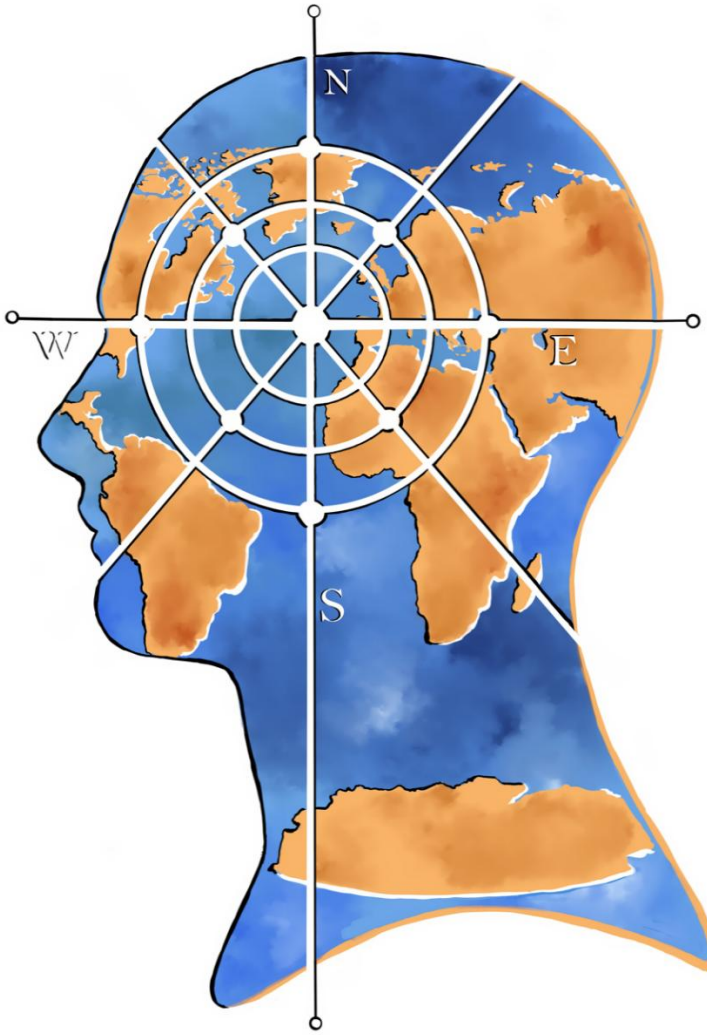
Volume 8 Number 2 December 2025 ISSN 2585-383X

JOURNAL *of* EUROPEAN *and* AMERICAN INTELLIGENCE STUDIES

AN INTERNATIONAL PEER-REVIEWED JOURNAL



Research Institute for European and American Studies – RIEAS
School of Law and Government, Dublin City University
International Centre for Policing & Security, University of South Wales

JOURNAL *of* EUROPEAN *and* AMERICAN INTELLIGENCE STUDIES

AN INTERNATIONAL PEER-REVIEWED JOURNAL

Volume 8 Number 2 December 2025 ISSN 2585-383X

JOURNAL *of* EUROPEAN *and* AMERICAN INTELLIGENCE STUDIES

An international peer-reviewed journal

Sponsored by the **Research Institute** *for* **European** *and* **American Studies – RIEAS**

Under the editorial direction of the

School of Law and Government, Dublin City University

and the **International Centre for Policing & Security, University of South Wales**

Subscriptions

***Individual rate**: 100 Euros | 100 US Dollars | 85 UK Pounds*
***Institutional rate**: 200 Euros | 200 US Dollars | 170 UK Pounds*

Mission and Scope

The *Journal of European and American Intelligence Studies* (*JEAIS*, formerly the *Journal of Mediterranean and Balkan Intelligence —JMBI*) is published by the Research Institute for European and American Studies (RIEAS) under the editorial direction of the Department of Security and Intelligence Studies at Coastal Carolina University. It is an international academic-led scholarly publication that focuses on the field of intelligence and related areas of study and practice, such as terrorism and counterterrorism, domestic and international security, geopolitics, and international relations. The journal's rationale is driven by the global nature of security challenges, where we are called more than ever to communicate and work collaboratively to solve our common problems. Thus, the *JEAIS* aspires to promote an international dialogue between diverse perspectives and experiences, based on original research on the broader European and American practice and study of intelligence. The *JEAIS* is an all-inclusive academic platform that allows accomplished and emerging scholars and practitioners from both the public and private sectors to share their knowledge, ideas and approach to intelligence studies. By crafting each journal issue through a rigorous and highly selective screening process of potential contributors, and an exhaustive review process, the *JEAIS* adheres to its mission, which is three-fold: (a) to provide an equal opportunity for academics and practitioners of intelligence to discuss and challenge established and emerging ideas; (b) to address existent knowledge gaps by advancing new knowledge; and (c) to shape the evolution of intelligence scholarship beyond traditional communities of research.

Editorial Offices

John M. Nomikos, Founding/Honorary Editor: PO Box 427, 17402 Alimos, Greece.
E-mail: secretary@rieas.gr | Tel: +30-210-991-1214 | RIEAS: www.rieas.gr.

Jonathan Smith, Managing Editor: Department of Intelligence and Security Studies, Coastal Carolina University, PO Box 261954, Conway, SC 29528-6054, United States
E-mail: jonsmith@coastal.edu | Tel: +1- 843-349-6573 | CCU: www.coastal.edu.

JOURNAL *of* EUROPEAN *and* AMERICAN INTELLIGENCE STUDIES

An international peer-reviewed journal

Sponsored by the **Research Institute** *for* **European** *and* **American Studies – RIEAS**

Under the editorial direction of the

School of Law and Government, Dublin City University
and the **International Centre for Policing & Security, University of South Wales**

Contents

Editor's Note

Joseph Fitsanakis

Professor, Department of Intelligence and Security Studies, Coastal Carolina University

We are pleased to welcome readers to Volume 8, Issue 2, of the *Journal of European and American Intelligence Studies* (*JEAIS*). This issue brings together a diverse and methodologically rich set of contributions that collectively interrogate some of the most enduring and contested problems in intelligence studies: analytic rigor and judgment under uncertainty, the communication of probability and confidence, the structural conditions shaping contemporary information environments, and the strategic implications of information operations for democratic governance. Across empirical, conceptual, and critical traditions, the articles in this volume reflect the field's continued maturation and its willingness to engage both the internal mechanics of intelligence work and the broader ecosystems in which intelligence operates.

The issue opens with Gideon Manger and Sanne van der Weide's empirical examination of the relationship between analytical rigor and predictive accuracy in intelligence assessments. Drawing on an original dataset of assessments produced during analyst qualification training within the Netherlands Armed Forces, the authors directly address a question that has long preoccupied both scholars and practitioners: whether adherence to established tradecraft standards measurably improves forecasting outcomes. Their findings offer important nuance. While rigor is positively associated with successful predictions, its relationship with the precision of probabilistic judgments proves weaker than expected. Particularly noteworthy is their analysis of "50–50" assessments, which emerge as both methodologically less rigorous and substantively less useful for intelligence consumers. The article makes a valuable contribution by empirically grounding debates about analytic standards, probability expression, and evaluative frameworks—while also raising important questions about the cross-cultural transferability of analytic rating scales.

Jeremy Levin's article continues the focus on probability and judgment but approaches the problem from a conceptual and methodological standpoint. Levin challenges the uncritical application of quantified probability to qualitative analytic judgments, particularly in contexts characterized by limited data and narrow historical baselines. He proposes a distinction between communicated probability and analytic certainty, arguing that the latter more accurately captures the logic underpinning many intelligence judgments. By introducing argument mapping as a tool for calibrating analytic certainty,

Levin offers a framework designed to enhance transparency, replicability, and collaborative reasoning. This contribution speaks directly to ongoing debates about how intelligence organizations should reason, communicate uncertainty, and maintain rigor when statistical approaches are insufficient or misleading.

Shifting from analytic cognition to structural power, Elena Botts' "Opaque Architectures" offers a critical examination of the convergence between media consolidation, cultural funding, and intelligence cooptation in contemporary information environments. Drawing on cases from the Euro-American and Russian contexts, Botts argues that state-affiliated financing mechanisms increasingly function as instruments of epistemic enclosure rather than mere support for cultural production. The article advances the concept of an "epistemic cartel" to describe a durable infrastructure of perception management in which transparency is redefined through state-sanctioned visibility. This theoretically dense and provocative contribution extends intelligence studies into dialogue with media theory, political economy, and critical security studies, underscoring the field's relevance to broader questions of democratic accountability and knowledge production.

Alan Cunningham's article returns the focus to contemporary strategic competition by examining Russian information operations and their impact on American foreign policy discourse. Emphasizing the role of domestic intermediaries in amplifying disinformation, Cunningham situates political security as a multidimensional challenge encompassing both human and national security concerns. The article highlights the permeability of democratic systems to sustained influence campaigns and argues for a more systematic integration of political security into policy planning. In doing so, it contributes to a growing body of literature that treats information operations not as episodic disruptions, but as enduring features of modern conflict.

The issue concludes with Adam Hanzel's review of Simon Ball's *Death to Order: A History of Modern Assassination*. Hanzel situates Ball's work as a rare and comprehensive treatment of assassination as a transnational political practice, emphasizing its analytical value for scholars and practitioners across multiple disciplines. The review complements the issue's broader themes by reminding readers of the historical continuities that underpin contemporary security practices.

Taken together, the contributions in Volume 8, Issue 2, reflect the intellectual breadth of intelligence studies today. They demonstrate the field's capacity to integrate empirical evaluation, conceptual innovation, and critical analysis, while remaining attentive to the practical and ethical stakes of intelligence work. We hope readers find this issue both challenging and illuminating, and that it stimulates further research and debate across the many domains in which intelligence intersects with policy, society, and power.

The Effect of Analytical Rigor on Accuracy of Intelligence Forecasts

Gideon Manger[i] and **Sanne van der Weide**[ii]

[i] PhD candidate, Leiden University, Leiden, The Netherlands; and Netherlands Defence Academy, Breda, the Netherlands
[ii] Graduate, International Master in Security, Intelligence, and Strategic Studies, University of Glasgow, Glasgow, United Kingdom; Dublin City University, Dublin, Ireland; and Charles University, Prague, Czech Republic

Abstract

Our study examines the relationship between the rigor of intelligence processes and the accuracy of predictive intelligence assessments. While much discussion has focused on evaluating the performance of intelligence organizations through predictive accuracy, the mechanisms behind the varying accuracy of predictive intelligence assessments remain unclear. One critical gap is the relationship between rigorous intelligence processes and predictive accuracy. Intelligence processes are rigorous when they adhere to established quality principles commonly used in intelligence tradecraft. Predictive accuracy reflects the intelligence professionals' skill in predicting event probabilities and successfully anticipating events. Data on rigor and predictive accuracy were collected from 105 intelligence assessments produced during qualification training for intelligence analysts in the Netherlands Armed Forces (2018–2023). While the results support the hypothesized association of rigor and successful predictions, the correlation between rigor and the precision of probability assessments is weak. This might indicate that the U.S. rating scale used in this study is inappropriate for assessing rigor in non-U.S. intelligence organizations. Another possibility is that the large number of 50-50 assessments may have diluted the relationship. Furthermore, our results indicate that the degree of rigor in 50-50 assessments and prediction failures is significantly lower than in successful predictions, suggesting that 50-50 assessments are uninformative for intelligence consumers and associated with poorer intelligence tradecraft.

Introduction

Predictive assessments of future developments inform national security decision-making. By generating accurate predictive assessments of future developments, intelligence forecasts aim to enhance national security decisions, inform policy, formulate strategies, and facilitate decision-making[1]. These predictions are based on the rigorous execution of data collection, analysis, and evaluation processes. Generally, rigorous processes are believed to produce accurate predictive intelligence assessments in intelligence products.

The compelling idea that intelligence processes influence predictive assessments has generated significant academic interest. Yet, the assumption that executing these processes rigorously enhances accuracy is unexplored.[2] Research on the process rigor emphasizes either employing specific methods during analysis[3] or evaluating rigor through their compliance with established standards[4] to increase predictive accuracy. Measuring rigor through methods is problematic due to the rare use of standardized methods.[5] Intelligence organizations have developed measurement instruments for rigor using predefined standards[6] to boost the rigor of intelligence processes. However, it remains unclear how well intelligence processes adhere to these standards or whether adherence impacts on predictive accuracy. Similarly, research on forecasting accuracy has identified factors influencing accuracy, such as judgment skill, group sizes, analytic cognitive styles, and update frequency[7]. Additionally, descriptive studies on

[1] David R. Mandel, "Accuracy of Intelligence Forecasts From the Intelligence Consumer's Perspective," *Policy Insights from the Behavioral and Brain Sciences* 2, no. 1 (2015), https://doi.org/10.1177/2372732215602907.

[2] Gideon Manger, "Unravelling effectiveness in intelligence: a systematic review," *Intelligence and National Security* 39, no. 7 (2024): 1149, https://doi.org/10.1080/02684527.2024.2370132.

[3] e.g. Stephen J. Coulthart, "An Evidence-Based Evaluation of 12 Core Structured Analytic Techniques," *International Journal of Intelligence and CounterIntelligence* 30, no. 2 (2017), https://doi.org/10.1080/08850607.2016.1230706; Lars C Borg, "Reduced analytic uncertainty through increased analytic rigour: effects of using structured analytic techniques in estimative intelligence" (Brunel University London, 2022).

[4] e.g. A Barnett et al., "Analytic Rigour in Intelligence," (Melbourne, Australia: Hunt Laboratory for Intelligence Research, University …, 2021); Daniel J Zelik, Emily S Patterson, and David D Woods, "Measuring attributes of rigor in information analysis," in *Macrocognition metrics and scenarios: Design and evaluation for real-world teams*, ed. E. S. Patterson & J. Miller (Aldershot, UK: Ashgate: CRC Press, 2010).

[5] Rubén Arcos and José-Miguel Palacios, "EU INTCEN: a transnational European culture of intelligence analysis?," *Intelligence and National Security* 35, no. 1 (2019): 85, https://doi.org/10.1080/02684527.2019.1649912; Stephen Coulthart, "Why do analysts use structured analytic techniques? An in-depth study of an American intelligence agency," *Intelligence and National Security* 31, no. 7 (2016): 940, https://doi.org/10.1080/02684527.2016.1140327.

[6] Alexandru Marcoci, Ans Vercammen, and Mark Burgman, "ODNI as an analytic ombudsman: is Intelligence Community Directive 203 up to the task?," *Intelligence and National Security* 34, no. 2 (2019), https://doi.org/10.1080/02684527.2018.1546265; Barry Zulauf, "From a Former ODNI Ombudsperson Perspective: Safeguarding Objectivity in Intelligence Analysis," *Studies in Intelligence* 63, no. 3 (2021).

[7] E.g. Pavel Atanasov et al., "Small Steps to Accuracy: Incremental Belief Updaters Are Better Forecasters," *Proceedings of the 21st ACM Conference on Economics and Computation* (2020), https://doi.org/10.1145/3391403.33995; Barbara A. Mellers et al., "How generalizable is good judgment? A multi-task, multi-benchmark study," *Judgment and Decision Making* 12, no. 4 (2017), https://doi.org/10.1017/S1930297500006240; Michael Horowitz et al., "What makes foreign policy teams tick: Explaining variation in group performance at geopolitical forecasting," *The Journal of Politics* 81, no. 4 (2019), https://doi.org/10.1177/1555343414554702; Joshua C. Poore et al., "Personality, Cognitive Style, Motivation, and Aptitude Predict Systematic Trends in Analytic

forecasting accuracy in intelligence organizations aid in benchmarking and comparing judgment accuracy across samples.[8] However, it remains unclear how predictive accuracy is influenced by rigor[9]. There is, in other words, a lack of evidence connecting rigor to predictive accuracy.

The discussion surrounding rigor and predictive accuracy assumes a relationship between rigorous intelligence processes and the accuracy of predictions. To address this gap, this study seeks to answer the following research question: Does the rigor of intelligence processes affect the accuracy of intelligence predictions? By answering this question, our study aims to contribute to intelligence theory and practice. Its theoretical contribution is to explore the unexplored relationship between rigor and the accuracy of intelligence forecasts. In this manner, it contributes to broader debates in intelligence analysis, particularly regarding performance evaluation and methodological standards. Lastly, the study aims to help practitioners optimize processes to improve forecasting accuracy.

Background and Literature Review

In 2004, the United States Congress passed the Intelligence Reform and Terrorism Prevention Act (IRTPA) in reaction to flawed intelligence reporting leading up to the 9/11 attacks and the U.S. invasion of Iraq. The act required the U.S. intelligence community to uphold stricter quality standards in its intelligence processes. Although not mentioned explicitly in the act, more stringent quality standards are imposed, at least partly, to improve forecasting accuracy. In 2007, two initiatives were taken to strengthen U.S. intelligence. One was the Intelligence Community Directive 203 (ICD203), laying out quality standards for intelligence products.[10] The other was the Aggregative Contingent Estimation (ACE) project,[11] which sought to improve the accuracy, precision, and timeliness of intelligence predictions by organizing forecasting tournaments. This dual approach of examining rigor and accuracy separately reflects past efforts to study these two aspects. One of the first efforts to assess forecasting accuracy was the CIA's validity studies regarding its estimates in the 1950s.[12] The ACE project may be viewed as a continuation of these initiatives. Early attempts to determine the rigor of

Forecasting Behavior," *Journal of Cognitive Engineering and Decision Making* 8 (2014), https://doi.org/10.1177/1555343414554702.

[8] E.g. David R. Mandel and Alan Barnes, "Accuracy of forecasts in strategic intelligence," *Proceedings of the National Academy of Sciences* 111, no. 30 (2014), https://doi.org/10.1073/pnas.1406138111; David R. Mandel and Alan Barnes, "Geopolitical Forecasting Skill in Strategic Intelligence," *Journal of Behavioral Decision Making* 31, no. 1 (2018), https://doi.org/10.1002/bdm.2055; David R. Mandel and Daniel Irwin, "Tracking accuracy of strategic intelligence forecasts: Findings from a long-term Canadian study," *Futures & Foresight Science* 3, no. 3-4 (2021), https://doi.org/10.1002/ffo2.98.

[9] Alexandru Marcoci et al., "Better together: Reliable application of the post-9/11 and post-Iraq U.S. intelligence tradecraft standards requires collective analysis," *Frontiers in psychology* 9 (2019): 8, https://doi.org/10.3389/fpsyg.2018.02634.

[10] ODNI, "Intelligence Community Directive (ICD) 203- Analytic Standards," ed. Office of the Director of National Intelligence (2022). https://www.dni.gov/files/documents/ICD/ICD-203.pdf.

[11] "Aggregative Contingent Estimation (ACE)," 2011, accessed 04 NOV 2024, 2024, https://www.iarpa.gov/research-programs/ace.

[12] CIA, "A Study of National Intelligence Estimates on the USSR, 1950–57"; Central Intelligence Agency, (1958).

intelligence processes focused on developing intelligence processes[13], followed by introducing procedures and techniques[14], and more recently by ODNI's efforts to assess rigor through predefined standards. These efforts sought to prevent intelligence failures and boost intelligence's forecasting accuracy.

Both approaches have provided intelligence studies scholars with insights into the dynamics of both rigor and accuracy. Moreover, efforts to explore rigor and accuracy have spilled over to scholars outside the U.S. Research has explored the accuracy of Canadian intelligence predictions.[15] Australian, British, and Norwegian scholars have sought to examine the measurement of rigor.[16] These efforts have not only extended the debates of rigor and accuracy beyond the borders of the U.S. They have also shown that concepts of rigor initially devised by ODNI might be applied to the intelligence organizations of U.S. allies.

Measuring Rigor of Intelligence

One can view rigor through adherence to methodologies or predefined standards. Heuer inspires proponents of the methods perspective.[17] Proponents stress the importance of constraining cognitive biases, scrutinizing various hypotheses, and cultivating systematic critical thinking in intelligence analysis using structured analytic techniques (SATs). Intelligence organizations are believed to have increased the use of SATs to enhance rigorous analysis, reduce the risk of intelligence failures, and make analysts' reasoning more transparent to intelligence consumers.[18] This is partly based on the observation that SATs have been prominent in educating and training intelligence personnel. However, whether standardized methods improve quality in intelligence processes remains unclear because analysts rarely use them.[19]

Some scholars reject using generic methodologies of any kind in intelligence processes. Breakspear[20], for example, argues that implementing a standardized approach risks making intelligence practices more rigid and procedural. Fearing this could lead to the

[13] Sherman Kent, *Strategic intelligence for American world policy* (Princeton, N.J.: Princeton University Press, 1966), 1 online resource (xxv, 226 pages).
[14] Richards J. Heuer, *Psychology of intelligence analysis* ([Washington, D.C.]: Center for the Study of Intelligence, Central Intelligence Agency, 1999), 1 online resource (xxv, 184 pages) : illustrations.
[15] Mandel and Barnes, "Accuracy of forecasts in strategic intelligence."; Mandel and Barnes, "Geopolitical Forecasting Skill in Strategic Intelligence."; Mandel and Irwin, "Tracking accuracy of strategic intelligence forecasts: Findings from a long-term Canadian study."
[16] Barnett et al., "Analytic Rigour in Intelligence."; Borg, "Reduced analytic uncertainty through increased analytic rigour: effects of using structured analytic techniques in estimative intelligence."; Marcoci et al., "Better together: Reliable application of the post-9/11 and post-Iraq U.S. intelligence tradecraft standards requires collective analysis."; Luke Thorburn et al., "The IC Rating Scale as a Measure of Analytic Rigor"; SSRN, (2022).
[17] Heuer, Psychology of intelligence analysis.
[18] Stephen J Artner, Richard S Girven, and James Bruce, Assessing the value of structured analytic techniques in the U.S. intelligence community (RAND, 2017).
[19] Michael J Ard, "Structured Analytic Techniques: A Pragmatic Approach," International Journal of Intelligence and CounterIntelligence (2023): 619.
[20] Alan Breakspear, "A New Definition of Intelligence," Intelligence and National Security 28, no. 5 (2013), https://doi.org/10.1080/02684527.2012.699285.

institutionalization of methods based on the mistaken belief that there is a single correct way to conduct intelligence. The efficacy of institutionalizing SATs as a generic methodology is specifically contested, and their absence in intelligence organizations [21] may serve as evidence of their inadequacy. This viewpoint asserts that every intelligence process is unique, arguing that standardized procedures are unsuitable. However, intelligence processes must still adhere to specific quality standards. Advocates of a standards-based approach emphasize that the effectiveness of intelligence relies on standards of objectivity, neutrality, accuracy, persuasiveness, timeliness, and relevance. General standards are outlined in ICD203[22] and the Logical, Objective, Thorough, Stringent, Acute (LOTSA) dimensions of rigor[23]. For both ICD 203 and the LOTSA dimensions, adherence to standards is measured in the intelligence product. The core idea is that the outcomes of intelligence processes must logically align with these standards. While methodological rigor measures rigor directly from a workflow or process, standards rigor can be utilized more broadly by measuring rigor from the product of intelligence processes. In this way, standards can accommodate divergent processes. For this study, rigor is defined as the degree to which intelligence reflects attributes of specific quality standards. Due to the limited application of the LOTSA dimensions to measure rigor, this study uses the ICD203 criteria to assess rigor.

The ICD203 describes nine quality criteria for U.S. intelligence (see *Table 1*). To assess the adherence to these standards, a rating scale has been developed to determine nine items of analytic tradecraft on a four-point scale and evaluate each item as poor, fair, good, or excellent.[24] One issue with this instrument is the ODNI Rating Scale's (henceforth Rating Scale) validity and reliability. Marcoci et al. find that the inter-rater reliability of the rating scale is poor for individual raters and good to excellent when groups of three raters apply the instrument.[25] Whether the Rating Scale has validity as an indicator of rigor remains open; they argue that more research is needed to evaluate whether it is a valid and reliable quality assessment tool. Questioning the validity of ICD203, Gentry argues that despite its best intentions to improve rigor in intelligence processes, implementing these measures has had a marginal impact.[26] Although he sees some improvement in new and weak analysts, these standards do not guide the more experienced intelligence officer. Unfortunately, there is no meaningful empirical evidence to support this claim. Although the validity of ICD203 standards remains to be tested, they have been used for more than a decade in the U.S. intelligence community, indicating their validity and reliability.

[21] Arcos and Palacios, "EU INTCEN: a transnational European culture of intelligence analysis?," 85; Coulthart, "Why do analysts use structured analytic techniques? An in-depth study of an American intelligence agency," 940.
[22] ODNI, "Intelligence Community Directive (ICD) 203- Analytic Standards."
[23] Barnett et al., "Analytic Rigour in Intelligence."
[24] ODNI, "Rating Scale for Evaluating Analytic Tradecraft Standards," (2007).
[25] Marcoci et al., "Better together: Reliable application of the post-9/11 and post-Iraq U.S. intelligence tradecraft standards requires collective analysis."
[26] John A. Gentry, "Has the ODNI Improved U.S. Intelligence Analysis?," *International Journal of Intelligence and CounterIntelligence* 28, no. 4 (2015), https://doi.org/10.1080/08850607.2015.1050937.

Criterion	Description
1	Properly describes quality and credibility of underlying sources, data, and methodologies.
2	Properly expresses and explains uncertainties associated with major analytic judgments.
3	Properly distinguishes between underlying intelligence information and analysts' assumptions and judgments.
4	Incorporates analysis of alternatives.
5	Demonstrates customer relevance and addresses implications.
6	Uses clear and logical argumentation.
7	Explains change to or consistency of analytic judgments.
8	Makes accurate judgments and assessments.
9	Incorporates effective visual information where appropriate.

Table 1: *ICD203 standards of analytic tradecraft adapted from ODNI and Marcoci* [27]

Measuring forecasting accuracy of intelligence

Measuring accuracy in intelligence is challenging. When producing predictive judgments, intelligence organizations often operate in conditions where perfect clarity is unattainable.[28] There is an inherent uncertainty in intelligence work, complicating efforts to make definitive predictions. Forecasting accuracy is hindered by data ambiguity, long-term uncertainty, and the need for robust tracking mechanisms. While aggregation of judgments can enhance accuracy[29], systemic challenges such as biases and delayed prediction verification hamper evaluation. Intelligence predictions are often long-term, complicating immediate assessment of accuracy and timely performance-tracking.[30] To mitigate these challenges, Chang proposes that predictions be aimed at specific events, using specified probabilities and delineated time

[27] ODNI, "Intelligence Community Directive (ICD) 203- Analytic Standards."; Marcoci et al., "Better together: Reliable application of the post-9/11 and post-Iraq U.S. intelligence tradecraft standards requires collective analysis.".
[28] Richard K Betts, "Analysis, war, and decision: Why intelligence failures are inevitable," *World politics* 31, no. 1 (1978): 87, https://doi.org/10.2307/2009967.
[29] James E. Kajdasz, "A Demonstration of the Benefits of Aggregation in an Analytic Task," *International Journal of Intelligence and CounterIntelligence* 27, no. 4 (2014), https://doi.org/10.1080/08850607.2014. 924814.
[30] Mark M. Lowenthal and Ronald A. Marks, "Intelligence Analysis: Is It As Good As It Gets?," *International Journal of Intelligence and CounterIntelligence* 28, no. 4 (2015), https://doi.org/10.1080/08850607.2015. 1051410.

horizons.[31] Another challenge keeping intelligence organizations from openly assessing their judgment accuracy is secrecy. Intelligence organizations are reluctant to share how precisely they can predict future events for fear of making adversaries aware of their capabilities. Furthermore, intelligence organizations risk partially revealing their areas of interest. The last challenge is publicizing their judgment accuracy; intelligence organizations may expose themselves to public criticism if accuracy is misaligned with political or public expectations. These challenges have restricted data access for scholars.

In the absence of access to intelligence organizations, scholars have examined the dynamics of forecasting accuracy in similar contexts. Tetlock and Gardner identify traits of superforecasters from forecasting tournaments.[32] Their research has focused on the differences that distinguish consistent high performers, attributing superior forecasting to cognitive abilities, task-specific skills, motivation[33], and update frequency[34]. Other researchers have found that the application of probabilistic reasoning[35], accountability for forecasting outcomes[36], personality, cognitive style, motivated cognition, and motivational focus[37] were found to affect accuracy in forecasting tournaments. In contrast, domain expertise was not shown to lead to improved predictive accuracy.[38] Moreover, individual performance often lags group performance, with no individual forecaster outperforming collective judgments.[39] Research into accuracy has yielded many fascinating insights. However, whether predictive accuracy is affected by the degree of rigor remains to be determined.

Accumulated knowledge regarding predictive accuracy makes measuring accuracy relatively straightforward. On the one hand, predictive accuracy can be measured by assessing the extent to which assessments are 'getting it right'. This is referred to as discrimination. Discrimination indicates a forecaster's ability to distinguish between the occurrence and non-occurrence of an event.[40] Perfect discrimination would entail that a forecaster perfectly distinguishes the occurrence and non-occurrence of future events. A simple measure of discrimination is the success-failure

[31] Welton Chang, "Getting It Right Assessing the Intelligence Community's Analytic Performance," *American Intelligence Journal* 30, no. 2 (2012), http://www.jstor.org/stable/26202020.
[32] Philip E. Tetlock and Dan Gardner, *Superforecasting: The art and science of prediction* (London: Random House Business, 2019).
[33] Barbara Mellers et al., "Identifying and cultivating superforecasters as a method of improving probabilistic predictions," *Perspectives on Psychological Science* 10, no. 3 (2015), https://doi.org/10.1177/1745691615577794.
[34] Atanasov et al., "Small Steps to Accuracy: Incremental Belief Updaters Are Better Forecasters."
[35] Mellers et al., "How generalizable is good judgment? A multi-task, multi-benchmark study."
[36] Welton Chang et al., "Accountability and adaptive performance under uncertainty: A long-term view," *Judgment and Decision Making* 12, no. 6 (2017), https://doi.org/10.1017/S1930297500006732; P. E. Tetlock and B. A. Mellers, "Intelligent Management of Intelligence Agencies: Beyond Accountability Ping-Pong," *American Psychologist* 66, no. 6 (2011), https://doi.org/10.1037/a0024285.
[37] Poore et al., "Personality, Cognitive Style, Motivation, and Aptitude Predict Systematic Trends in Analytic Forecasting Behavior."
[38] Philip E Tetlock, *Expert Political Judgment* (Princeton: Princeton University Press, 2017). https://doi.org/10.1515/9781400888818.
[39] Horowitz et al., "What makes foreign policy teams tick: Explaining variation in group performance at geopolitical forecasting."
[40] Chang, "Getting It Right Assessing the Intelligence Community's Analytic Performance," 102; Tetlock, *Expert Political Judgment*, 47.

distinction. Failure indicates the inability to predict, and success refers to the ability to predict the event. Failure or success can be established through a logical test. If an event is predicted not to occur and does not occur, then this would be considered a prediction success. If the event occurs, it should be regarded as a failure to predict. If the event is predicted to occur, non-occurrence constitutes failure, and occurrence constitutes success. On the other hand, intelligence predictions are generally probabilistic and thus indicate the degree of uncertainty regarding occurrence and non-occurrence. In this sense, forecasting aims to provide correct probabilities of future events unfolding. This is referred to as calibration and is defined as the degree to which predicted probabilities for an event match the actual rate of occurrence[41]. With perfect calibration, forecasters can determine the likelihood of occurrence and non-occurrence. Through calibration, forecasters also predict how often they cannot discriminate between occurrence and non-occurrence. Intelligence forecasters cannot achieve both perfect discrimination and perfect calibration. They attempt to maximize both. A commonly used metric to measure calibration and discrimination in prediction is the Brier Score, originally used to evaluate meteorological forecasts.[42] It measures how well the communicated probabilities describe what occurs[43]. Tetlock and his many collaborators have championed using Brier Scores to measure judgment accuracy. As a result, the metric is widely used in accuracy research. This study will use the success-failure distinction and the Brier Score as measures of forecasting accuracy because differing approaches may yield better insight into the relationship with rigor.

The rigor-accuracy relationship

Although the relationship between rigor and accuracy has been hypothesized, whether improved rigor leads to greater accuracy of intelligence predictions still needs to be addressed. The rationale behind our expectation is that a relationship between rigor and accuracy is supported by intelligence studies literature. For one, Heuer asserts that analytical rigor is crucial for helping decision-makers understand their environment and reduce uncertainty by estimating future outcomes.[44] In addition, the ability to provide accurate predictions is hypothesized to be associated with rigorous intelligence processes.[45] The underlying assumption for both claims is that

[41] Chang, "Getting It Right Assessing the Intelligence Community's Analytic Performance," 102; Tetlock, *Expert Political Judgment*, 47.
[42] Glenn W Brier, "Verification of forecasts expressed in terms of probability," *Monthly weather review* 78, no. 1 (1950).
[43] R.L. Winkler et al., "Scoring rules and the evaluation of probabilities," *Test* 5, no. 1 (1996): 3, https://doi.org/10.1007/BF02562681.
[44] Heuer, *Psychology of intelligence analysis*, 53.
[45] Donald J. Calista, "Enduring Inefficiencies in Counterintelligence by Reducing Type I and Type II Errors Through Parallel Systems: A Principal-Agent Typology," *International Journal of Intelligence and CounterIntelligence* 27, no. 1 (2014), https://doi.org/10.1080/08850607.2014.842809; Ehud Eiran, "The Three Tensions of Investigating Intelligence Failures," *Intelligence and National Security* 31, no. 4 (2016), https://doi.org/10.1080/02684527.2015.1044293; Kira Vrist Rønn, "(Mis-) Informed Decisions? on Epistemic Reasonability of Intelligence Claims," *International Journal of Intelligence and CounterIntelligence* 27, no. 2 (2014), https://doi.org/10.1080/08850607.2014.842813; Marcoci et al., "Better together: Reliable application of the post-9/11 and post-Iraq U.S. intelligence tradecraft standards requires collective analysis."; Marcoci, Vercammen, and Burgman, "ODNI as an analytic ombudsman: is Intelligence Community Directive 203 up to the task?."; Tetlock and Mellers, "Intelligent Management of Intelligence Agencies: Beyond Accountability Ping-Pong," 549.

improvements in rigor may lead to bias reduction, which should positively impact predictive accuracy.[46] This assertion is supported by various case studies[47]. Consequently, this paper poses the following hypotheses regarding the expected relationship between rigor and accuracy:

- *Hypothesis 1*: Higher degrees of rigor in intelligence processes are associated with more accurate probability judgments in intelligence forecasts.

- *Hypothesis 2*: Higher degrees of rigor in intelligence processes are associated with successful prediction in intelligence forecasts.

Materials and Methods

Materials

This study utilizes intelligence reports from junior analysts during a four-week analysis training course at the Intelligence and Security Academy (ISA). ISA is the Royal Netherlands Armed Forces' training institute for military intelligence personnel. The analysis course is mandatory for intelligence analysts in the Netherlands Armed Forces, and participants include both military and civilian intelligence personnel. Participants learn a standardized analysis process using SATs during the course. The course features two weeks of workshops on analysis and techniques, and group work on a real-world intelligence question in groups of three to four. This is followed by a week of individual work, where the participants repeat the analysis process. Both group and individual reports follow the same structure and aim to make predictive judgments on real-world security issues. Participants are restricted to open-source reporting. By the end, participants compile a 3,000-word intelligence report, deliver a 5-minute briefing, and present their work on the analysis process and corresponding SATs, all in Dutch. After the course, intelligence, briefing, and process reports are archived. Reports used in this study lack classified information. Permission to use archived reports for data collection is granted as they include no personal information.

248 intelligence reports produced between February 2018 and March 2024 were recovered from the archive. 30 reports were excluded because the predicted event deadline had not yet passed. An additional 13 reports were excluded because they lacked a written report, and 6 were excluded because a probabilistic statement was missing. Of the remaining reports, not all contained predictive judgments. A further 92 reports were excluded due to the descriptive nature of their assessments. Moreover, two reports made a conditional prediction. These assessments did not predict a specific, verifiable event and were excluded. This left 105 reports for our study.

[46] Manger, "Unravelling effectiveness in intelligence: a systematic review," 1148.

[47] e.g. Rubén Arcos and José-Miguel Palacios, "The impact of intelligence on decision-making: the EU and the Arab Spring," *Intelligence and National Security* 33, no. 5 (2018), https://doi.org/10.1080/02 684527.2018. 1434449; Stephen Mettler, "Return of the Bear Learning from Intelligence Analysis of the USSR to Better Assess Modern Russia," *American Intelligence Journal* 35, no. 2 (2018); James J. Wirtz, "The Art of the Intelligence Autopsy," *Intelligence and National Security* 29, no. 1 (2014), https://doi.org/10.1080/02684527 .2012.748371.

Procedures

Data were collected in two stages. First, rigor was measured using the Rating Scale, involving seven of the nine predetermined categories. Two categories were omitted as inapplicable: "Explains change/consistency of analytical judgments" due to unavailable prior judgments and "Incorporates visual information," since written reports lacked visual data. Marcoci et al. found that interrater reliability of the rating scale was acceptable when independently rated by three raters.[48] They also omitted criterion 7, 'Explains change/consistency of analytical judgments', because of the absence of a prior forecast. In addition, they found that dropping the category 'Incorporates visual information' improved reliability. One criterion was adapted. Criterion 8, which rates the degree to which the judgments are accurate, consists of two elements: a. the degree to which judgments are correct, and b. the degree to which judgments describe key factors anticipating events. Since the degree to which judgments are correct is part of our accuracy measure, we only assessed the degree to which judgments describe key factors anticipating events for criterion 8. Each category had four subcodes—poor, fair, good, excellent—with scores from 0 (poor) to 3 (excellent). Consequently, the Rigor Score of each report was determined by adding the numerical scores of criteria 1 to 6 and criterion 8. The total Rigor Score for each report ranged from 0 to 21. The codebook and the coding instructions are included in the supporting information.[49]

In our study, two raters evaluated reports after calibration training to ensure consistent application of the Rating Scale. The training consisted of three phases. In the pre-calibration phase, each rater independently rated five reports based on their interpretation of the Rating Scale. Comparing Rigor Scores revealed differences in scale interpretation. A brief review addressed these differences, refining the rating instructions for increased uniformity. Next, each coder rated another ten reports using the improved instructions. Further differences between raters prompted additional refinements. Ten new reports were coded independently in the final phase to assess inter-rater reliability, achieving good results with a Krippendorff alpha of 0.84. The remaining reports were divided equally among the raters.

The second stage assessed the accuracy of predictive judgments in the reports through five steps. First, the predictive judgments were extracted from the reports. These judgments expressed the expected event and a verbal probability of that event's occurrence. Second, judgments were classified as easy or difficult according to the methodology by Mandel and Barnes[50]. Third, a *Nexis database*[51] search was conducted to assess whether the event occurred

[48] Marcoci et al., "Better together: Reliable application of the post-9/11 and post-Iraq U.S. intelligence tradecraft standards requires collective analysis."
[49] Manger, Gideon, and Sanne van der Weide. 2025. "Supplemental Materials for "The Effect of Analytical Rigor on Accuracy of Intelligence Forecasts"." DataverseNL. https://doi.org/10.34894/JLKINY.
[50] Mandel and Barnes, "Accuracy of forecasts in strategic intelligence."
[51] The Nexis *database* offers news and business publications from diverse sources, providing access to twenty years of archives. It is operated by LexisNexis. https://www.lexisnexis.co.uk/products/nexis.html

during the forecast period, assigning a score of 1 for occurrence and 0 for non-occurrence. To verify data integrity, two researchers conducted the database search independently of each other. The results differed for five assessments, which could be attributed to mistakes made during manual data entry and were corrected accordingly. Fourth, verbal probabilities were converted into numeric ranges used by the Netherlands armed forces (see *Table 2*), with the median as the forecast probability. Finally, the accuracy of each forecast was calculated

Verbal expressions probability	Percentages
Confirmed	>95%
Highly likely	90-95%
Likely	60-90%
Even chance	40-60%
Unlikely	10-40%
Highly unlikely	5-10%
Remote	<5%

Table 2: *Verbal and corresponding numeric expressions of probability in the Netherlands Armed Forces*

Accuracy is calculated using the Brier Score and a success-failure distinction. The Brier Score assesses prediction accuracy while considering the assigned probabilities. With binary outcomes, in this case, occurrence or non-occurrence, the Brier Score is the squared error of the forecasted probability and observed outcome.[52] The squared error for each forecast is considered a strictly proper scoring rule because it incentivizes careful and honest forecasting.[53] The Brier Score for each forecast is calculated as follows:

$$BS = (f_i - o_i)^2$$

Where:

- f_i is the forecasted probability,
- o_i is event occurrence, 0 for non-occurrence, and 1 = occurrence.

[52] Brier, "Verification of forecasts expressed in terms of probability."
[53] Winkler et al., "Scoring rules and the evaluation of probabilities," 2.

The Brier Score ranges from 0 (perfect accuracy) to 1 (perfect inaccuracy). For example, when the forecasted probability is 0.25 and the event did not occur within the predicted timeframe, the Brier Score is $(0.25\text{-}0)^2 = 0.0625$. In case of occurrence, the Brier Score is $(0.25\text{-}1)^2 = 0.5625$.

In the success-failure distinction, failure refers to the inability to predict the occurrence of the forecasted event, and success refers to the ability to predict the event. Failure or success is established through a logical test. This is considered a success if the probability is lower than 50% and the event did not occur (occurrence = 0). If the event did occur (occurrence = 1), then it is a failure. If the forecast probability is greater than 50%, non-occurrence constitutes failure, and occurrence constitutes success. Using the success-failure distinction also excludes 50% of forecasts, as they give no guidance on the expected occurrence value.

Analysis

First, we summarize our data and present descriptive statistics to describe the characteristics of our data. To understand what the mean Brier Score (BS) indicates, we break it down into three components to gain insights into the specific strengths and weaknesses of our participants' probabilistic predictions. The components of the mean Brier Score include the variance index (VI), the calibration index (CI), and the discrimination index (DI)[54]. VI reflects the inherent uncertainty of the forecasting environment, with a range from 0 (no uncertainty) to 0.25 (perfect uncertainty). DI measures a forecaster's ability to distinguish between events and non-events, evaluating how effectively forecasters assign higher probabilities to events that occur and lower probabilities to non-occurrences. Good discrimination means the forecasts effectively separate likely occurrences from non-occurrences. Lastly, CI measures the alignment between predicted probabilities and observed frequencies, assessing how closely an event's predicted probabilities match its observed frequencies. Perfect calibration occurs when the proportion of events matches the forecasted probability. The mean Brier Score decomposition for our study is calculated as follows:

$$\overline{\mathrm{BS}} = VI - DI + CI$$

$$VI = \bar{o} \cdot (1 - \bar{o})$$

$$DI = \frac{1}{N} \sum_{k=1}^{K} N_k \cdot (o_k - \bar{o})$$

$$CI = \frac{1}{N} \sum_{k=1}^{K} N_k \cdot (f_k - o_k)$$

[54] J. Yates and Frank, "External correspondence: Decompositions of the mean probability score," *Organizational behavior and human performance.* 30, no. 1 (1982), https://doi.org/10.1016/0030-5073(82)90237-9.

Where:

- $\bar{o}$ is the mean occurrence,
- K is the number of forecast categories,
- N_k is the number of forecasts in category k,
- o_k is the relative frequency of event occurrence in category k,
- f_k is the forecast probability of category k.

Since the value of DI cannot exceed that of VI, this value should be interpreted in relation to VI. Discrimination skill is more aptly determined by normalizing it with VI[55]. Normalized discrimination is calculated as follows:

$$\eta^2 = \mathrm{DI}/\mathrm{VI}$$

Next, we assessed Rigor Scores, Brier Scores, and Failure/Success across groups, individuals, genders, forecast difficulty, and years to examine other factors influencing our results. Lastly, two statistical tests were conducted to test our hypothesis. A correlational analysis investigated the relationship between rigor and Brier Scores. Since the Rigor Score is an interval variable, Spearman's rank correlation is suitable[56] for assessing the correlation between rigor and the Brier score. Moreover, we conducted tests to determine if the Rigor Score significantly differs between failure and success cases. The Wilcoxon signed-rank test is used to assess differences between groups. To determine the effect size between groups, we calculate the rank biserial. All calculations were performed in R. The R script and dataset are available in the supporting information[57].

Results

Descriptives of rigor and forecasting accuracy

Table 3 summarizes descriptive statistics. The mean Rigor Score indicates a fair degree of rigor, supported by the median ratings of the underlying criteria: five medians are 'fair' and two are 'good.' However, *Table 3* shows that one criterion has a mean that could be considered below the threshold of 'fair': the use of sources and methodologies (criterion 1). Qualitative data analysis reveals it received a low rating for two primary reasons. First, it might be caused by inaccuracies in the representation of source information in some products. When raters identified this, it directly resulted in a 'poor' rating for sourcing. Moreover, scores may be low because intelligence personnel attempt to conceal details about their sourcing methods to maintain operational security. This impedes raters from assessing this criterion. While assessing source quality without revealing sensitive information presents challenges, sharing such details could theoretically help reassure clients regarding the rigor of the methodology.

[55] Ilan Yaniv, J Frank Yates, and J Keith Smith, "Measures of discrimination skill in probabilistic judgment," *Psychological bulletin* 110, no. 3 (1991), https://doi.org/10.1037/0033-2909.110.3.611.

[56] Patrick Schober, Christa Boer, and Lothar A Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesthesia & analgesia* 126, no. 5 (2018), 10.1213/ANE.0000000000002864.

[57] Manger and van der Weide, "Supplemental materials for "The Effect of Analytical Rigor on Accuracy of Intelligence Forecasts"."

Variable	Mean	Median	SD	SE	Min	Max
Rigor Score	8.221	8.000	2.698	0.254	2.00	15.00
Brier Score	0.185	0.063	0.231	0.023	0.006	0.856
Failure	0.225	0.000	0.420	0.047	0.00	1.00
Occurrence	0.301	0.000	0.461	0.043	0.00	1.00
Criterion 1 (sourcing)	0.646	1.000	0.533	0.050	0.00	2.00
Criterion 2 (uncertainties)	1.327	1.000	0.647	0.061	0.00	2.00
Criterion 3 (assumptions)	0.965	1.000	0.566	0.053	0.00	2.00
Criterion 4 (alternatives)	1.168	1.000	1.085	0.102	0.00	3.00
Criterion 5 (relevance)	0.912	1.000	0.851	0.080	0.00	3.00
Criterion 6 (argumentation)	1.540	2.000	0.613	0.058	0.00	3.00
Criterion 8 (accuracy)	1.664	2.000	0.763	0.072	0.00	3.00

Table 3: *Descriptive statistics*

We use two variables to measure the accuracy of forecasts in our data: the Brier Score and the classification of forecast outcomes into failure and success. To understand what the mean Brier Score (BS) indicates, it is decomposed into the variance index (VI), calibration index (CI), and discrimination index (DI). Our data shows VI = 0.19, signaling a relatively high uncertainty. Our data shows DI = 0.05. The normalized discrimination in our study is 0.24, indicating that discrimination accounts for 24% of the occurrence variance, reflecting limited discrimination skills. The discrimination diagram (*Figure 1*) further illustrates the forecasters' abilities to distinguish occurrence from non-occurrence. It shows that forecasters effectively discriminate for probabilities below 0.5, where the 'false negative-true negative' ratio is about 0.06. However, for probabilities above 0.5, the 'false positives–true positives' ratio is 0.44. The difference in ratios indicates that forecasters are more proficient at predicting non-occurrence than occurrence.

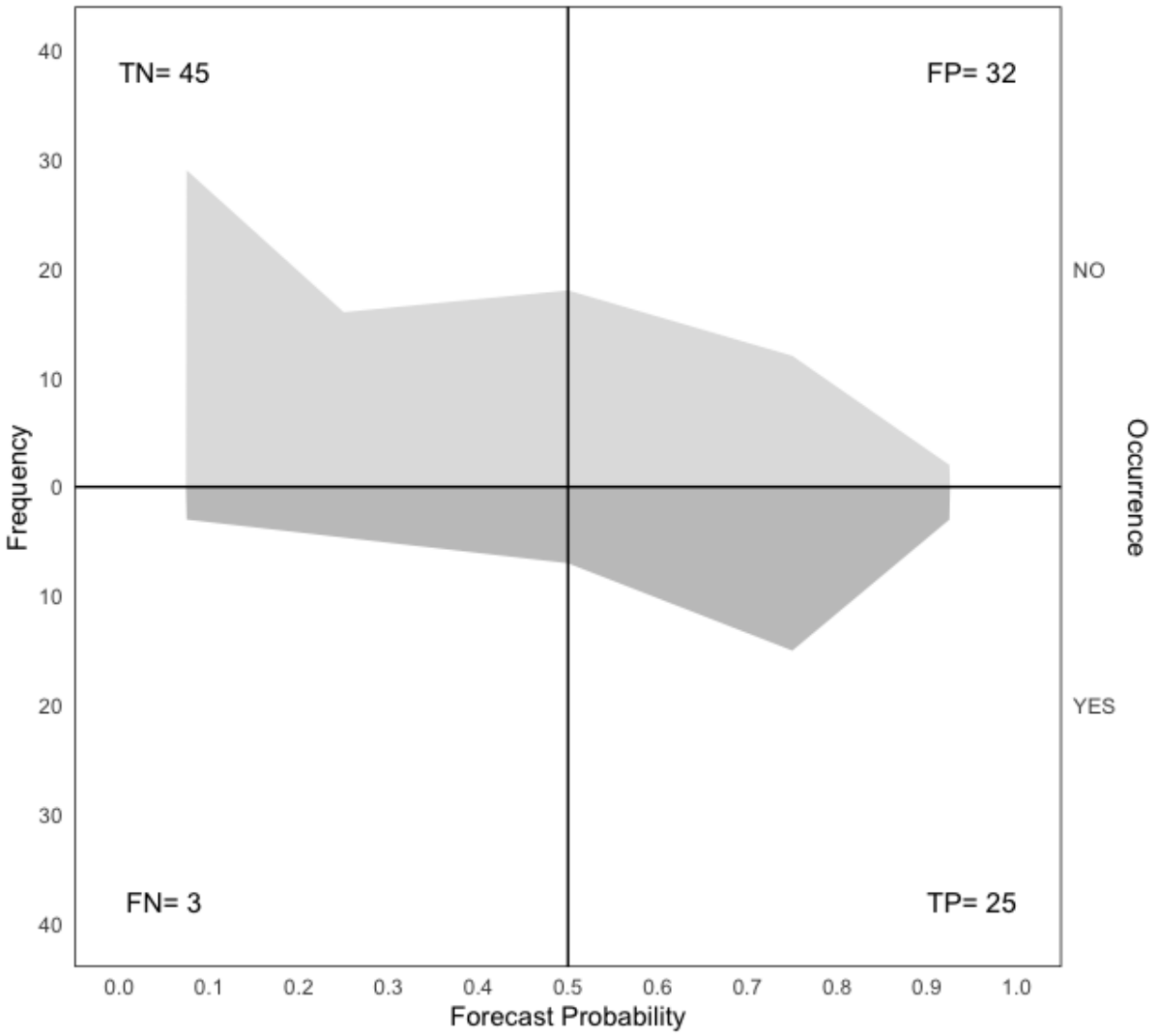


Figure 1: *Discrimination diagram. FN, false negatives; FP, false positives; TN, true negatives; TP, true positives. Illustration generated using the ggplot2*[58] *library in R*

Next, we examine the alignment between predicted probabilities and observed frequencies by calculating the CI. In perfect calibration, CI equals 0; our data shows CI = 0.02, indicating forecasters' calibration skill negatively impacts accuracy marginally. In *Figure 2*, we visualize calibration with the calibration curve of our data, where perfect calibration appears as a 45-degree line, indicating an alignment between predicted probabilities and actual outcomes. Deviations suggest miscalibration, possibly due to overconfidence or underconfidence. The calibration curve falling below the perfect line for forecasts below 0.5 indicates participants' underconfidence in predicting event likelihoods. In contrast, participants show overconfidence for estimates of 0.5 and above.

[58] Hadley Wickham, "Getting Started with ggplot2," *ggplot2: Elegant graphics for data analysis* (2016), https://doi.org/10.1007/978-3-319-24277-4_2.

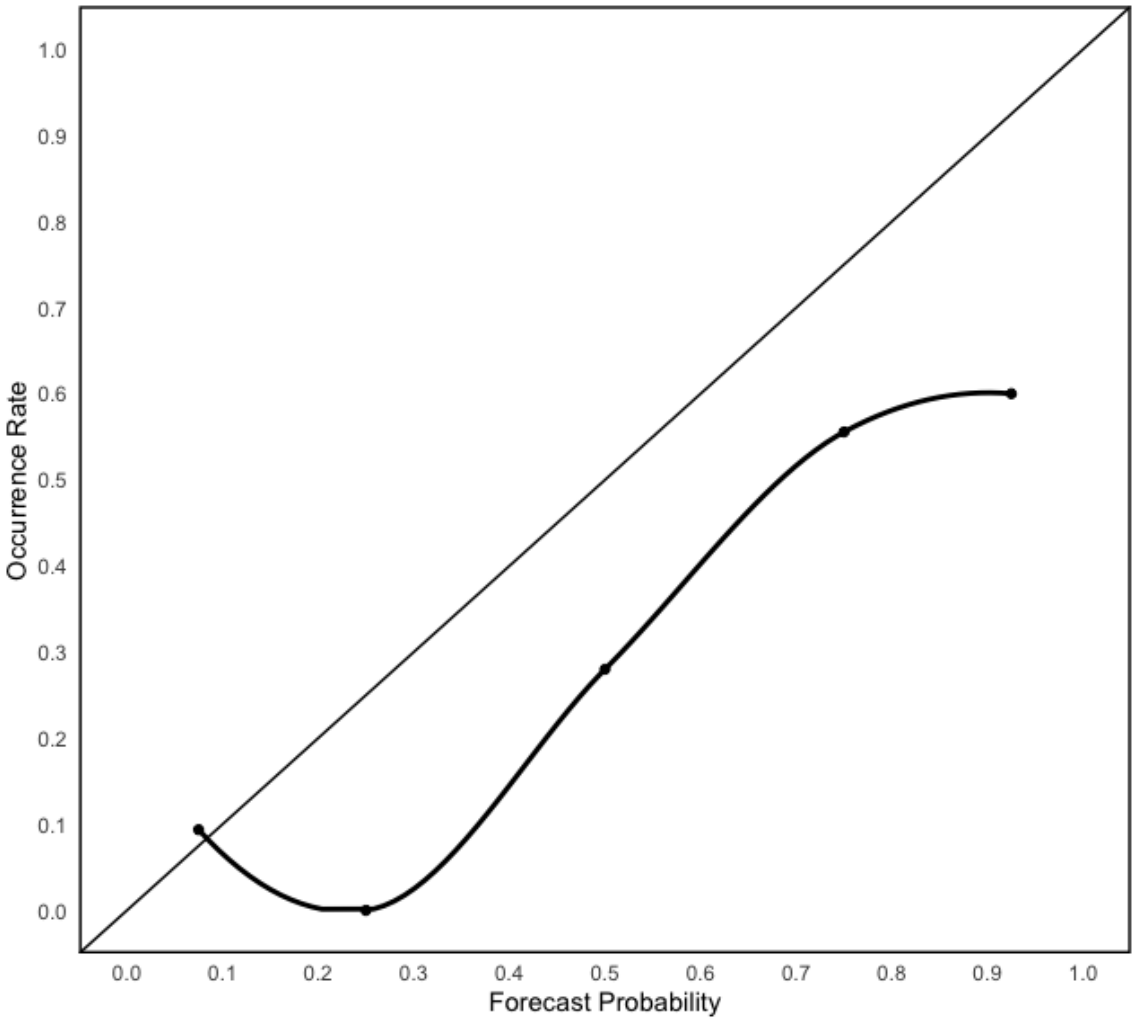


Figure 2: *Model-based calibration curves. Illustration generated using the ggplot2[59] library in R*

The second measure of accuracy is the failure/success distinction. The frequency distribution of prediction failures and successful predictions is shown in *Figure 3*. We observe a substantial presence of 50-50 assessments in our data. Predictions that approximate probabilities of 0.5 provide little value to intelligence consumers unless they convey systemic uncertainty. In such cases, a 50-50 assessment is justified only if it reflects substantial uncertainties that prevent definitive claims about event occurrence. This systemic uncertainty should be observable in the distribution of outcomes. However, the data shows that the systemic uncertainty claimed by these assessments is not fully represented. We expect the mean occurrence rate of 50-50 assessments to be 0.5; it is lower at $\bar{o}_{50-50}$ = 0.28, raising doubts about whether these assessments identified systemic uncertainty. Yet, 50-50 assessments cannot be considered failures or successful predictions. Furthermore, we suspect that the unhelpful nature of 50-50 assessments may be mirrored in the Rigor Score, implying that the degree of rigor of 50-50 assessments is lower than that of prediction failure or success.

[59] Wickham, "Getting Started with ggplot2."

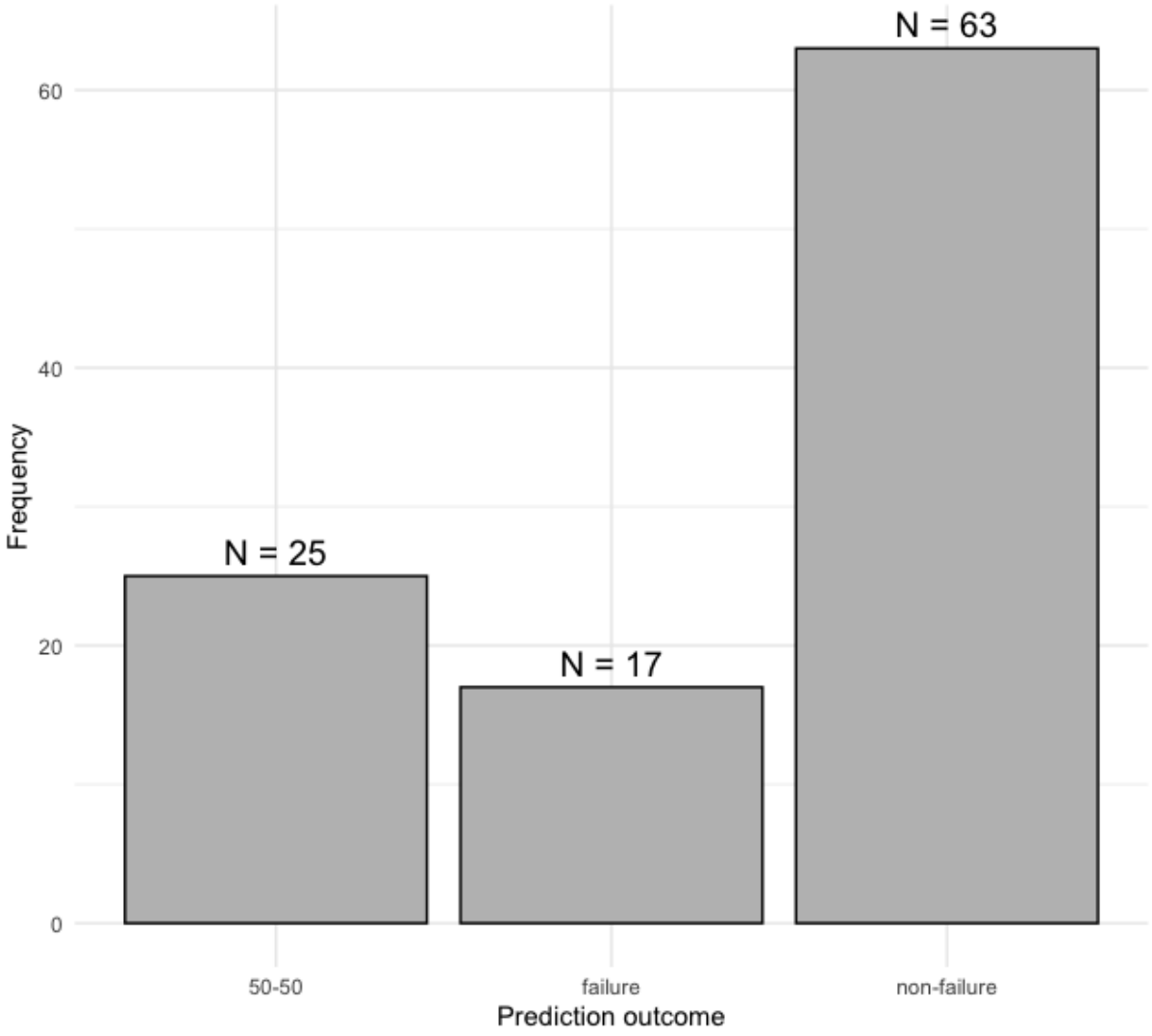


Figure 3: *Frequency distribution Failure, Success, 50-50. Illustration generated using the ggplot2*[60] *library in R*

Next, we assess Rigor Scores, Brier Scores, and Failure/Success across groups, individuals, genders, forecast difficulty, and years. Our tests reveal no significant differences in Rigor Scores between groups or genders. While difficulty does not impact Rigor Scores, Brier Scores, and Failure varied significantly, accuracy is higher for easy forecasts than for difficult ones. This suggests that difficulty may reduce accuracy without affecting process quality. This is supported by Spearman's correlation between Brier Score and Rigor Score, which shows no significant correlation between challenging and easy forecasts. When comparing Rigor Scores across years,

Table 4 shows significant differences in 2021 ($\overline{RS}_{2021}$ = 6.65) and 2023 ($\overline{RS}_{2023}$ = 9.27). This suggests that factors, such as curriculum changes or alterations in instructional team composition, may have influenced variations in Rigor Scores. Both years saw changes in team composition,

[60] Wickham, "Getting Started with ggplot2."

but the curriculum was largely stable. However, the reason for these differences in Rigor Scores cannot be determined with any degree of reliability from our data. When comparing Failure across years, we find that prediction accuracy for 2021 differs significantly ($\bar{F}_{2021}$ = 0.45, p = 0.02). We suspect that the high uncertainty in 2021, as indicated by the occurrence variance of VI_{2021} = 0.245, is the primary cause of the diminished discrimination skill. What causes occurrence variance or uncertainty to change over the years is beyond the scope of this paper.

	Rigor Score		**Brier Score**		**Failure**	
	W	p-value	W	p-value	W	p-value
groups vs. individuals	706.5	0.96	548.5	0.15	380	0.61
gender	628.5	0.87	654.5	0.65	327.5	0.75
difficulty	1015.5	0.74	1238.5	0.04*	648	0.02*
2023	1673	0.003**	1183.5	0.60	730	0.98
2022	1097	0.89	1160	0.54	471	0.28
2021	475.5	0.001***	1149	0.10	750	0.019*
2020	138.5	0.78	106	0.36	61	0.47
2019	493	0.84	349.5	0.16	283	0.43
2018	263.5	0.84	193	0.38	118	0.29

Table 4: *Differences of Brier Score, Rigor Score, and Failure between groups and individuals, gender, forecast difficulty, and years. * p<0.05, ** p<0.01, *** p<0.001*

Relating rigor and forecasting accuracy

Figure 4 presents the correlation table, summarizing the relationship between analytical rigor and accuracy of predicted probabilities. It reveals a significant inverse relationship ($r_s(105) = -0.36$, $p<0.001$), indicating that higher analytical Rigor Scores are associated with lower Brier Scores. Despite the correlation being significant, it is weak to moderate. This provides marginal support for hypothesis 1. Closer examination of the correlation table reveals that four underlying criteria of rigor are unrelated to the accuracy of probability judgments. Firstly, in the description of sourcing and methodologies (criterion 1). The lack of correlation between this criterion and both accuracy measures indicates that descriptions of source quality and methods have little effect on accuracy, as their (mis)representation may not be considered an indicator of the quality of the sources or methods used. Secondly, there is no significant relationship between facts,

assumptions, and judgments (criterion 3) and accuracy measures. We suspect this is partly due to inadequate source referencing, as our data shows a weak but significant correlation between distinguishing between facts, assumptions, and judgments (criterion 3) and source referencing (criterion 1). When sources are described well, it may be easier to separate facts from assumptions, even without clear statements. Third, criterion 4 (alternatives analysis) shows no significant correlation to the Brier Score. This might be explained by the fact that incorporating alternatives into the analysis does not aid in accurately determining event probabilities. Lastly, there is no correlation between the Brier Score and the assessment of relevance and implications (criterion 5). While the reports aimed to address real intelligence questions, participants worked with fictional customers, complicating the relevance assessment.

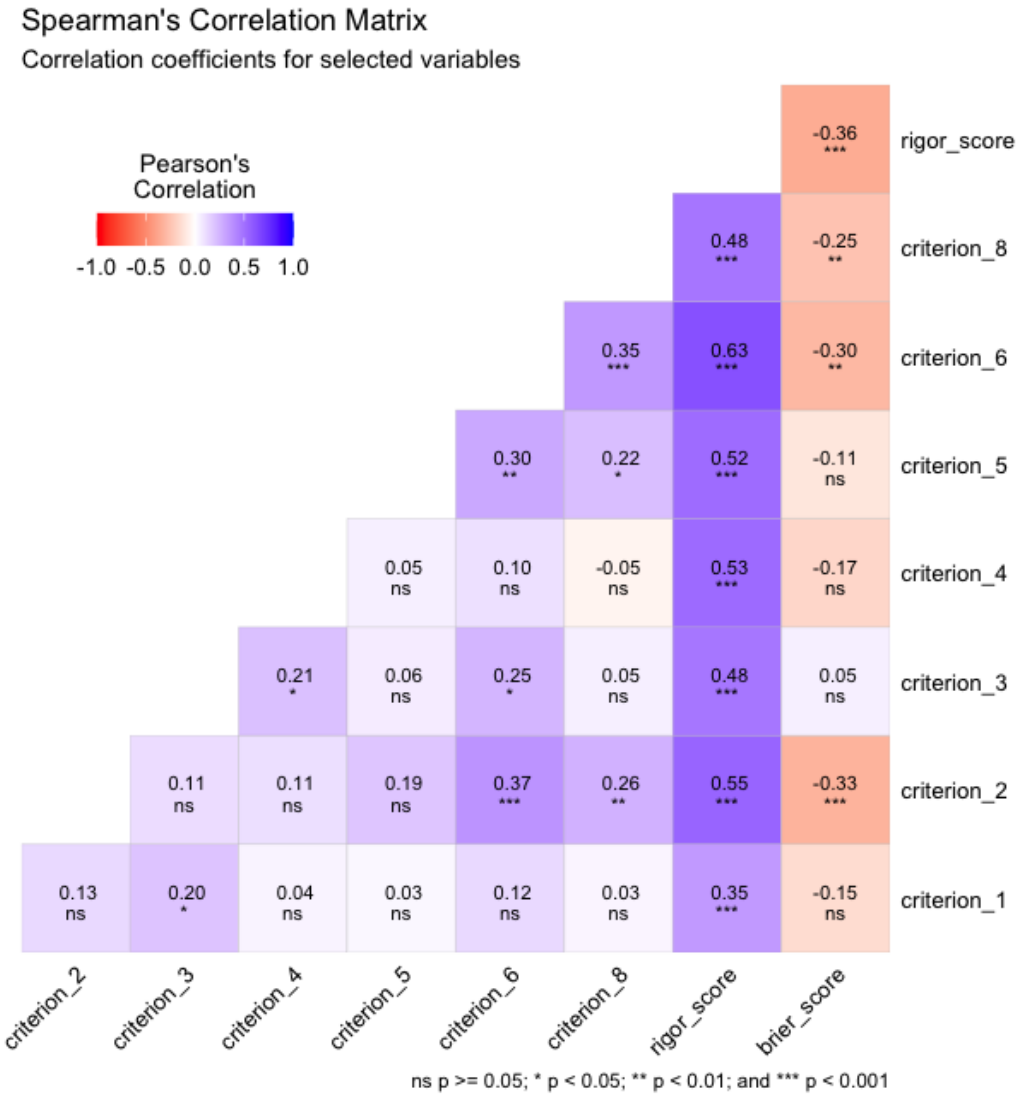


Figure 4: *Correlation matrix. Illustration generated using the metan*[61] *library in R*

[61] Tiago Olivoto and Alessandro Dal'Col Lúcio, "metan: An R package for multi-environment trial analysis," *Methods in Ecology and Evolution* 11, no. 6 (2020), https://doi.org/https://doi.org/10.1111/2041-210X.13384.

Next, we test our second hypothesis, with results summarized in *Table 5*. The hypothesis is explored in three stages. First, we compare the Rigor Scores of failures to those of successful predictions. The Rigor Scores for failures differ significantly from those in successful predictions. More specifically, the results suggest that higher degrees of rigor in intelligence processes are associated with successful prediction in intelligence forecasts. Moreover, the effect size of $r_b = -0.50$ indicates that the relationship between accuracy and rigor becomes stronger when 50-50 assessments are excluded from the data, qualifying this effect size as large[62].

Second, we suspect that the uninformative nature of the 50-50 assessment could be reflected in the Rigor Score, suggesting that the scores from 50-50 assessments differ from those in successful prediction or failure. If 50-50 assessments are associated with poor rigor, this may compound their uninformative nature as they do not reflect systemic uncertainty due to lower levels of rigor. To test this, we compared the Rigor Scores of 50-50 assessments to those of successful predictions. When the rigor in 50-50 assessments resembles the rigor of successful predictions, this should be evident in the Rigor Score. The results indicate that Rigor Scores of 50-50 assessments differ significantly from successful predictions. The effect size ($r_b = -0.32$) is qualified as medium[63]. It indicates that Rigor Scores for 50-50 assessments are lower than those for successful predictions, further underscoring the uninformative nature of 50-50 assessments. Not only are these assessments uninformative, but they also demonstrate less rigor than successful predictions. Lastly, we examine whether the Rigor Scores of 50-50 assessments differed from those of failures. The results show that the Rigor Scores between 50-50 assessments and failures do not differ significantly ($p = 0.26$). It does not indicate that 50-50 assessments should be classified as failures. Nonetheless, our results suggest that the difference in rigor between 50-50 assessments and failures is less pronounced than the differences observed between 50-50 assessments and successful predictions or between failures and successful predictions. Consequently, 50-50 assessments may be uninformative and the result of lower degrees of rigor. Results from the three tests strongly support our second hypothesis, predicting that higher levels of rigor in intelligence processes are associated with successful predictions. Moreover, both failures and 50-50 assessments are associated with similarly low levels of rigor.

	N	W	p-value	Effect size
success vs. failure	80	268	< 0.002***	-0.50
success vs. 50-50	88	599.00	0.01**	-0.32
failure vs. 50-50	42	257	0.26	0.21

Table 5: *Test results for the Wilcoxon test in Rigor Scores between success, failure, and 50-50. * $p<0.05$, ** $p<0.01$, *** $p<0.001$*

[62] Jacob Cohen, "A power primer," *Psychological Bulletin* 112, no. 1 (1992): 157, 10.1037/0033-2909.112.1.155.
[63] Cohen, "A power primer," 157.

Discussion

The results support our hypotheses to varying degrees. While substantial evidence supports the association of rigor in intelligence processes and successful predictions in intelligence forecasts, the correlation between rigor and the precision of probability assessments is weak. When examining the distribution of Rigor Scores across failures and successful predictions, we observe that not all successful predictions have high degrees of rigor. Roughly a quarter exhibit poor rigor. The data show that failure may occur despite high degrees of rigor, and successful predictions occur despite low degrees of rigor. A relatively high degree of poor rigor in successful prediction may indicate two things. First, rigor may not be a reliable predictor of failure. The weak correlation of Brier Scores with the ICD203 rigor scale might support this. However, the weakness of the correlation might also be attributed to the abundance of 50-50 forecasts in the data. Second, the weak correlation between the Rigor and the Brier Score might indicate that the Rating Scale might be inappropriate for non-U.S. intelligence products to assess rigor. ICD203 standards were developed in the U.S. and were never intended for use by others. Although the quality standards may appear universally applicable, this is not necessarily the case. More research applying the Rating Scale to intelligence products of other non-U.S. products is needed to investigate whether ICD203 standards are universally applicable. Nonetheless, the results suggest that the assumption that increased rigor in intelligence processes may positively influence forecasting accuracy holds some validity.

Upon closer examination, the Rigor Scale results appear to align with the findings of Thorburn et al. Their study evaluated analytical rigor across eight dimensions using the ICD203 Rating Scale with a mean score of 10.2 out of a maximum score of 24.[64] While our assessment of rigor focused on seven dimensions, the mean score in our study was 8.22 out of a maximum score of 21. Both studies awarded approximately 40% of the maximum score, with this study slightly below and theirs slightly above. Thorburn's dataset, consisting of fictional reports generated by professional analysts and public participants, may reflect operational practices to some degree due to the involvement of trained professionals; however, it is questionable whether it accurately reflects real-world applications. Similarly, our data consists of reports based on fictional intelligence requirements. The differences in rigor between our study and real intelligence reports remain a subject for further exploration.

Our data shows that, generally, forecasters differentiated fairly between assumptions and facts but often did not explore uncertainties or evaluate the strengths and weaknesses of alternative scenarios. Forecasters performed relatively well in terms of dimensions such as accurate judgments and logical argumentation. These areas likely benefited from targeted training during the training course, including workshops on argument mapping and integrating competing hypotheses into structured arguments.[65] Despite the training emphasizing the need for multiple mutually exclusive hypotheses development in scenario-building methods, many reports poorly

[64] Thorburn et al., "The IC Rating Scale as a Measure of Analytic Rigor."
[65] Ariel Kruger, Luke Thorburn, and Timothy van Gelder, "Using argument mapping to improve clarity and rigour in written intelligence products," *Intelligence and National Security* 37, no. 5 (2022), https://doi.org/10.1080/02684527.2022.2026584.

presented alternatives to the primary scenario. The lack of development and exploration of mutually exclusive hypotheses may be interpreted as a sign of limited critical thinking skills. Additionally, many reports misrepresented sources or included references that could not be traced back to the original material, underscoring the need for a more robust integration of source reliability assessments into the analytical process. Due to the lack of source reliability assessments, the impact of source reliability on accuracy remains unclear. Research is needed to determine whether source quality affects forecasting accuracy. An implication for training programs is that dimensions frequently rated as "poor" may require greater emphasis to address deficiencies. Future research is needed to explore how such changes impact accuracy.

Next, we look to other studies to see how forecasters in our study fare in terms of their ability to make accurate predictions. Two studies may provide a meaningful comparison. First, we compare prediction accuracy to a study by Mellers et al.[66], who examined prediction accuracy in a two-year forecasting tournament. In contrast to our study, participants in their study could update their beliefs as often as they wished before the tournament's close. They found that untrained forecasters had a mean Brier Score of 0.22 in year 1 and 0.23 in year 2 for forecasts made in the tournament's first week. Trained forecasters fared marginally better at 0.20 in year 1 and 0.21 in year 2. Superforecasters, the top 2% performers, did much better with a Brier Score of 0.125 for predictions made in the tournament's first week. This indicates that the accuracy achieved through the training course outperformed the untrained average. Mellers et al. also examined the difference between groups and individuals.[67] In contrast to our findings, they found that accuracy differed significantly between groups and individuals over the entire tournament duration. Next, we can compare our findings with those of a second study. Mandel and Barnes studied the accuracy of intelligence forecasts in the Canadian government's strategic intelligence analysis unit.[68] Forecasters in their sample had a mean Brier Score of 0.07, outperforming forecasters in our study, surpassing Mellers et al's superforecasters. However, we must note that the forecasters in our study are thought to have limited experience. This lack of forecasting experience may lead to lower accuracy in our study. In addition, forecasters in our study were not specifically knowledgeable about the topics of their forecasts, as were analysts in the Canadian study. Another factor that may explain the difference in accuracy is that forecasters in our study made their predictions in a training setting, limiting them to open sources. This starkly contrasts with the Canadian analysts who could utilize their organization's resources. We expect that some of the differences in forecasting accuracy might be explained by the sources and resources available to analysts. Another reason our accuracy results deviate from the Canadian data is that customers held these forecasters accountable for their accuracy. Mandel and Barnes attributed high accuracy to the detailed feedback analysts receive. Other studies have shown that when forecasters are held accountable, they demonstrate superior forecasting accuracy compared to their counterparts who are not held accountable.[69] More specifically, outcome accountability leads to better adaptive performance than process

[66] Barbara Mellers et al., "Psychological strategies for winning a geopolitical forecasting tournament," *Psychological science* 25, no. 5 (2014), 10.1177/0956797614524255.
[67] Mellers et al., "Psychological strategies for winning a geopolitical forecasting tournament."
[68] Mandel and Barnes, "Accuracy of forecasts in strategic intelligence."
[69] Chang et al., "Accountability and adaptive performance under uncertainty: A long-term view."

accountability, with this effect becoming more pronounced over time. Forecasters in our study were not held accountable based on their outcome, but explicitly on their process. This may have negatively affected the accuracy in our study.

This study's main limitation is the use of ICD203 as a measure of rigor. Assessing rigor with standards other than ICD203 might yield different outcomes. Additionally, ICD203 was developed as a policy instrument rather than a scientific measurement tool. This could explain the weak correlation with accuracy measures. However, there are few alternative measurement tools available to assess rigor. As discussed earlier, LOTSA dimensions and Zelnik et al.'s rigor score may serve as potential alternative measurement instruments. Yet, little is known about their validity or reliability. We contend that one method to test the validity or reliability of rigor measures is to examine their relationship to accuracy. Consequently, more empirical research is needed regarding the relationship between rigor and accuracy using alternative measurement instruments for rigor.

A second limitation is the method used to assess the accuracy of predictions. In this study, we were only able to establish accuracy measures for single-event predictions. The methods employed in this study did not allow for the verification of assessments related to intentions and conditional predictions. This meant that more than half (51.8%) of the available assessments were excluded because we were unable to determine whether the prediction occurred within the specified timeframe. On the one hand, this could mean that over half of the assessments were irrelevant. If one cannot establish whether predicted events will occur, what is the sense of predicting them? By extension, these assessments would not be relevant to customers and could be an indicator of inefficiencies in the intelligence process. On the other hand, the exclusion of more than half of the available assessments could be attributed to our methodological inability to determine accuracy for conditional and non-predictive assessments. If this is the case, it should prompt us to develop methods that enable us to detect the occurrence of these types of assessments.

Conclusion

This study explores the relationship between the rigor of intelligence processes and forecasting accuracy. Rigor in intelligence processes refers to the strict adherence to pre-determined methodologies or standards during these processes. Forecasting accuracy refers to the ability to predict future developments. The study aims to contribute to theory and practice in forecasting accuracy and intelligence by examining this relationship. The theoretical contribution is that it explored the unexplored relationship between processes and the accuracy of intelligence forecasts. We found that rigor is only weakly to moderately correlated with the precision of probability assessments. We suspect the large number of 50-50 assessments may have weakened the relationship between rigor and accuracy. Another reason might be that the ICD203 rating scale is not an appropriate instrument to assess rigor in non-U.S. intelligence organizations. In contrast, the results suggest that successful prediction is significantly associated with higher degrees of rigor when determining accuracy through failure or success.

The study aims to provide practitioners with insights on optimizing their processes for greater forecasting accuracy, potentially informing improvements in or the applicability of institutional standards like ICD203 as mechanisms to enhance the quality of outcomes. We conclude that 50-50 assessments are not only uninformative for intelligence consumers but also associated with lower degrees of rigor. In addition, we find that subject matter expertise may positively impact forecasting accuracy. Lastly, the study suggests that assessing rigor can provide insights into how to improve training programs.

Acknowledgments

We would like to thank Robert Beeres and Sebastiaan Rietjens for their feedback on earlier versions of this paper. In addition, we want to thank Milou Cazemier for her research assistance. Lastly, we thank the participants of the December 2024 Professional Military Education conference in Stockholm, Sweden, and the panel members of ISA 2025 in Chicago, U.S., for their insightful remarks.

Bibliography

Arcos, Rubén, and José-Miguel Palacios. "Eu Intcen: A Transnational European Culture of Intelligence Analysis?". *Intelligence and National Security* 35, no. 1 (2019): 72-94. https://doi.org/10.1080/02684527.2019.1649912.

———. "The Impact of Intelligence on Decision-Making: The EU and the Arab Spring." *Intelligence and National Security* 33, no. 5 (2018/07/29 2018): 737-54. https://doi.org/10.1080/02684527.2018.1434449.

Ard, Michael J. "Structured Analytic Techniques: A Pragmatic Approach." *International Journal of Intelligence and CounterIntelligence* (2023): 1-17. https://doi.org/10.1080/08850607.2023.2241308.

Artner, Stephen J, Richard S Girven, and James Bruce. *Assessing the Value of Structured Analytic Techniques in the U.S. Intelligence Community.* RAND, 2017.

Atanasov, Pavel, Jens Witkowski, Lyle H. Ungar, Barbara A. Mellers, and Philip E. Tetlock. "Small Steps to Accuracy: Incremental Belief Updaters Are Better Forecasters." *Proceedings of the 21st ACM Conference on Economics and Computation* (2020). https://doi.org/10.1145/3391403.33995.

Barnett, A, T Primoratz, R de Rozario, M Saletta, L Thorburn, and T van Gelder. "Analytic Rigour in Intelligence." Melbourne, Australia: Hunt Laboratory for Intelligence Research, University …, 2021.

Betts, Richard K. "Analysis, War, and Decision: Why Intelligence Failures Are Inevitable." *World politics* 31, no. 1 (1978): 61-89. https://doi.org/10.2307/2009967.

Borg, Lars C. "Reduced Analytic Uncertainty through Increased Analytic Rigour: Effects of Using Structured Analytic Techniques in Estimative Intelligence." Brunel University London, 2022.

Breakspear, Alan. "A New Definition of Intelligence." *Intelligence and National Security* 28, no. 5 (2013): 678-93. https://doi.org/10.1080/02684527.2012.699285.

Brier, Glenn W. "Verification of Forecasts Expressed in Terms of Probability." *Monthly weather review* 78, no. 1 (1950): 1-3.

Brouwer, P.A. , and M. Scholten. *Joint Doctrine Publicatie 2 Inlichtingen*. The Hague, 2012.

Calista, Donald J. "Enduring Inefficiencies in Counterintelligence by Reducing Type I and Type Ii Errors through Parallel Systems: A Principal-Agent Typology." *International Journal of Intelligence and CounterIntelligence* 27, no. 1 (2014/03/01 2014): 109-31. https://doi.org/10.1080/08850607.2014.842809.

Chang, Welton. "Getting It Right Assessing the Intelligence Community's Analytic Performance." *American Intelligence Journal* 30, no. 2 (2012): 99-108. http://www.jstor.org/stable/26202020.

Chang, Welton, Pavel Atanasov, Shefali Patil, Barbara A Mellers, and Philip E Tetlock. "Accountability and Adaptive Performance under Uncertainty: A Long-Term View." *Judgment and Decision Making* 12, no. 6 (2017): 610-26. https://doi.org/10.1017/S1930297500006732.

Cohen, Jacob. "A Power Primer." *Psychological Bulletin* 112, no. 1 (1992). 10.1037/0033-2909.112.1.155.

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): 933-48. https://doi.org/10.1080/2684527.2016.1140327.

Coulthart, Stephen J. “An Evidence-Based Evaluation of 12 Core Structured Analytic Techniques.” *International Journal of Intelligence and CounterIntelligence* 30, no. 2 (2017/04/03 2017): 368-91. https://doi.org/10.1080/08850607.2016.1230706.

Eiran, Ehud. “The Three Tensions of Investigating Intelligence Failures.” *Intelligence and National Security* 31, no. 4 (2016/06/06 2016): 598-618. https://doi.org/10.1080/02684527.2015.1044293. https://doi.org/10.1080/02684527.2015.1044293.

Gentry, John A. “Has the Odni Improved U.S. Intelligence Analysis?”. *International Journal of Intelligence and CounterIntelligence* 28, no. 4 (2015/10/02 2015): 637-61. https://doi.org/10.1080/08850607.2015.1050937.

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

Horowitz, Michael, Brandon M Stewart, Dustin Tingley, Michael Bishop, Laura Resnick Samotin, Margaret Roberts, Welton Chang, Barbara Mellers, and Philip Tetlock. “What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance at Geopolitical Forecasting.” *The Journal of Politics* 81, no. 4 (2019): 1388-404. https://doi.org/10.1177/1555343414554702.

“Aggregative Contingent Estimation (Ace).” 2011, accessed 04 NOV 2024, 2024, https://www.iarpa.gov/research-programs/ace.

Kajdasz, James E. “A Demonstration of the Benefits of Aggregation in an Analytic Task.” *International Journal of Intelligence and CounterIntelligence* 27, no. 4 (2014): 752-63. https://doi.org/10.1080/08850607.2014.924814.

Kent, Sherman. *Strategic Intelligence for American World Policy*. Princeton, N.J.: Princeton University Press, 1966.

Kruger, Ariel, Luke Thorburn, and Timothy van Gelder. “Using Argument Mapping to Improve Clarity and Rigour in Written Intelligence Products.” *Intelligence and National Security* 37, no. 5 (2022): 607-26. https://doi.org/10.1080/02684527.2022.2026584.

Lowenthal, Mark M., and Ronald A. Marks. “Intelligence Analysis: Is It as Good as It Gets?”. *International Journal of Intelligence and CounterIntelligence* 28, no. 4 (2015/10/02 2015): 662-65. https://doi.org/10.1080/08850607.2015.1051410.

Mandel, David R. “Accuracy of Intelligence Forecasts from the Intelligence Consumer’s Perspective.” *Policy Insights from the Behavioral and Brain Sciences* 2, no. 1 (2015): 111-20. https://doi.org/10.1177/2372732215602907.

Mandel, David R., and Alan Barnes. “Accuracy of Forecasts in Strategic Intelligence.” *Proceedings of the National Academy of Sciences* 111, no. 30 (2014): 10984-89. https://doi.org/10.1073/pnas.1406138111.

———. “Geopolitical Forecasting Skill in Strategic Intelligence.” *Journal of Behavioral Decision Making* 31, no. 1 (2018): 127-37. https://doi.org/10.1002/bdm.2055.

Mandel, David R., and Daniel Irwin. “Tracking Accuracy of Strategic Intelligence Forecasts: Findings from a Long-Term Canadian Study.” *Futures & Foresight Science* 3, no. 3-4 (2021). https://doi.org/10.1002/ffo2.98.

Manger, Gideon. “Unravelling Effectiveness in Intelligence: A Systematic Review.” *Intelligence and National Security* 39, no. 7 (2024): 1136-57. https://doi.org/10.1080/02684527.2024.2370132.

Manger, Gideon, and Sanne van der Weide. 2025. “Supplemental Materials for "The Effect of Analytical Rigor on Accuracy of Intelligence Forecasts”.” DataverseNL. https://doi.org/10.34894/JLKINY

Marcoci, Alexandru, Mark Burgman, Ariel Kruger, Elizabeth Silver, Marissa McBride, Felix Singleton Thorn, Hannah Fraser, *et al.* “Better Together: Reliable Application of the Post-9/11 and Post-Iraq Us Intelligence Tradecraft Standards Requires Collective Analysis.” *Frontiers in psychology* 9 (2019): 2634. https://doi.org/10.3389/fpsyg.2018.02634.

Marcoci, Alexandru, Ans Vercammen, and Mark Burgman. “Odni as an Analytic Ombudsman: Is Intelligence Community Directive 203 up to the Task?”. *Intelligence and National Security* 34, no. 2 (2019/02/23 2019): 205-24. https://doi.org/10.1080/02684527.2018.1546265.

Mellers, Barbara A., Joshua D. Baker, Eva Chen, David R. Mandel, and Philip E. Tetlock. “How Generalizable Is Good Judgment? A Multi-Task, Multi-Benchmark Study.” *Judgment and Decision Making* 12, no. 4 (2017): 369-81. https://doi.org/https://doi.org/10.1017/S1930297500006240.

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, *et al.* “Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions.” *Perspectives on Psychological Science* 10, no. 3 (2015): 267-81. https://doi.org/10.1177/1745691615577794.

Mellers, Barbara, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott*, et al.* "Psychological Strategies for Winning a Geopolitical Forecasting Tournament." *Psychological science* 25, no. 5 (2014): 1106-15. 10.1177/0956797614524255.

Mettler, Stephen. "Return of the Bear Learning from Intelligence Analysis of the Ussr to Better Assess Modern Russia." *American Intelligence Journal* 35, no. 2 (2018): 33-41.

ODNI. "Intelligence Community Directive (Icd) 203- Analytic Standards." edited by Office of the Director of National Intelligence, 2022. https://www.dni.gov/files/documents/ICD/ICD-203.pdf.

———. "Rating Scale for Evaluating Analytic Tradecraft Standards." (2007).

Olivoto, Tiago, and Alessandro Dal'Col Lúcio. "Metan: An R Package for Multi-Environment Trial Analysis." *Methods in Ecology and Evolution* 11, no. 6 (2020): 783-89. https://doi.org/https://doi.org/10.1111/2041-210X.13384.

Poore, Joshua C., Clifton Forlines, Sarah Miller, John Regan, and John M. Irvine. "Personality, Cognitive Style, Motivation, and Aptitude Predict Systematic Trends in Analytic Forecasting Behavior." *Journal of Cognitive Engineering and Decision Making* 8 (2014): 374 - 93. https://doi.org/10.1177/1555343414554702.

Schober, Patrick, Christa Boer, and Lothar A Schwarte. "Correlation Coefficients: Appropriate Use and Interpretation." *Anesthesia & analgesia* 126, no. 5 (2018): 1763-68. 10.1213/ANE.0000000000002864.

Tetlock, P. E., and B. A. Mellers. "Intelligent Management of Intelligence Agencies: Beyond Accountability Ping-Pong." *American Psychologist* 66, no. 6 (2011): 542-54. https://doi.org/10.1037/a0024285.

Tetlock, Philip E. *Expert Political Judgment.* Princeton: Princeton University Press, 2017. https://doi.org/10.1515/9781400888818.

Tetlock, Philip E., and Dan Gardner. *Superforecasting: The Art and Science of Prediction.* London: Random House Business, 2019.

Thorburn, Luke, Timothy van Gelder, Christine Shahan Brugh, Sue Mi Kim, Michele Kolb, and Barry Zulauf. "The Ic Rating Scale as a Measure of Analytic Rigor." *Available at SSRN 4519926* (2022).

Vrist Rønn, Kira. "(Mis-) Informed Decisions? On Epistemic Reasonability of Intelligence Claims." *International Journal of Intelligence and CounterIntelligence* 27, no. 2 (2014): 351-67. https://doi.org/10.1080/08850607.2014.842813.

Wickham, Hadley. "Getting Started with Ggplot2." *ggplot2: Elegant graphics for data analysis* (2016): 11-31. https://doi.org/10.1007/978-3-319-24277-4_2.

Winkler, R. L., Javier Muñoz, José L. Cervera, José M. Bernardo, Gail Blattenberger, Joseph B. Kadane, Dennis V. Lindley*, et al.* "Scoring Rules and the Evaluation of Probabilities." *Test* 5, no. 1 (1996/06/01 1996): 1-60. https://doi.org/10.1007/BF02562681.

Wirtz, James J. "The Art of the Intelligence Autopsy." *Intelligence and National Security* 29, no. 1 (2014): 1-18. https://doi.org/10.1080/02684527.2012.748371.

Yaniv, Ilan, J Frank Yates, and J Keith Smith. "Measures of Discrimination Skill in Probabilistic Judgment." *Psychological bulletin* 110, no. 3 (1991): 611-17. https://doi.org/10.1037/0033-2909.110.3.611.

Yates, J. , and Frank. "External Correspondence: Decompositions of the Mean Probability Score." *Organizational behavior and human performance.* 30, no. 1 (1982): 132-56. https://doi.org/10.1016/0030-5073(82)90237-9.

Zelik, Daniel J, Emily S Patterson, and David D Woods. "Measuring Attributes of Rigor in Information Analysis." In *Macrocognition Metrics and Scenarios: Design and Evaluation for Real-World Teams*, edited by E. S. Patterson & J. Miller, 65-84. Aldershot, UK: Ashgate: CRC Press, 2010.

Zulauf, Barry. "From a Former Odni Ombudsperson Perspective: Safeguarding Objectivity in Intelligence Analysis." *Studies in Intelligence* 63, no. 3 (2021): 35-39.

JOURNAL *of* EUROPEAN *and* AMERICAN INTELLIGENCE STUDIES

An international peer-reviewed journal

Sponsored by the **Research Institute *for* European *and* American Studies – RIEAS**

Under the editorial direction of the

School of Law and Government, Dublin City University

and the **International Centre for Policing & Security, University of South Wales**

CALL FOR PAPERS

The *Journal of European and American Intelligence Studies* (*JEAIS*) is seeking papers focusing on the field of intelligence and related areas of study and practice, such as terrorism and counter-terrorism, domestic and international security, geopolitics, and international relations. The papers should contain or examine original research on the broader European and American practice and study of intelligence, but also highlight intelligence themes from other regions of the world, to include Africa and Asia, as well as Oceania. Submissions will undergo rigorous and highly selective screening, as well as an exhaustive review process. Particular attention will paid by the editors to papers that discuss and challenge established or emerging ideas, address existent knowledge gaps by advancing new knowledge on intelligence-related topics, and examine intelligence scholarship beyond traditional communities of research.

Relevant Topics Include:

international intelligence cooperation ▪ intelligence operations ▪ intelligence analysis ▪ inter-agency cooperation ▪ practitioner-policymaker relations ▪ military intelligence ▪ politicization of intelligence ▪ counterintelligence ▪ intelligence and sub-state violence ▪ intelligence, race, and gender ▪ intelligence and privacy ▪ intelligence and democracy ▪ intelligence ethics

Abstracts of up to 250 words may be submitted to:
secretary@rieas.gr and *christian.kaunert@southwales.ac.uk*

Deadline for Abstract Submissions: Monday February 16, 2026
Deadline for Paper Submissions: Monday May 4, 2026

Instructions for Authors

- Submitted manuscripts must be maximum 8,000 words, excluding a 250-word abstract (required) and any footnotes, as well as references. Manuscripts that exceed the word limit will be automatically rejected and returned to their authors.

- Manuscripts will be accepted for submission and evaluation with the understanding that their content is unpublished, original work by their authors, and have not been submitted for publication elsewhere.

- All accepted manuscripts and artwork become the property of the publisher, which is the Research Institute for European and American Studies (RIEAS).

- The entirety of manuscripts, including title page, abstracts, tables, legends, and references, should be typewritten and submitted in a Word-type file. No portable document format (PDF) documents will be accepted.

- Submissions should be 1.5-spaced and use Times New Roman size 12 as their standard font.

- All margins should be at least one inch in length, and all pages should be numbered consecutively throughout the manuscript.

- Titles must be as brief and clear as possible. On the title page, please include full names of authors, their academic and/or other professional affiliations, their contact information (including email accounts) and their complete mailing address for correspondence.

- All references should be numbered consecutively and listed as footnotes at the end of every page. In the text, references should be cited by a superior character of the corresponding number.

- For further information on writing style, consult *The Chicago Manual of Style*, 17th edition.