ ❖