INTELLIGENCE · ANALYSIS AND TRADECRAFT

Measurement of Analytic Rigor and Quality in Intelligence Products

How can an organisation tell whether its intelligence analysis is any good? This article explains the two main approaches to measuring analytic quality, assessing the process by which analysis is produced and scoring the accuracy of its judgements, and shows how they can be combined in government and business.

ModQu research article · 2,288 words · 22 references (APA 7th edition)

Intelligence organisations produce a great deal of analysis, but they have often found it hard to say how good that analysis is. A report may be well written and well received yet based on weak evidence. A judgement may turn out to be right for the wrong reasons, or wrong despite careful work. Without a way to measure quality, organisations cannot tell whether training, new methods or reorganisation make any difference, and analysts cannot learn which of their habits serve them well.

This article explains the two main approaches to measuring analytic quality. The first assesses the **process**: whether the analysis was carried out in a rigorous way, according to recognised standards. The second assesses **outcomes**: whether the judgements turned out to be accurate. Each has strengths and limits, and the most useful systems combine them.

Why measurement became a priority

Major intelligence failures in the early 2000s prompted a new emphasis on analytic standards. The commission that investigated American intelligence on Iraqi weapons of mass destruction found that analysts had relied on weak sources, failed to state their assumptions and allowed uncertainty to disappear from final judgements (Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction, 2005). It recommended stronger standards and independent review of analytic products.

Marchio (2014), tracing the history of analytic tradecraft in American intelligence, shows that attention to standards had risen and fallen over the decades, often in response to failures, and had rarely been sustained. After 2004, the newly created Office of the Director of National Intelligence made standards formal and lasting. Fingar (2011), who led American analytic production during that period, describes efforts to set common standards across agencies and to evaluate products against them.

A major review by the National Research Council (2011) went further. It argued that intelligence organisations should apply the methods of the behavioural and social sciences to their own work: testing which analytic methods improve accuracy, keeping score of judgements and learning systematically from results. A companion volume collected the underlying research (Fischhoff & Chauvin, 2011).

Why rigour is hard to see

Part of the difficulty in measuring quality is that weak analysis often looks like strong analysis. Heuer (1999) observed that analysts are usually unaware of the mental shortcuts that shape their conclusions, and that a confident, well-organised report can rest on a single unexamined assumption. Readers face the same problem: they see the conclusion and the supporting points, not the alternatives that were never considered or the sources that were never checked. A report that cites many sources may seem thorough even if all of them repeat one original claim.

For this reason, quality cannot be judged from the finished product alone. Reviewers need to know something about how the analysis was produced: what was searched, what alternatives were considered, which assumptions were tested and who challenged the reasoning. Pherson and Heuer (2021) recommend that analysts keep a brief record of these steps, such as the hypotheses considered, the key assumptions and the sources consulted, so that reviewers and later readers can see the work behind the conclusion. Such records make process review possible and also help analysts themselves notice when they have cut corners.

Process standards

ICD 203

The best-known set of standards is Intelligence Community Directive 203, which sets out the analytic standards that American intelligence products must meet (Office of the Director of National Intelligence, 2015a). It requires analysis to be objective, independent of political considerations, timely, based on all available sources and consistent with a set of tradecraft standards. These include properly describing the quality and credibility of sources; expressing and explaining uncertainty; distinguishing between underlying information and the analyst's assumptions and judgements; considering alternative explanations; demonstrating relevance to the customer; using clear and logical argument; explaining changes in judgements over time; making accurate judgements; and using effective visual information.

A related directive, ICD 206, sets requirements for describing sources in analytic products, so that readers can see what evidence supports each key judgement (Office of the Director of National Intelligence, 2015b).

Agencies evaluate samples of finished products against these standards, rating each product on each criterion. The results show where tradecraft is weak and allow trends to be tracked over time.

Attributes of rigour

Researchers in cognitive systems engineering have studied rigour more directly. Patterson et al. (2001) observed professional analysts working on an unfamiliar problem under time pressure, with access to a large collection of documents. They found that analysts often relied on a small number of documents that appeared early in their search, missed key information and reached conclusions without noticing gaps or conflicts. The risk, they concluded, was not obviously poor analysis but **shallow** analysis that looked complete.

Building on this work, Zelik et al. (2010) proposed a set of eight attributes that together describe the rigour of an analysis. Each can be rated as low, moderate or high.

Eight attributes of analytic rigour

Each can be rated low, moderate or high for a given piece of analysis

Attribute	Question to ask
Hypothesis exploration	Were several explanations considered and tested?
Information search	Was the search broad and deep enough for the question?
Information validation	Were key facts checked and corroborated?
Stance analysis	Were the perspectives and motives of sources considered?
Sensitivity analysis	Was it tested how far the conclusion depends on key assumptions?
Specialist collaboration	Were relevant experts consulted?
Information synthesis	Were the pieces brought together into a coherent account?
Explanation critiquing	Was the reasoning reviewed and challenged by others?

Based on Zelik et al. (2010); questions by ModQu

ModQu.com

Figure 1. Eight attributes of analytic rigour. Rating each attribute shows where an analysis is strong and where it may be shallow. Based on Zelik et al. (2010); questions by ModQu.

The framework has several uses. It allows reviewers to judge whether the effort put into an analysis matched the importance of the question. It helps analysts check their own work before it is issued. And it gives decision-makers a way to ask how far a conclusion can be trusted, based on how it was reached.

Outcome measures

Process standards describe how analysis was done but not whether it was right. A complementary approach measures accuracy directly. This requires judgements that can be scored: statements about future or unknown events, with clear criteria and, ideally, numerical probabilities.

Scoring probability judgements

The standard measure is the **Brier score**, originally proposed for weather forecasts (Brier, 1950). For each judgement, it takes the squared difference between the forecast probability and the outcome, coded as 1 if the event happened and 0 if it did not, and averages across judgements. Lower scores are better. The Brier score rewards both **calibration**, meaning that events forecast at 70 per cent happen about 70 per cent of the time, and **discrimination**, meaning that higher probabilities are given to events that occur than to those that do not.

Evidence from real intelligence

For many years it was assumed that intelligence judgements could not be scored, because they were expressed in vague words and dealt with unique events. Canadian research showed otherwise. Mandel and Barnes (2014) examined about 1,500 forecasts made by a Canadian strategic intelligence unit, which used numerical probabilities. They found that the forecasts were well calibrated and discriminated well between events that happened and those that did not; if anything, analysts were somewhat underconfident. Barnes (2016) described how the unit had introduced numerical probabilities and how this made scoring possible.

Friedman and Zeckhauser (2012) emphasise an important point: the quality of an estimate cannot be judged from a single outcome. A forecast of 80 per cent that does not come true is not necessarily wrong, since events forecast at 80 per cent should fail one time in five. Quality can only be judged across many forecasts.

Forecasting tournaments

The most extensive evidence on what improves accuracy comes from forecasting tournaments sponsored by the American intelligence research agency. Mellers et al. (2014) found that training in probabilistic reasoning, working in teams and tracking the most accurate forecasters all improved accuracy. Tetlock and Gardner (2015) describe the habits of the best performers, the "superforecasters". These tournaments showed that accuracy on geopolitical questions can be measured, compared and improved, and that the methods for doing so can be adopted by intelligence organisations.

Comparing the approaches

Two ways to measure analytic quality

Process measures can be applied at once; outcome measures need time and many judgements

	Process measures	Outcome measures
What they assess	How the analysis was done	Whether the judgements were right
Examples	Standards reviews; rigour ratings	Brier scores; calibration of forecasts
Strength	Usable immediately, on any product	Measures what ultimately matters
Limit	Good process can still give wrong answers	Needs many scored judgements and time

Compiled from Zelik et al. (2010), ICD 203 (ODNI, 2015), Brier (1950) and Mandel and Barnes (2014); table by ModQu

ModQu.com

Figure 2. Two ways to measure analytic quality. Process measures give immediate feedback on every product; outcome measures show whether judgements are actually accurate. Compiled from Zelik et al. (2010), ICD 203, Brier (1950) and Mandel and Barnes (2014); table by ModQu.

Process measures can be applied immediately to any product, including descriptive and explanatory analysis that makes no forecasts. They give analysts specific guidance on what to improve. But a product can meet every standard and still be wrong, and standards themselves may not be the ones that matter most for accuracy.

Outcome measures assess what ultimately matters, but they need many scored judgements and time for outcomes to become known. They apply only to judgements that can be checked. They also say little about why a judgement was right or wrong.

Testing the methods themselves

Measurement also allows organisations to test the methods they teach. Structured analytic techniques are widely used in intelligence, but Coulthart (2017) found that few had been rigorously evaluated. Chang et al. (2018) argued that many techniques were adopted because they seemed sensible rather than because evidence showed that they improved accuracy, and that some might even increase noise. Marrin (2012) called for closer links between intelligence practice and academic research, so that methods could be tested before being widely adopted.

Mandel et al. (2018) showed how measurement can identify what works. Testing several ways of improving analysts' probability judgements, they found that some simple statistical corrections, such as making judgements logically consistent and averaging across analysts, produced clear gains in accuracy. Such findings can only emerge where judgements are recorded and scored.

Noise audits

Kahneman et al. (2021) suggest another measurement tool: the **noise audit**. Several analysts are given the same case independently, and their judgements are compared. Wide variation shows that the organisation's judgements depend heavily on which analyst happens to handle a case, which is a quality problem even if the average judgement is correct. Noise audits require no waiting for outcomes and can be run on realistic cases at any time.

Matching rigour to stakes

Not every product needs the highest level of rigour. A short note on a minor development does not justify the same effort as an assessment that will inform a major investment or a national security decision. Zelik et al. (2010) suggest that the right question is whether the rigour of an analysis was sufficient for the decision it supported. A useful practice is to agree, at the start of each major piece of work, how important the decision is and therefore what level of rigour is expected on each attribute. Reviewers can then judge the analysis against that expectation, rather than against an ideal that no product under time pressure could meet.

Building a measurement system

Drawing on these approaches, an organisation can build a practical system for measuring analytic quality:

1. **Adopt clear standards.** Use a short list of tradecraft standards, adapted from ICD 203 or similar, that every product should meet.
2. **Review samples of products.** Rate a regular sample of products against the standards and the attributes of rigour, using reviewers who did not produce them.
3. **Record judgements in scorable form.** Ask analysts to state key judgements with numerical probabilities, clear criteria and dates.
4. **Score and report accuracy.** When outcomes are known, calculate Brier scores and calibration, for individuals and teams, and share the results.
5. **Run occasional noise audits** on representative cases.
6. **Use results to improve methods.** Compare the accuracy of judgements produced with and without particular techniques or training.
7. **Keep measurement developmental.** Use results to help analysts learn, not mainly to rank or punish them, or analysts will avoid making scorable judgements.

Applying the approach in business

Business research and intelligence teams can use the same principles. Market forecasts, competitor assessments and survey-based projections can be recorded with probabilities or ranges and scored later. Reports can be reviewed against a short checklist: are sources described, are assumptions stated, are alternatives considered, is uncertainty expressed clearly? Even a simple spreadsheet of forecasts and outcomes, reviewed once a quarter, gives a team evidence of how accurate it is and where it tends to be overconfident or biased.

Common mistakes

- **Measuring only what is easy.** Counting reports produced or readers' satisfaction says little about quality.
- **Judging from single outcomes.** One wrong forecast does not show poor analysis; one right forecast does not show good analysis.
- **Treating standards as a checklist to pass.** Standards are meant to improve reasoning, not to be satisfied with boilerplate.
- **Avoiding numbers.** Vague language cannot be scored, and what cannot be scored cannot be improved systematically.
- **Using results to punish.** Analysts who fear scoring will hedge, making their judgements less useful.

An illustration

A Canadian insurance company's risk intelligence team decides to measure its own performance. It adopts six tradecraft standards and asks a colleague from another department to review a sample of reports each quarter. It also begins recording every forecast it makes, with a probability and a date. After eighteen months it has scored more than 200 forecasts.

The results are instructive. Reviews show that the team rarely considers alternative explanations. Scoring shows that its forecasts are generally well calibrated, except for those about regulatory change, where it has been consistently overconfident. The team introduces a brief alternatives section in every report and asks a regulatory specialist to review forecasts in that area. The following year, its forecasts on regulatory questions are better calibrated, and senior managers say they trust the team's reports more because they can see its track record.

Conclusion

Measuring the quality of intelligence analysis was long considered impractical. It is now clear that it can be done. Process measures, such as tradecraft standards and attributes of rigour, show whether analysis is carried out carefully and where it is shallow. Outcome measures, based on scoring probability judgements, show whether judgements are accurate and allow organisations to test which methods work. Neither is sufficient alone. Together, applied in a developmental spirit, they allow analysts and organisations to learn from experience and to demonstrate the value of their work.

References

Barnes, A. (2016). Making intelligence analysis more intelligent: Using numeric probabilities. *Intelligence and National Security, 31*(3), 327–344. https://doi.org/10.1080/02684527.2014.994955

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Chang, W., Berdini, E., Mandel, D. R., & Tetlock, P. E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security, 33*(3), 337–356. https://doi.org/10.1080/02684527.2017.1400230

Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction. (2005). *Report to the President of the United States*. U.S. Government Printing Office.

Coulthart, S. J. (2017). An evidence-based evaluation of 12 core structured analytic techniques. *International Journal of Intelligence and CounterIntelligence, 30*(2), 368–391. https://doi.org/10.1080/08850607.2016.1230706

Fingar, T. (2011). *Reducing uncertainty: Intelligence analysis and national security*. Stanford University Press.

Fischhoff, B., & Chauvin, C. (Eds.). (2011). *Intelligence analysis: Behavioral and social scientific foundations*. National Academies Press. https://doi.org/10.17226/13062

Friedman, J. A., & Zeckhauser, R. (2012). Assessing uncertainty in intelligence. *Intelligence and National Security, 27*(6), 824–847. https://doi.org/10.1080/02684527.2012.708275

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence, Central Intelligence Agency.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences, 111*(30), 10984–10989. https://doi.org/10.1073/pnas.1406138111

Mandel, D. R., Karvetski, C. W., & Dhami, M. K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making, 13*(6), 607–621.

Marchio, J. (2014). Analytic tradecraft and the intelligence community: Enduring value, intermittent emphasis. *Intelligence and National Security, 29*(2), 159–183. https://doi.org/10.1080/02684527.2012.746415

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

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

National Research Council. (2011). *Intelligence analysis for tomorrow: Advances from the behavioral and social sciences*. National Academies Press. https://doi.org/10.17226/13040

Office of the Director of National Intelligence. (2015a). *Intelligence Community Directive 203: Analytic standards*.

Office of the Director of National Intelligence. (2015b). *Intelligence Community Directive 206: Sourcing requirements for disseminated analytic products*.

Patterson, E. S., Roth, E. M., & Woods, D. D. (2001). Predicting vulnerabilities in computer-supported inferential analysis under data overload. *Cognition, Technology & Work, 3*(4), 224–237. https://doi.org/10.1007/s10111-001-8004-y

Pherson, R. H., & Heuer, R. J., Jr. (2021). *Structured analytic techniques for intelligence analysis* (3rd ed.). CQ Press.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown.

Zelik, D. J., Patterson, E. S., & Woods, D. D. (2010). Measuring attributes of rigor in information analysis. In E. S. Patterson & J. E. Miller (Eds.), *Macrocognition metrics and scenarios: Design and evaluation for real-world teams* (pp. 65–83). Ashgate.