Alexandru Marcoci, Ans Vercammen and Mark Burgman]

ODNI as an analytic ombudsman: is Intelligence Community Directive 203 up to the task?

Article (Accepted version) (Refereed)

ODNI as an analytic ombudsman: Is Intelligence Community Directive 203 up to the task?

Alexandru Marcoci, Ans Vercammen and Mark Burgman

Forthcoming in *Intelligence and National Security*

Abstract. In the wake of 9/11 and the assessment of Iraq's WMD, several inquiries placed the blame primarily on the Intelligence Community. Part of the reform that followed was a codification of analytic tradecraft standards into *Intelligence Community Directive (ICD) 203* and the appointment of an analytic ombudsman in the newly created Office of the Director of National Intelligence charged with monitoring the quality of analytic products from across the intelligence community. In this paper we identify three assumptions behind ICD203: (1) tradecraft standards can be employed consistently; (2) tradecraft standards sufficiently capture the key elements of good reasoning; (3) good reasoning leads to more accurate judgments. We then report on two controlled experiments that uncover operational constraints in the reliable application of the ICD203 criteria for the assessment of intelligence products. Despite criticisms of the post-9/11 and post-Iraq reform, our results highlight that ICD203, properly applied, holds potential to improve precision and accountability of intelligence processes and products.

1. Introduction

In his Preface to Roberta Wohlstetter's *Pearl Harbor: Warning and Decision*, Thomas Schelling writes:

> Surprise, when it happens to a government, is likely to be a complicated, diffuse, bureaucratic thing. It includes neglect of responsibility, but also responsibility so poorly defined or so ambiguously delegated that action gets lost. It includes gaps in intelligence, but also intelligence that, like a string of pearls too precious to wear, is too sensitive to give to those who need it. It includes the alarm that fails to work, but also the alarm that has gone off so often it has been disconnected … finally, as at Pearl Harbor surprise may include some measure of genuine novelty introduced by the enemy, and possibly some sheer bad luck." [1]

Despite Schelling's analysis of strategic surprise as a state failure, in the wake of 9/11 and the October 2002 National Intelligence Estimate (NIE) on *Iraq's Continuing Programs for Weapons of Mass Destruction*,[2] several inquiries have placed the blame primarily on the Intelligence Community (IC) and called for fundamental reform. Part of the analytic transformation that followed was a codification of the analytic standards into a policy document – Intelligence Community Directive (ICD) 203 – and the appointment of an analytic ombudsman in the newly created Office of the Director of National Intelligence (ODNI) charged with ensuring that intelligence reports from across the IC comply with these standards. ICD203 (in its second, January 2, 2015 iteration) contains four analytic standards: objectivity, political independence, timeliness and good tradecraft. The last is broken down into nine analytic tradecraft standards: (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; and (9) Incorporates effective visual information where appropriate.

Commenting on the reforms, Robert Cardillo, who served in ODNI as Deputy Director of National Intelligence for Intelligence Integration, wrote that ICD203 "injects rigor into our processes and products and holds analysts and managers accountable for results".[3] However, this position hinges on whether the analytic tradecraft standards can be operationalised effectively. Yet, to date and as far as we are aware, no empirical investigation of these standards has been conducted. Thus, despite the optimism shown by Cardillo and others, without explicit testing, the effectiveness of ICD203 in the promotion of analytic excellence remains a potentially perilous assumption. This paper represents the first empirical investigation into the ICD203 criteria, providing insight into the assumptions that underlie their effective operationalisation, and the conditions under which these assumptions may not hold.

In this study we draw on a set of hypothetical reasoning problems that emulate many of the challenges that arise in real intelligence problems. These problems have been developed by the Intelligence Advanced Research Programs Activity (IARPA) as part of a multi-year program on improving the reasoning of teams of intelligence analysts. The first stage of this program involves experimental assessments of teams' abilities to solve constrained reasoning problems – that is, problems whose descriptions

contain all the information required to solve them. The problems have a normative standard against which the quality of the solution can be evaluated. This paper comprises a first fundamental step towards establishing whether ICD203 can be an effective instrument in the assessment of the quality of reasoning of analytic reports, initially focusing on constrained problems which enable a stricter level of experimental control.

Our focus is exclusively on the evaluation of intelligence products, rather than on their production. As such, the purpose of the paper is to evaluate how ICD203 supports ODNI in its mandated role of analytic ombudsman. As we reveal in our discussion, the results have relevance for the wider application of ICD203.

We begin with a theoretical investigation of ICD203 and its aims. We identify three assumptions behind the optimistic discourse surrounding ICD203: (1) that the standards can be used consistently by different users (i.e. evaluators of analytic reports); (2) that scoring highly on the analytic tradecraft standards correlates with good reasoning; and (3) that good reasoning correlates with accurate judgment. We begin by asking whether the analytic tradecraft standards can be followed in a consistent (i.e. reliable) manner by different evaluators and whether training enhances reliability. We consider the experts involved in the development and testing of the problems used in the IARPA study to provide an external benchmark of good reasoning with respect to the problems they developed. Then, by comparing evaluations made using ICD203 with their evaluations, we establish whether the application of the tradecraft standards conforms to an external, independent reference point for good reasoning. Our study thus aims to explore the first two assumptions, leaving the third assumption for a future study.

In this paper we argue that:

i. Both inter-rater agreement and correlation with expert judgments are sensitive to and may be substantially improved by training.
ii. Evaluations based on application of ICD203 are at least *fairly reliable* and that even novice evaluators' judgments guided by the analytic tradecraft standards are at least *moderately correlated* with expert judgments of the same analytical reports (the italicized terms are technical and they will be explained in more detail in Section 6).
iii. The scoring system used to aggregate the evaluators' judgments on the individual tradecraft standards also contributes importantly to measures of validity and consistency. This points to an often-overlooked issue: that determining whether tradecraft standards engender good reasoning is partly a function of how well-calibrated evaluators are and what weight one assigns to the different standards.

The three points above are substantive and novel in the literature on analytic standards that we review in this paper.

First, the analytic standards contained in ICD203 are framed in natural language and are deployed in the assessment of intelligence products in the form of an assessment rubric. The large literature on the use of assessment rubrics in (higher) education conclusively shows the benefits of evaluator training and calibration.[4] Nevertheless, to date, no study has shown the impact of training on application of intelligence analytic standards. We compare expert evaluations with the results of a baseline experiment (without evaluator training) and an intervention (with basic evaluator training) in terms of reliability and correlation. This accomplishes two important

goals: it establishes that findings from the educational assessment literature carry over to the use of assessment rubrics in intelligence analysis, and it provides a baseline for the expected effect training can have on these kinds of rubrics. In other words, future studies might examine alternative, more comprehensive training methods to assess whether this training pays off in comparison to the basic training we provided our evaluators in this study.

Second, we show that with minimal training (to be defined more precisely, below) assumptions (1) and (2) are warranted: evaluators become more reliable in using the rubric and their judgments correlate more closely to the judgments of experts. This suggests that it may be the right kind of instrument for evaluating the quality of intelligence reports in the IC.

We stop short of saying that ICD203 actually is the right instrument for two reasons: as we discuss below, the results show that there may still be some need for revision in the way the standards are defined. Moreover, this study does not address accuracy (assumption (3)), and accuracy is essential for determining whether ICD203 can aid ODNI in performing its role as analytic ombudsman. Nevertheless, the results of this study offer a second benefit: they provide a benchmark in terms of reliability and correlation with good reasoning. In other words, we now know what a revised list of analytic standards has to outperform.[5]

Third, this study draws attention to an oft-overlooked problem. Any research on an assessment rubric will be sensitive to the scoring method used for that rubric and we can only talk about reliability, correlations to expert judgments and accuracy with specific reference to that method. We have not found any discussion of this important point in the literature. To illustrate its importance, we investigate two different

scoring methods in this paper. The significant divergence of the results we obtain should be enough to convince the IC that this is an important, open question that warrants immediate additional research.

This paper is organised as follows. In section 2 we introduce the controversies surrounding ICD203 and explain why its contribution to the quality of reasoning in intelligence analysis is treated by some scholars with skepticism. In section 3 we unpack the three assumptions outlined above that underpin ICD203 and we explore the issues they engender. In section 4 we introduce the operationalisation of ICD203 used by ODNI. In section 5 we present the IARPA project from which this study draws its analytic reports. Section 6 presents the materials, methods and results of the two experiments we conducted on ICD203. We conclude with a general discussion in section 7.

2. Background

Vital government actions hinge on intelligence analyses. In the wake of the failure to foresee the 9/11 terrorist attacks on the US and the misreading of Iraq's capability to deploy weapons of mass destruction, several inquiries have requested a reform of the standards of analysis and a strengthening of quality control processes employed by the intelligence community.[6] In consequence, the 2004 *Intelligence Reform and Terrorism Prevention Act*[7] (IRTPA) identified the need for the creation and adoption of a list of analytic tradecraft standards focused on:

> whether the product or products concerned were based on all sources of available intelligence, properly describe the quality and reliability of underlying sources,

> properly caveat and express uncertainties or confidence in analytic judgments, properly distinguish between underlying intelligence and the assumptions and judgments of analysts, and incorporate, where appropriate, alternative analyses.[8] (Section 1019.b.2.A)

Furthermore, IRTPA called for the creation of an ombudsman for analytic integrity whose mandate was to ensure that "finished intelligence products produced […] are timely, objective, independent of political considerations, based upon all sources of available intelligence, and employ the standards of *proper analytic tradecraft*" (Section 1019.a, our emphasis). The Office of the Director of National Intelligence (ODNI) was created and within it an Office for Analytic Integrity and Standards (AIS) run by a Deputy Director of National Intelligence for Analysis (DDNI/A).[9] Under the supervision of the first DDNI/A, Thomas Fingar, the standards listed in Section 1019.b.2.A were developed and published as Intelligence Community Directive (ICD) 203, which in turn formed part of the foundation of what Fingar called an "Analytic Transformation".[10] The term has been used to cover several transformations under the guidance of ODNI including increased collaboration between the different agencies comprising the intelligence community (IC), rotations among agencies, the introduction of a basic course for all analysts (Analysis 101), the evaluation of personnel, the monitoring of the analytic standards of ICD203 throughout the IC, and annual reviews of analytic products against the standards laid out in ICD203. In this paper we primarily focus on this last aspect of the analytic transformation.

The analytic tradecraft standards enshrined in ICD203 are *prima facie* uncontroversial.[11] Take the idea that analytic products should "properly caveat and express uncertainties or confidence in analytic judgments."[12] According to a well-

known mantra,[13] the role intelligence is to eliminate (or at least reduce) uncertainty[14] and the discussion of how uncertainties should be caveated, expressed and distinguished from confidence judgments goes back to at least Sherman Kent[15]. Yet, several authors have claimed that ICD203 genuinely transformed analytical practice. For instance, Immerman claims that "[i]n a remarkably brief time … intelligence analysis has experienced genuine reform, some of which is radical and even revolutionary."[16] Cardillo believes one of the main pillars of this reform is the promulgation of ICD203, which "codified good analytic tradecraft — much discussed but seldom formally documented in the 50-year history of the IC"[17] and is confident "[g]reat progress has been achieved".[18] Gentry also reports a generally positive view of ICD203 and the work of AIS among intelligence officials he interviewed[19] but notes that younger analysts seem to be more enthusiastic about the reforms.

Despite these affirmations, critics take their cue from Betts's classic analysis of intelligence failures:

> [i]n the best-known cases of intelligence failure, the most crucial mistakes have seldom been made by collectors of raw information, occasionally by professionals who produce finished analyses, but most often by the decision makers who consume the products of intelligence services. Policy premises constrict perception, and administrative workloads constrain reflection. Intelligence failure is political and psychological more often than organizational… [o]bservers who see notorious intelligence failures as egregious often infer that disasters can be avoided by perfecting norms and procedures for analysis and argumentation. This belief is illusory. Intelligence can be improved marginally, but not radically, by altering the analytic system.

> The illusion is also dangerous if it abets overconfidence that systemic reforms will increase the predictability of threats.[20]

Their main claim is that the failures of 9/11 and the October 2002 NIE were due to policy makers rather than analysts. For instance, Gentry unpacks the many ways in which a state can fail to produce and utilize intelligence to defend itself against strategic surprises. Out of the six categories of intelligence-related state failures, only two are related to analysis. Gentry argues that none of them apply to Pearl Harbour or 9/11. Instead he focuses on organizational and managerial reform, suggesting incentives for increased collaboration between agencies, and better communication to policy makers and the public of the possibilities and limitations of intelligence analysis.[21]

Immerman[22] and Phythian[23] delve deeper into the history of the October 2002 NIE and suggest that the politics of intelligence agencies played a more substantial role than analysts' tradecraft errors. One of the most striking examples Phythian gives relates to Curve Ball, arguably the most important source for the case that Iraq had a mobile biological weapons production program. According to the US Senate Intelligence Committee report, the only US intelligence official to personally make contact with Curve Ball contacted the CIA and raised concerns regarding his credibility.[24] In reply, the deputy chief of the CIA's Iraqi Task Force wrote: "[L]et's keep in mind the fact that this war's going to happen regardless of what Curve Ball said or didn't say, and that the Powers That Be probably aren't terribly interested in whether Curve Ball knows what he's talking about. However, in the interest of Truth, we owe somebody a sentence or two of warning, if you honestly have reservations."[25]

Amy Zegart also places the blame primarily on policy-makers. She argues that the failure of the IC to adapt to the new threats emerging after the end of the Cold War is primarily due to "the nature of bureaucratic organizations, … the self-interest of presidents, legislators, and government bureaucrats, … and the fragmented structure of the federal government".[26] In a different paper, Zegart surveyed all intelligence reform committees before 9/11 and found that many of the recommendations they made had not been acted on prior to 9/11 due to politics. For instance, 28% of all recommendations for reform pre-9/11 focused on the need for more collaboration between the different elements of the IC and the Government. The biggest challenge in putting this recommendation into practice, according to Zegart, has been the fact that the official responsible for setting "broad strategies" and coordinating "efforts across [the IC]" - the Director of the Central Intelligence Agency - "held direct control over only 15 percent of the intelligence budget (the Secretary of Defense controlled the rest) and had weak management authority for allocating money, people, and programs to every agency outside the CIA".[27] The main message from Zegart's analysis is that the IC recognized the relevant threats and warned decision-makers of them but that decision-makers failed to act accordingly.

This is not to say that the failure to predict the 9/11 attacks and to correctly assess Iraq's WMD capabilities weren't, in part, due also to "intelligence failures". However, in the literature on ICD203 we surveyed for the purposes of this paper, the overall consensus seemed to be that the two failures that lead to the adoption of the new standards can *primarily* be explained away as state failures (see Schelling above). What is more, some authors even worry the new standards may produce perverse outcomes, reflected in Betts's remark that "perceived intelligence failures often generate 'reforms' that produce other pathologies".[28]

To sum up, one strand of criticism against ICD203 is that analytic reform was not necessary given the failures of anticipating 9/11 and the October 2002 NIE and that without more emphasis on how intelligence is managed and consumed it will not be sufficient either.

A second strand of criticism against ICD203 is that the standards contained within it are not appropriate. This argument takes many forms. First, anecdotal evidence suggests the elements of analytic standards in ICD203 have been largely in use long before 2004[29] and that their emphasis in IRTPA reflects the lack of intelligence knowledge of the drafters of the act.[30] Second, the standards are accused of being too broad, to the point that they are commonsensical and uninformative.[31] The implication is that complicated procedures were added to analysts' workflows to ensure they follow principles that they were already following (either because they were standard practice or because they are too commonsensical). Third, some critics emphasise that the goal of analysis is wisdom, insight and knowledge, rather than process, data and clarity of expression. In consequence, good analysis cannot be operationalised and it cannot "be tightened or tweaked to improve the outcome".[32]

In this paper we do not take a position on whether 9/11 and the October 2002 NIE were due primarily to analytic or policy failure. Instead we remark that even the staunchest opponents of ICD203 and the Analytic Transformation championed by ODNI such as Gentry agree that "[t]he room for analytical error is large for cognitive, psychological, and institutional reasons that the intelligence literature discusses at length, and that most analysts understand in principle even if they sometimes err in practice".[33] In a similar vein, Betts believed that "minor improvements are possible by reorganizing to correct pathologies".[34] Hence, the question we are interested in is

whether ICD203 can overcome at least some of the cognitive reasons for analytical error and correct at least some of the pathologies present in the analytic cycle. It is unlikely that ICD203 is a panacea for all analytic pathologies and that by itself could prevent major future strategic surprises. In this sense we agree with its critics. On the other hand, it is possible that it makes incremental improvements to the production and evaluation of analytical reports. In this sense we side with its supporters. Importantly, the latter's claims are empirical and are yet to be tested.

3. Unpacking Intelligence Community Directive 203

The idea that ICD203 would improve analysis relies on three assumptions: (1) that the tools to assist analysts with the application of the tradecraft standards are reliable; (2) that analytic tradecraft standards sufficiently capture the key elements of good reasoning; and (3) that good reasoning leads analysts to perform better, which in turn will reduce intelligence failure rates.

The first assumption has received very little attention so far. The crux of it is: do different people apply the guidelines in ICD203 consistently? One could examine this question from the perspective of the production of reports (can analysts following the tradecraft standards in ICD203 apply them consistently?) and the perspective of their evaluation (can evaluators apply the tradecraft standards consistently?). In this paper, we focus on the latter perspective.

The second assumption can be framed as an issue of criterion validity of the tradecraft standards, that is, the degree to which the ICD203 criteria and their application conform to an "established" reference point of reasoning quality. In particular, we are

interested in concurrent validity, where the performance of ICD203 is compared against a second, concurrently, but independently obtained metric of quality of reasoning. One can investigate the concurrent validity of ICD203 from the perspective of the production of reports as well as from their evaluation by AIS. The former would correspond to an answer to the question: Do reports following the guidelines in ICD203 exhibit better reasoning? The latter would answer the question: Is there a strong correlation between evaluations of reports based on ICD203 and based on a different (well-established) method for assessing quality of reasoning? In this paper we focus on the latter perspective.

There are reasons to be pessimistic that a strong correlation exists. Take for instance ICD203's insistence on Structured Analytical Techniques (SATs): "Analysts must perform their functions with objectivity and with awareness of their own assumptions and reasoning. They must employ reasoning techniques and practical mechanisms that reveal and mitigate bias." (ICD203.6.a) SATs have become ubiquitous recommendations for improving analytic tradecraft. [35] SATs are defined as "mechanism[s] by which internal thought processes are externalized in a systematic and transparent manner so that they can be shared, built on, and easily critiqued by others."[36] But do they deliver on this promise of mitigating bias? A recent meta-analyis by Coulthart[37] suggests that the best answer we can give is 'maybe, in some cases'. Coulhart finds that out of the 12 SATs mentioned in the *Analytic Tradecraft Primer*, only 6 have been studied (45 studies available), but that only 3 of them had "credible research studies".[38] Out of the three, only one had consistently proven its efficiency in experiments.

Chang et al. note that although the declared purpose of SATs is to reduce "systematic biases and random noise … no one knows how close the current generation of SATs comes to achieving either of them".[39] They believe that despite the lack of evidence, we should question the theory behind them. First, there isn't enough evidence to show that even SATs successful at mitigating one type of cognition bias (say overconfidence) won't generate its opposite (e.g. underconfidence). Second, SATs that focus on decomposing complex judgments into simpler ones may create more opportunities for making errors and the "cumulative nature of error in multi-stage assessments will make those judgments less, not more, consistent".[40]

Finally, commenting on ODNI's function as an analytic ombudsman, Fingar claims that "it's clear, because we've now got data on thousands of products that in aggregate, we're getting better."[41] But better in what sense? Marrin[42] distinguishes between three aims of the intelligence community; (predictive) accuracy, reduction of strategic surprise and significance for policy-making, and argues that currently there is no consensus on what analysts should strive for.

For instance, some authors, such as Marchio[43] and Friedman and Zeckhauser[44] seem to assume that the goal is estimative accuracy.[45] But is this right? Tetlock and Mellers purport the intelligence community has "tacitly placed a massive institutional bet on the validity of its home-grown theory of good judgment: namely, that accuracy should be a positive function of how well analysts conform to the process standards embodied in its performance management guidelines."[46] Similar concerns have been raised by Lowenthal and Gentry in their critiques of the presumed analytic transformation engendered by IRTPA, ODNI and ICD203.[47] Even authors such as Fingar and Friedman and Zeckhauser are aware of this concern.[48]

Nevertheless, there is little evidence in the scholarly literature relevant to the issue of accuracy,[49] perhaps because the task of determining when an analytic judgment is accurate may be ambiguous and there is a longstanding view that it is intractable. Indeed Sherman Kent believed it was "almost impossible to answer…".[50] Several standard reasons are usually invoked to support Kent's view: "since estimates are probabilistic, we can never really say whether they are 'right' or 'wrong'",[51] "every intelligence question is unique. Therefore, even if we wanted to evaluate accuracy across estimates, broad patterns are not meaningful",[52] "even if we wanted to measure estimative accuracy, there is no rigorous way to keep score"[53] and "these methods all require analysts to state probabilities quantitatively"[54] and "evaluating estimative accuracy would be too expensive".[55] Friedman and Zeckhauser argue convincingly that these challenges can be overcome (at least partially). Marrin[56] and Marrin and Clemente[57] point out similar hurdles in medicine and that there may be useful cross-fertilization between fields on how to measure accuracy and quality of judgments. Nevertheless, investigating whether tradecraft correlates with accuracy can only come after the community settles the meaning of the latter. We leave this for a future paper.

None of these assumptions have been properly, empirically tested. In this paper we offer the first empirical investigation of assumptions (1) and (2) under controlled conditions. We focus our attention on an operationalisation of the tradecraft standards currently employed by AIS (see Section 4) and draw on a set of constrained reasoning problems developed by the CREATE research program and on analytic reports developed in response to those problems (see Section 5).

4. ICD203 and the Rating Scale for Evaluating Analytic Tradecraft Standards

ICD203 outlines the standards for good reasoning in intelligence but does not provide specific guidelines for how to apply them. For the purposes of testing reliability and accuracy, the analytic tradecraft standards need to be operationalised. We use a modified version of the *Rating Scale for Evaluating Analytic Tradecraft Standards* (*Rating Scale* from now on), the rubric used by experts from AIS to evaluate analytic products. Our simplified Rating Scale comprises 8 criteria that capture the tradecraft standards outlined in ICD203, each with detailed interpretation and examples:

Table 1: The *Rating Scale* criteria

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 authors' assumptions and judgments
4	Incorporates analysis of alternatives
5	Demonstrates relevance and addresses implications
6	Uses clear and logical argumentation
7	Makes accurate judgments and assessments
8	Incorporates effective visual information where appropriate

The criterion "Explains change to or consistency of analytical judgements" was omitted because it requires the report writer to have an understanding of previous

analyses, which is irrelevant for the kind of problems we used in this experiment (see below). Furthermore, we made a small number of minor textual changes to the *Rating Scale* motivated by participants' unfamiliarity with the jargon used in the original document, and to match the specific nature of the more constrained problems we used in this experiment. For instance, the word "product", typically referring to an analysis of intelligence information was changed to the more general "report". This study reports on two experiments conducted during the first beta test of the SWARM platform (July-August 2017),[58] one of the products of the CREATE program (see below).

5. CREATE and SWARM

As part of the effort to improve the quality of intelligence analysis, the US intelligence community's research arm, the Intelligence Advanced Research Projects Activity (IARPA) funded the CREATE program (Crowdsourcing Evidence, Argumentation, Thinking and Evaluation). Under the umbrella of CREATE, research teams are investigating how group reasoning can be harnessed to improve intelligence analysis. The Smartly-assembled Wiki-style Argument Marshalling (SWARM) project,[59] based the University of Melbourne, Australia, is developing an online platform where analysts collaborate to generate solutions to reasoning problems. The platform aims to (a) use the power of distributed processing within a network of individual thinkers, and (b) improve reasoning quality and the aggregation of solutions into a final, agreed solution. Users write individual analytical reports that outline the outcome (the solution to the problem) and the process (the underlying reasoning). Users then comment and/or make edits to one another's contributions, to

improve individual solutions. The platform encourages users to rate others' solutions and the aggregate (average) quality rating (on a scale 0-100) determines the rank of each solution. The top-rated solution becomes the template for a final, collectively drafted report.

The system is under development and is designed initially to assist groups of users to solve constrained reasoning problems, those that "can be solved using provided materials supplemented only with unproblematic, well-known background facts and reasonable assumptions. In addressing constrained problems, subjects use only common knowledge and provided problem-specific information" (CREATE BAA: 8). These reasoning problems were developed by the performer teams in CREATE and problem developers at *Good Judgment Inc*.[60]

This study draws on reports submitted by users of the SWARM platform during beta testing. Beta testers were members of the SWARM research team with extensive experience in critical thinking and argumentation, structured analytical techniques and the systematic evaluation of written reports such as peer-review and university coursework assessments. For the purposes of this study, the beta-testers were assumed to be highly-experienced evaluators. The average quality score (on a scale from 0-100) submitted by these evaluators for each report was therefore taken to be a close approximation of the true value of the quality of reasoning of that report. There are two reasons why this cohort was treated as highly-experienced. First, they had intimate knowledge of the problems and they had spent an extensive amount of time reading and discussing those problems and the submitted analytic reports. Second, they collectively had considerable professional experience in the assessment of

written reports on the basis of analytic rigor, logic, argument structure and general clarity of presentation. Their evaluations of the reports were found to be reliable, suggesting at least internal consistency of these "expert judgements" (Appendix C).

6. Evaluating the reliability and concurrent validity of ICD203

In the course of two experiments, we investigated whether novice evaluators (those with substantial backgrounds in the practical assessment of reasoning, but with no prior knowledge of AIS's *Rating Scale*) could apply it reliably. We also assessed whether their ratings matched the expert evaluations of the same analytic products. The first is a matter of inter-rater reliability and gives us an insight into assumption 1. The second offers an insight into whether assumption 2 holds.

Our expectation was that in Experiment 1, without evaluator training, both inter-rater reliability and correlation with expert judgements would be low. This will provide a benchmark against which the effect of evaluator training could be evaluated.

The following two subsections detail the results and the methodology behind the two experiments. We note that despite having access to some of the practices in place in AIS for the evaluation of intelligence reports, we are not privy to the full protocols currently in use. Therefore, our study cannot perfectly replicate professional practice. Moreover, there may be other agencies in the US and elsewhere that are performing similar quality control exercises using similar instruments. For these agencies and others interested in this topic to benefit from our results, they need to be able to compare their internal procedures to the procedures we followed and hence need full access to our entire experimental protocol.

6.1 Experiment 1

Participants. We recruited 41 postgraduate (masters and doctoral) students from *Imperial College London* and the *London School of Economics and Political Science* to form our population of 'novice' evaluators. The only requirement was that they were experienced / familiar with assessment rubrics in education (but not intelligence). We chose to enlist naive evaluators from outside the IC in order to measure the effect calibration training may have on inter-rater reliability. Any specific intelligence analysis training would confound the attribution of experimental effects between the baseline (experiment 1) and the intervention (experiment 2). Moreover, any application of the rubric might likewise have been tainted by analysts' preconceptions of the "gold standards" of intelligence analysis.

The sample consisted of 20 men and 21 women, with an average age of 25.9 years (SD=5.2 years). Most had a social science background (63%), with a minority having a natural science (20%) or an arts & humanities (17%) background. Participants were informed of the aims and procedures and gave written consent to participate. All data were collected anonymously.

Materials and Procedure. We obtained 27 reports generated by beta-testers on the SWARM platform in response to 6 constrained reasoning problems. These included reports generated by individuals and final collaborative reports submitted by teams. They varied in quality of reasoning based on the collective assessment of SWARM users (range=13-94, M=68.33, SD=17.58). Reports were allocated to individual evaluators on a constrained-random basis, with the constraint that each product was evaluated by three evaluators. Before performing the evaluations, participants were

given the opportunity to familiarise themselves with the *Rating Scale*, take notes and ask general questions about terminology (approx. 30 min.). No further instructions were given for the application of the *Rating Scale* to specific reports. Participants were given two products each (except for one participant who only received one product) and instructed to use a scoring sheet to assess each product (Appendix A). Evaluators had to indicate for every criterion in the *Rating Scale* whether the product was of Poor, Fair, Good or Excellent quality.

Analysis. To compute reliability and validity measures, the qualitative assessments on the individual criteria were aggregated into an overall assessment of quality of reasoning for each product. We investigated two aggregation methods.[61] In the first, each assessment level was assigned a value (from 0=poor to 3=excellent), and values were summed across criteria (Appendix B: Table 1). The second involved assigning weights to the different criteria depending on the particular feature they were probing: Use of Evidence and Reasoning were weighted more heavily than Communication in determining the overall quality of reasoning score (Appendix B: Table 2). We call the former the Equal Weights Scoring System, and the latter the Weighted Scoring System. To assess reliability of the rubric, we calculated the intraclass correlation (ICC), commonly used to indicate the consistency or reproducibility of quantitative measurements made by different observers rating the same object(s), in this case, the analytical reports. It generates a measure between 0-1, with higher values indicating greater consistency (see appendix C for more information).

Results. In the absence of training, the inter-rater reliability of the *Rating Scale* according to both the Equal Weights and the Weighted Scoring Systems in

Experiment 1 was low: ICC=0.294 (95% CI [-0.330,0.654]) and ICC ≈ 0 (95% CI [-1.607,0.322]), respectively. The individual report scores generated by novice evaluators using the *Rating Scale* correlated positively with SWARM-generated quality ratings. Nevertheless, the correlation was low-moderate for both the Equal Weights and the Weighted Scoring Systems: r=.293 and r=.262, respectively.

Discussion. The *Rating Scale* appears to have low inter-rater reliability and offers little help to novice users in matching the judgments of expert evaluators. The results of Experiment 1 are somewhat surprising, seeming to validate concerns regarding the ambiguity of language in ICD203 and refuting the argument claiming ICD203 is too commonsensical.[62] However, the *Rating Scale* was unfamiliar to participants before the experiment and the short familiarisation exercise at the beginning may have been insufficient. Evaluators working for AIS have extensive experience both in intelligence analysis and in evaluating analytic reports. In addition, inter-rater agreement often is improved by training and calibration exercises.[63] Therefore, to more closely reflect professional circumstances, we tested the impact of calibration training on the inter-rater agreement and validity of the *Rating Scale*.

6.2 Experiment 2

Participants. We recruited 36 postgraduate (masters) students from *Imperial College London* and the *London School of Economics* with appropriate backgrounds in reasoning evaluation but who had not participated in Experiment 1. The sample consisted of 21 men and 15 women, with an average age of 25.03 years (min=22, max=41, SD=3.86). Most had a social science background (80%), with a minority

having an arts & humanities (13.3%) or a natural science (6.7%) background. Participants were informed of the aims and procedures and gave written consent to participate. All data were collected anonymously.

Materials and Procedures. For the purposes of the second experiment, we selected a subset of 13 reports in response to 3 of the 6 constrained reasoning problems. These reports were chosen because they generated the lowest inter-rater reliability in Experiment 1. First, participants were given the opportunity to individually familiarise themselves with the *Rating Scale*, take notes and ask general questions about terminology, as in Experiment 1. Then, in contrast to Experiment 1, they were allocated to groups of 2-3 participants and were asked to read a constrained intelligence problem and its report (these had not been used in Experiment 1 but were generated for additional testing on the SWARM platform). Within these groups, evaluators were then asked to reach a consensus on the evaluation of the report given the *Rating Scale*, using the same scoring sheet as in Experiment 1 (45 min. in total). Following this calibration activity, evaluators individually completed evaluation of one of the 13 experimental reports. Members of the same calibration group rated the same report. The intention was to have 3 evaluators for each report, but due to no-shows, 3 out of the 13 groups consisted of just 2 evaluators. The analyses proceeded identically to those in Experiment 1.

Results. For the 10 reports that had three evaluators,[64] inter-rater reliability of the *Rating Scale* was good for the Equal Weights and fair for the Weighted Scoring Systems, ICC= 0.612 (95% CI [-0.101,0.894]) and ICC=0.491 (95% CI [-0.444,0.861]), respectively. Two-rater agreement assessed across all 13 reports was

fair for both the Equal Weights Scoring System: ICC=0.473 (95% CI [-0.661,0.837]), and the Weighted Scoring System: ICC=0.514 (95% CI [-0.532,0.850]).[65]

To examine the effect of the calibration exercise on agreement between evaluators in the Equal Weights Scoring System, we calculated the difference in report scores given by pairs of evaluators assessing the same report (e.g. for reports rated by 3 evaluators, there were 3 possible pairwise comparisons). In Experiment 1, average difference between pairs of evaluators was 5.4 points (SD=3.8) (M=5.6, SD = 4.2 for the subset of 13 reports also included in experiment 2). For Experiment 2, the average difference was reduced slightly to 4.1 points (SD=3.2). The distribution of pairwise differences shows that 90% of evaluator pairs in Experiment 1 were within 10 points of one another, and 90% of evaluator pairs in Experiment 2 were within 7 points of one another (Figure 2).

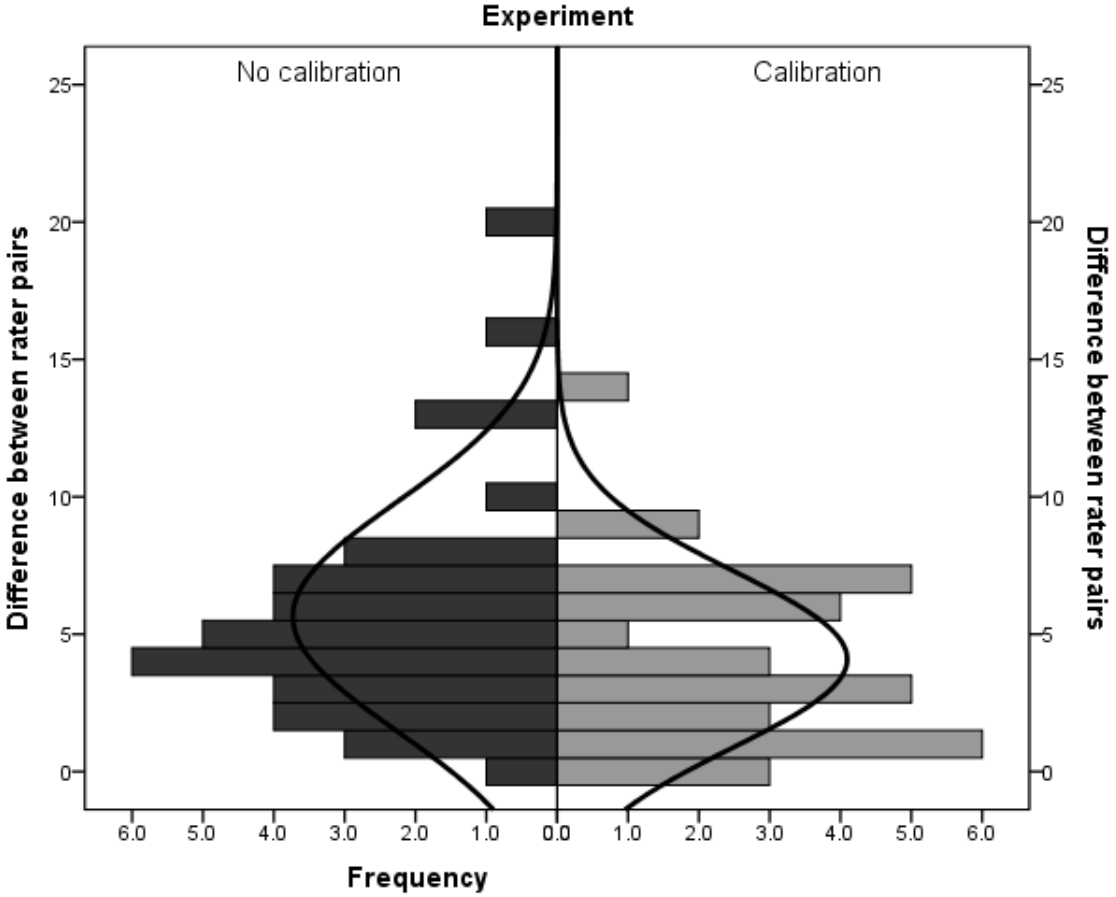

Figure 2. Distribution of score differences between pairs of evaluators assessing the same report using the Equal Weights Scoring System in Experiment 1 (without training/calibration) and in Experiment 2 (with training/calibration).

A similar conclusion could be drawn from examining the Weighted Scoring System, which generates a 5-level categorical assessment (from Poor to Excellent). We evaluated the number of category shifts between two evaluators assessing the same report. The majority (89%) of evaluator pairs were within 2 categories for Experiment 1. This increased to 100% in Experiment 2 (Figure 3).

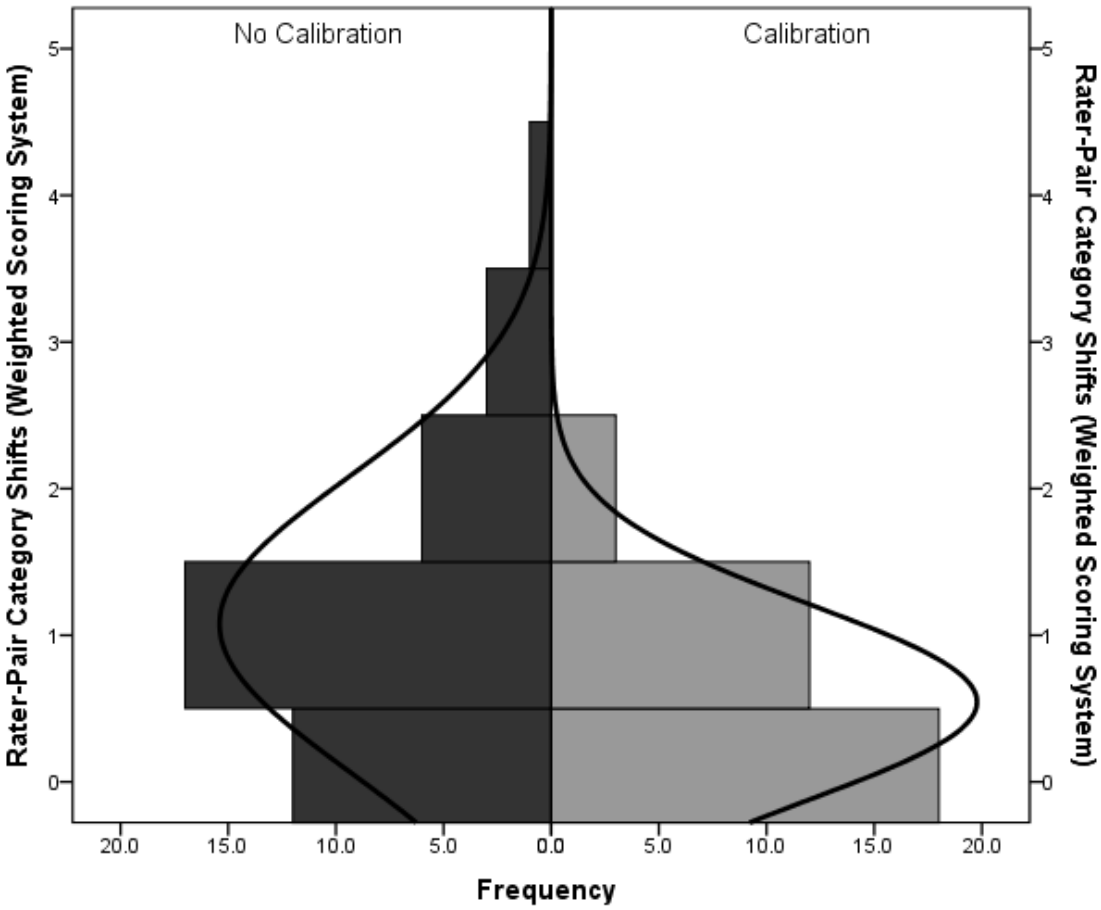

Figure 3. Distribution of category shifts between pairs of evaluators assessing the same report using the Weighted Scoring System in Experiment 1 (without calibration) and in Experiment 2 (with calibration).

To further examine whether specific criteria gave rise to more disagreement than others, we calculated the average difference between evaluator pairs in their assessment of individual criteria on a given report (Figure 4). Each criterion was rated on a 4-point nominal scale (from 0=poor to 3=excellent), so the maximum difference between evaluators was 3 points. The improvement in agreement between evaluators in experiment 2 compared to experiment 1 appears to be due to greater consistency in the proper expression and explanation of uncertainties (criterion 2), the incorporation

of alternatives (criterion 4), assessment of implications (criterion 5) and the appropriate incorporation of visual materials (criterion 8).

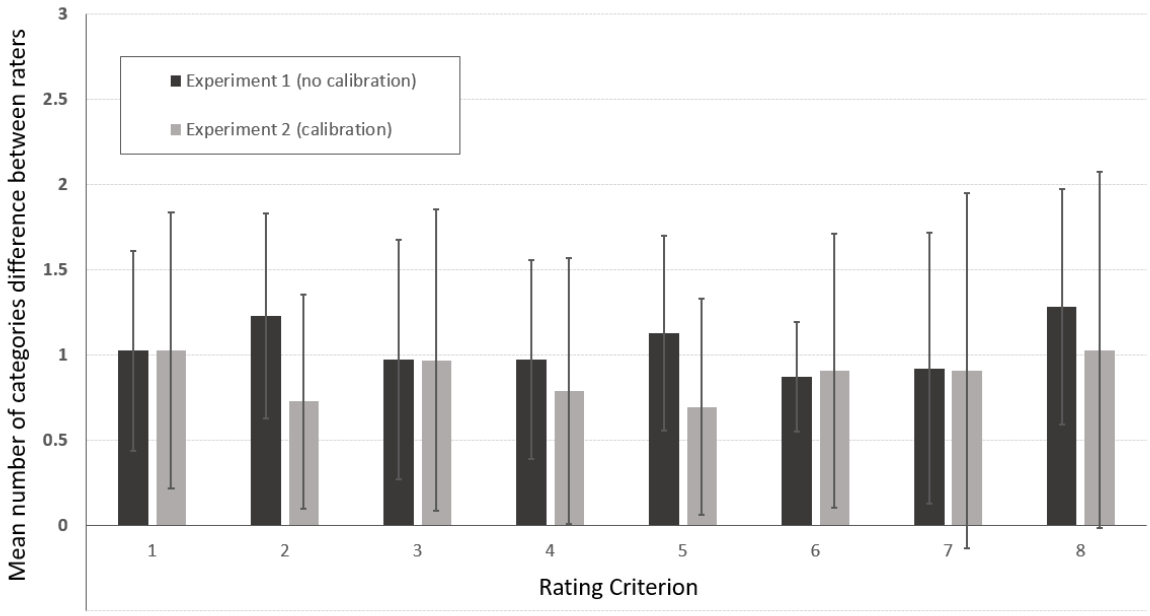

Figure 4. Average difference in assessment of the 8 criteria for evaluator pairs rating the same report.

The *Rating Scale* report scores generated by the novice evaluators correlated positively with the SWARM-generated quality ratings, and the association was stronger in Experiment 2 than in Experiment 1, reaching r=.614 and r=.512 for the Equal Weights and the Weighted scoring system respectively (Figure 5).

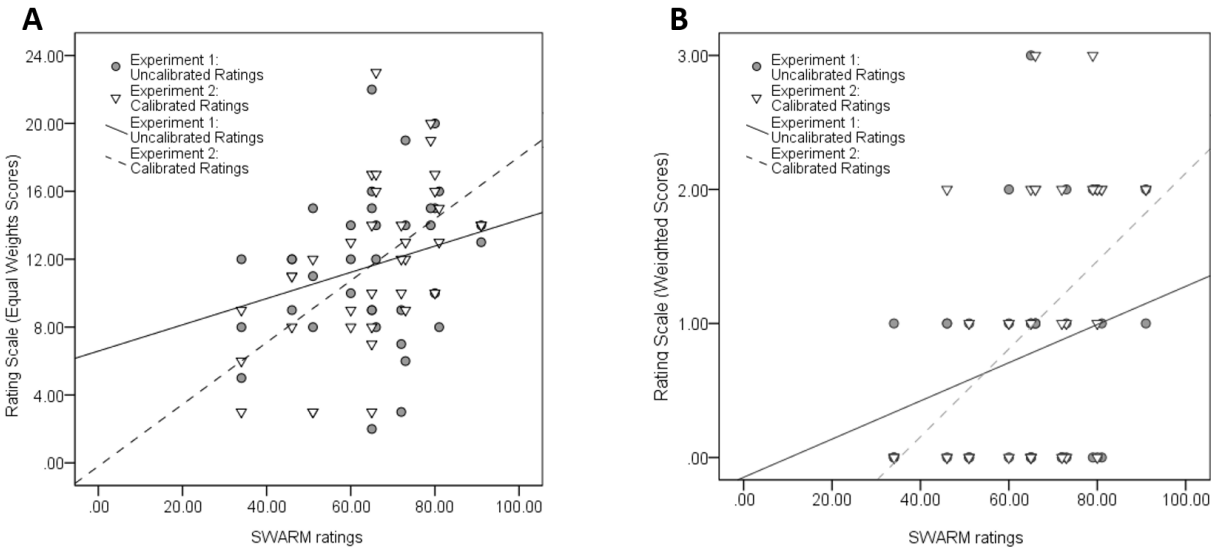

Figure 5. Correlations between expert-generated ratings on the SWARM platform and evaluations made by novice evaluators using the *Rating Scale*. Correlations are shown for the Equal Weight Scores (A) and the Weighted Scores (B).

Discussion. In these experiments, calibration-training substantially improved both the inter-rater reliability of the *Rating Scale* and the correlation between novice and expert evaluations. Novice evaluators who calibrated their evaluations on a training report were subsequently more likely to agree in their assessments of other reports. This suggests that the low values for both ICC and for the correlations with expert judgments obtained in Experiment 1 were due to the lack of familiarity of our evaluators with the criteria. Even the modest 45 minute calibration round applied in these trials helped novice evaluators to reach a fair to good (depending on the scoring method) inter-rater correlation and a strong correlation with expert judgments.

These results should be interpreted as good news for the champions of the analytic transformation for two reasons. First, part of the ODNI reforms was to require all analysts to go through a basic course, Analysis 101,[66] which is built around the

tradecraft standards. All analysts have extensive training in the tradecraft standards and therefore, the production and evaluation of products should be more reliable than can be seen in the present study. Furthermore, the practice of AIS is to have each product reviewed by two independent evaluators who then discuss each evaluation, which are further checked by a third evaluator (Guidelines to the *Rating Scale*). Consequently, there are ample opportunities for evaluators within AIS to calibrate their reasoning and interpretation of the language of the standards.

Nevertheless, agreement between evaluators did not increase for all criteria, suggesting that there is some reason for concern with respect to the language used in ICD203. Notably, there was substantial disagreement on arguably critical aspects including the description of sources, the accuracy of judgments and the use of clear logic and argumentation. Calibration appeared to reduce some ambiguity in the assessment of report writers' expressions of uncertainty associated with analytical judgment, how well alternative hypotheses were incorporated, and the use of supporting visual information where appropriate. This suggests that more specific training and/or more precise guidelines may be resolve differing interpretations of some of the more critical elements of quality of reasoning. Whether the training and opportunities to calibrate available to analysts are enough to overcome these challenges remains an open question.

7. General discussion and conclusions

Validating standards intended to capture quality of reasoning in intelligence reports is complicated by the absence of external standards. However, CREATE and SWARM

offer an opportunity for such validation. The constrained reasoning problems developed for Phase 1 of CREATE and used in this study have normative standards according to which solutions can be objectively assessed. Experts who are familiar with the problem and its solution can reliably and accurately judge whether the reasoning in a report matches to a sufficient degree the 'gold standard' for that particular problem.

While the constrained nature of the problems affords unprecedented opportunity for external validation of their analytical solutions, it also imposes limits on generalisability of the findings beyond such a highly controlled setting. Further study of the application of ICD203 in more realistic settings is therefore advised. If these results generalize, then groups of evaluators, trained in the same way and using the same ICD203-based rubric, would produce fairly consistent ratings on unconstrained, forecasting problems. Such an outcome would imply that the results of this study are likely to carry over to real-world applications of the *Rating Scale*.

On balance, our findings present a cautiously optimistic picture. We find that novice evaluators can assess the quality of reasoning in intelligence reports with acceptable levels of consistency. We stop short of claiming that therefore, ODNI can successfully perform its role as an analytic ombudsman as we do not yet have empirical confirmation that the third assumption (that reports that satisfy the ICD203 criteria consistently produce accurate predictions of outcomes) behind ICD203 holds. The bad news is that when focusing on the reliability of individual users who have to use the Rating Scale on their own, the effect of the training on reliability is negligible. Worse, the absolute reliability values tend to be quite low. This suggests that if the

standards in ICD203 are essential for good tradecraft, then analysis of reports has to be done in teams and no assessment should be produced by a lone analyst.

The results reported in this study suggest that the standard of good reasoning that the intelligence community has adopted as a consequence of their analytic transformation and deployed in their quality control program are more reliable and valid than implied by its critics. At the same time our results also highlight potential areas of improvement. We find that reliable and accurate identification of good reasoning relies not only on the standards in ICD203 but also on the way the evaluators are trained and calibrated, and the scoring method used to aggregate their judgments on individual criteria into an overall score for a report. Consequently, more attention should be given to these issues.

Critics of ICD203 accuse reformers, policy makers and the public of imposing too high a standard on intelligence analysts. But they themselves impose too high a standard on reforms. Betts believed improvements can only come at the margins. Our studies indicate that ICD203 has delivered at least that.

Funding. This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research projects Activity (IARPA), under Contract [2017-16122000002]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or

Bibliography

1. Artner, S., R. Girven, and J. Bruce. *Assessing the Value of Structured Analytical Techniques in the U.S. Intelligence Community*. RAND Corporation, 2016.
2. Betts, R. "Analysis, War, and Decision: Why Intelligence Failures Are Inevitable." *World Politics* 31, no. 1 (1978): 61-89.
3. Cardillo, R. "A Cultural Evolution." *Studies in Intelligence* 54, no. 3 (2010): 43-49.
4. Chang, W., E. Berdini, D. R. Mandel, and P. E. Tetlock. "Restructuring structured analytic techniques in intelligence." *Intelligence and National Security* 33, no. 3 (2018): 337-356.
5. Cicchetti, D. "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instrument in Psychology." *Psychological Assessment* 6 (1994): 284-290.
6. Coulthart, S. "An Evidence-Based Evaluation of 12 Core Structured Analytic Techniques." *International Journal of Intelligence and CounterIntelligence* 30, no. 2 (2017): 368-391.
7. Fingar, T. "Keynote Address to the *2008 INSA Analytic Transformation Conference Orlando, Florida* (September 4, 2008)." https://www.dni.gov/files/documents/Newsroom/Speeches%20and%20Interviews/20080904_speech.pdf
8. Fingar, T. *Analytic Transformation: Unleashing the Potential of a Community of Analysts* (Washington, DC: Office of the Director of National Intelligence 2008). Accessed May 10, 2018. https://www.hsdl.org/?abstract&did=29867.
9. Fingar, T. *Reducing Uncertainty: Intelligence Analysis and National Security*. Stanford: Stanford Security Studies, 2011.

10. Friedman, J. A., and R. Zeckhauser. "Assessing Uncertainty in Intelligence." *Intelligence and National Security* 27, no. 6 (2012): 824–847.
11. Friedman, J. A., and R. Zeckhauser. "Why Assessing Estimative Accuracy is Feasible and Desirable." *Intelligence and National Security* 31, no. 2 (2016): 178-200.
12. Gentry, J. "Intelligence Failure Reframed." *Political Science Quarterly* 123, no. 2 (2008): 247-270.
13. Gentry, J. "Has the ODNI Improved U.S. Intelligence Analysis?" *International Journal of Intelligence and CounterIntelligence* 28, no. 4 (2015): 637-661.
14. Heuer, R. J., and R. H. Pherson. *Structured Analytic Techniques for Intelligence Analysis*. Sage Publications, 2010.
15. IARPA, *Broad Agency Announcement: Crowdsourcing Evidence, Argumentation, Thinking and Evaluation* (Washington, DC: Office of Anticipating Surprise 2016). Accessed March 3, 2018 https://www.fbo.gov/index?tab=documents&tabmode=form&subtab=core&tabid=b315926fb14cb6d31026476229e920ab.
16. Immerman, R. H. "Transforming Analysis: The Intelligence Community's Best Kept Secret." *Intelligence and National Security* 26, no. 2-3 (2011): 159-181.
17. Immerman, R. H. "Intelligence and the Iraq and Afghanistan Wars." *Political Science Quarterly* 131 (2016): 477-501.
18. Intelligence Community Directive (ICD) 203, *Analytic Standards*, (Washington, DC: Office of the Director of National Intelligence 2015).
19. Jonsson, A., and G. Svingby. "The use of scoring rubrics: Reliability, validity and educational consequences." *Educational Research Review* 2, no. 2 (2007): 130-144.
20. Judd, T. P., C. Secolsky, and C. Allen. "Being Confident About Results from Rubrics." National Institute for Learning Outcomes Assessment Viewpoint Blog.

2012. http://illinois.edu/blog/view/915/68373?displayOrderdesc&displayTypenone&displayColumncreated&displayCount1.

21. Kent, S. "Words of Estimative Probability." *Studies in Intelligence* 8, no. 4 (1964): 49-65.
22. Lowenthal, M. "Towards a Reasonable Standard for Analysis: How Right, How Often on Which Issues?" *Intelligence and National Security* 23, no. 3 (2008): 303-315.
23. Lowenthal, M. "A Disputation on Intelligence Reform and Analysis: My 18 Theses." *International Journal of Intelligence and CounterIntelligence* 261 (2012): 31-37.
24. Marchio, J. "Analytic Tradecraft and the Intelligence Community: Enduring Value, Intermittent Emphasis." *Intelligence and National Security* 29, no. 2 (2014): 159-183.
25. Marrin, S. "Training and Educating US Intelligence Analysis." *International Journal of Intelligence and CounterIntelligence* 22, no. 1 (2009): 131-146.
26. Marrin, S. "Evaluating the Quality of Intelligence Analysis: By What (Mis) Measure?" *Intelligence and National Security* 27, no. 6 (2012): 896-912.
27. Marrin, S., and J. Clemente. "Improving Intelligence Analysis by Looking at the Medical Profession." *International Journal of Intelligence and CounterIntelligence* 18, no. 4 (2005): 707-729.
28. Marrin, S., and J. Clemente. "Modelling and Intelligence Analysis Profession on Intelligence." *International Journal of Intelligence and CounterIntelligence* 19, no. 4 (2006/2007): 642-645.
29. Patterson, E., S. McNee, D. Zelik, and D. Woods. "Insights from Applying Rigor Metric to Healthcare Incident Investigations." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52, no. 21 (2008): 1766 - 1770.

30. Phythian, M. “The Perfect Intelligence Failure? U.S. Pre-War Intelligence on Iraqi Weapons of Mass Destruction.” *Politics & Policy* 34 (2006): 400-424.
31. Reddy, Y. M., and H. Andrade. “A Review of Rubric Use in Higher Education.” *Assessment and Evaluation in Higher Education* 35, no. 4 (2010): 435-448.
32. RiCharde, R. S. “The Humanities Versus Interrater Reliability.” *Assessment Update* 20, no. 4 (2008): 10-11.
33. Silberman, L. and C. Robb. *Report to the President of the United States*. Washington: USGPO, 2005.
34. Tetlock, P. E., and B. A. Mellers. “Intelligent management of intelligence agencies: Beyond accountability ping-pong.” *American Psychologist* 66, no. 6 (2011): 542-554.
35. Turbow, D. J., and J. Evener. “Norming a VALUE rubric to assess graduate information literacy skills.” *Journal of the Medical Library Association : JMLA* 104, no. 3 (2016): 209–214.
36. U.S. Government. *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*. Washington: Center for the Study of Intelligence Analysis, 2009.
37. van Gelder, T., and R. de Rozario. “Pursuing Fundamental Advances in Reasoning.” In *Artificial General Intelligence, Proceedings of the 10 International Conference, AGI-2017, Lecture Notes in Artificial Intelligence*, edited by T. Everitt, A. Potapov, and B. Goertzel, 259–262. Cham: Springer, 2017.
38. Wohlstetter, R. *Pearl Harbor: Warning and Decision*. Stanford: Stanford University Press, 1962.
39. Zegart, A. “September 11 and the Adaptation Failure of U.S. Intelligence Agencies.” *International Security* 29, no. 4 (2005): 78-111.

40. Zegart, A. “An Empirical Analysis of Failed Intelligence Reforms Before September 11.” *Political Science Quarterly* 121 (2006): 33-60.
41. Zelik, D., E. Patterson and David Woods. “Judging sufficiency: How professional intelligence analysts assess analytical rigor.” In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 51, no. 4,* 318-322. Los Angeles: SAGE Publications, 2007.
42. Zelik, D., E. Patterson and D. Woods. “Understanding rigor in information analysis.” In *8th International Conference on Naturalistic Decision Making*. Pacific Grove, 2007.
43. Zelik, D., E. Patterson and D. Woods. “Measuring Attributes of Rigor in Information Analysis.” In *Macrocognition Metrics and Scenarios: Design and Evaluation of Real-world Teams*, edited by E. Patterson and J. Miller. Aldershot: Ashgate, 2010.

Appendix A: The scoring sheet

REPORT 1 **Report number: …**				
CRITERION	**RATING**			
1	Poor	Fair	Good	Excellent
2	Poor	Fair	Good	Excellent
3	Poor	Fair	Good	Excellent
4	Poor	Fair	Good	Excellent
5	Poor	Fair	Good	Excellent
6	Poor	Fair	Good	Excellent
7	Unclear	Conditioned	Unconditioned	
8	Poor	Fair	Good	Excellent

Appendix B: The scoring systems

	Score			
	0	1	2	3
Criterion 1	Poor	Fair	Good	Excellent
Criterion 2	Poor	Fair	Good	Excellent
Criterion 3	Poor	Fair	Good	Excellent
Criterion 4	Poor	Fair	Good	Excellent
Criterion 5	Poor	Fair	Good	Excellent
Criterion 6	Poor	Fair	Good	Excellent
Criterion 7	Unclear		Conditioned	Unconditioned
Criterion 8	Poor	Fair	Good	Excellent

Table 1: The equal weights scoring system

A report is [row] if it is at least [cell] on [column]	**Evidence**	**Reasoning**				**Communication**		
	Criterion 1	Criterion 2	Criterion 4	Criterion 5	Criterion 6	Criterion 3	Criterion 7	Criterion 8
4	Excellent	Excellent	Excellent	Good	Good	Excellent	Excellent	Good
3	Good	Excellent	Good	Fair	Good	Excellent	Excellent	Fair
2	Fair	Good	Good	Fair	Fair	Good	Good	Fair
1	Fair	Good	Good	Poor	Fair	Fair	Fair	Fair
0								

Table 2: The weighted scoring system.

Appendix C: Details of the ICC method

ICC calculations

There are many metrics and little agreement about the best method to test the strength of the association between assessments made by (groups of) independent evaluators. The intuitively appealing correlation coefficient is insensitive to mean differences, i.e. bias. Simple measures of percentage agreement do not account for random (i.e. chance) agreement between evaluators.

The intraclass correlation coefficient (ICC), as used here, assesses reliability by comparing the variability of different ratings of the same item (here: reports) to the total variation across all ratings and all items. ICC values lie between 0.0 and 1.0, with higher values corresponding to greater consistency between evaluators. The ICC is responsive to both lower correlation between evaluators and larger mean differences, meaning it is sensitive to bias. It accounts for chance-level agreement between evaluators. It should be noted that some statistical programs produce negative values, which are reported as ~ 0, i.e. indicating very low reliability. We used the IRR package in R to calculate ICC values, with a One-Way Random-Effects Model in which each object is rated by a different set of evaluators who were randomly chosen from a larger population of possible evaluators. There are different sub-types of ICC, and here we report the ICC value that represents the degree of consistency between evaluators, as opposed to absolute numerical agreement. ICC values <.40 indicate poor inter-rater agreement, between .40-.59 fair agreement, and >.60 good agreement. Concurrent validity was assessed with Spearman rank order correlations between the average quality ratings given by experts on the SWARM platform and the Rating Scale assessments made by the novice evaluators.

ICC values for the "expert assessments"

The expert assessments of quality of reasoning were generally consistent. Examining the inter-rater reliability across the set of reports that had been rated by at least three experts during SWARM beta-testing, we found ICC=0.794, indicating very good agreement on quality of reasoning between expert assessors. The majority of reports were rated by more than 3 experts on the SWARM platform. To enable a direct

comparison with the ICC obtained in the samples of novice evaluators, we randomly selected 3 expert assessors per report. We repeated this process 1000 times, obtaining different permutations of 3 evaluators, with an average ICC=0.794 (SD=0.039).

[1] Wohlstetter, *Pearl Harbor: Warning and Decision*, viii.

[2] https://nsarchive2.gwu.edu/NSAEBB/NSAEBB129/nie.pdf (Accessed 10 May 2018).

[3] Cardillo, "A Cultural Evolution", 44.

[4] See fn. 63.

[5] One of the main goals of this study is to offer an empirical assessment of the rubric used by AIS in order to perform their task of evaluating the quality of reasoning in intelligence products. A future research project could build on the results presented here and assess how different refinements of the AIS rubric perform in terms of reliability and correlation to expert evaluations. Such a project should be informed by other research on evaluation rubrics, such as Zelik et al, "Judging Sufficiency" and Zelik et al., "Understanding Rigor in Information Analysis". Zelik and his collaborators interviewed professional information analysts and determined "consistent patterns" of "critical vulnerabilities in information analysis" (Zelik et al., "Measuring Attributes of Rigor", 3). They organised these patterns into 8 different attributes of shallow analysis and developed a metric based on them to evaluate the rigor of analysis in an intelligence report. Their research is relevant to the AIS rubric for three reasons. First, Zelik et al.'s metric is inferred from intelligence professionals with experience in evaluating the kind of reasoning encountered in intelligence products. Second, their metric also uses qualitative levels of satisfaction for each criterion (though they use three instead of four). Third, they investigated the robustness of their rigor metric by applying it in other contexts, such as accident investigation analyses (Patterson et al., "Insights from Applying Rigor"). We believe this is a promising avenue of research but we do not pursue it any further in this paper.

[6] E.g. *The 2004 Senate Select Committee on Intelligence*, *President Bush's Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction* (Silberman and Robb, *Report to the President*), etc. A detailed discussion of the main findings of these committees can be found in Phythian, "The Perfect Intelligence Failure?".

[7] https://www.dni.gov/files/documents/IRTPA%202004.pdf.

[8] Similar recommendation can be found in the recent *Report on the Iraq Inquiry* (Chilcot Inquiry) carried out in the UK which recommended that authors of intelligence products should appreciate "[t]he importance of precision in describing [their] position" especially in the executive summary of the report. Moreover, it recommended that analysts "need to identify and accurately describe the confidence and robustness of the evidence base", "to be explicit about the likelihood of events", "to be scrupulous in discriminating between facts and knowledge on the one hand and opinion, judgement or belief on the other", and finally "to avoid unwittingly crossing the line from supposition to certainty, including by constant repetition of received wisdom" (Section 4.2.900).

[9] This position no longer exists.

[10] Fingar's belief in the truly transformative nature of the reforms put forward by ODNI can be seen in his address to the 2007 *Analytic Transformation Symposium in Chicago, Illinois*: "The kind of changes embodied in the term analytic transformation, if they're going to be successful, we'll be revolutionary."

(https://www.dni.gov/files/documents/Newsroom/Speeches%20and%20Interviews/20070905_speech.pdf)

[11] As Lowenthal puts it "[t]hese are not groundbreaking nor are they especially remarkable standards. In fact, most of them are fairly commonsensical but still mandatory." (Lowenthal, "Towards a Reasonable Standard", 307)

[12] ICD203.

[13] Fingar, *Reducing Uncertainty.*

[14] "[T]he primary purpose of intelligence inputs into the decision-making process is to reduce uncertainty, identify risks and opportunities, and, by doing so, deepen understanding so that those with policymaking responsibilities will make "better" decisions." (Ibid., 25). Cf. Friedman and Zeckhauser who argue that this widespread belief about the aim of intelligence analysis is misguided as it "can impair the accuracy, clarity, and utility of intelligence estimates. These problems frequently fall under one of two complementary categories. Consequence neglect occurs when collectors, analysts, and consumers of intelligence focus too much on the probability of each possible scenario and too little on the magnitude of those scenarios' potential consequences. Probability neglect is the reverse problem, arising when intelligence focuses predominantly on the potential consequences of various possibilities while giving less attention to their respective likelihoods. When likelihoods and consequences are not identified separately and then considered together, estimative intelligence will be incomplete, unclear, and subject to misinterpretation." (Friedman and Zeckhauser, "Assessing Uncertainty in Intelligence", 824-5)

[15] Kent, "Words of Estimative Probability".

[16] Immerman, "Transforming Analysis", 163.

[17] Cardillo, "A Cultural Evolution", 44.

[18] Ibid., 43.

[19] Gentry, "Has the ODNI Improved", 641.

[20] Betts, "Analysis, War, and Decision", 61. This view seems to have been shared by Thomas Schelling (see quote above).

[21] Gentry, "Intelligence Failure Reframed", 266ff. A similar argument can be found in Immerman, "Intelligence and the Iraq", 477.

[22] Immerman, "Intelligence and the Iraq".

[23] Phythian, "The Perfect Intelligence Failure".

[24] Ibid., 248.

[25] Ibid., 249. A similar account can be found in Silberman and Robb, *Report to the President*, 100-105.

[26] Zegart, "September 11", 79-80. See also Zegart, "An Empirical Analysis", 59.

[27] Zegart, "An Empirical Analysis", 54.

[28] Gentry, "Has the ODNI Improved", 638.

[29] Gentry, "Has the ODNI Improved"; Marchio, "Analytic Tradecraft".

[30] Gentry, "Has the ODNI Improved"; Lowenthal, "A Disputation on Intelligence Reform".

[31] Gentry, "Has the ODNI Improved".

[32] Lowenthal, "A Disputation on Intelligence Reform", 32. See also Gentry, "Has the ODNI Improved".

[33] Gentry, "Intelligence Failure Reframed", 252.

[34] Betts, "Analysis, War, and Decision", 85.

[35] See IRTPA and the discussion in Marrin, "Training and Educating"; and Artner et al., *Assessing the Value*.

[36] Heuer and Pherson, *Structured Analytic Techniques*, 4.

[37] Coulthart, "An Evidence-Based Evaluation".

[38] Ibid., 369.

[39] Chang et al, "Restructuring Structured Analytic Techniques", 1.

[40] Ibid., 4.

[41] Fingar, *Keynote Address*, 5.

[42] Marrin, "Evaluating the Quality of Intelligence".

[43] Marchio, "Analytic Tradecraft".

[44] Friedman and Zeckhauser, "Why Assessing Estimative Accuracy".

[45] "While getting a judgment 'right' is what ultimately matters most, the recipients of IC analytic products recognize that strong analytic tradecraft is more likely to result in assessments that are relevant and rigorous – what they need and value most" (Marchio, "Analytic Tradecraft", 182)

[46] Tetlock and Mellers, "Intelligent Management of Intelligence Agencies", 549.

[47] Lowenthal, "Towards a Reasonable Standard"; Gentry, "Has the ODNI Improved".

[48] Fingar, *Reducing Uncertainty*, 109-111, 129-131; Friedman and Zeckhauser, "Assessing Uncertainty in Intelligence", fn. 9.

[49] Note, however, Immerman's remark that "based on a small set of studies undertaken, there does seem to be a correlation between outstanding tradecraft and the accuracy of the product." (Immerman, "Transforming Analysis", 172) We did not have access to the studies Immerman is referring to.

[50] *Apud*. Friedman and Zeckhauser, "Why Assessing Estimative Accuracy".

[51] Ibid., 185.

[52] Ibid., 186.

[53] Ibid., 187.

[54] Ibid., 189.

[55] Ibid., 191.

[56] Marrin, "Evaluating the Quality of Intelligence".

[57] Marrin and Clemente, "Improving Intelligence Analysis"; Marrin and Clemente, "Modelling and Intelligence Analysis".

[58] Institutional approval for this study was gained on 20-Sep-2017 (reference: 17IC4179) and the study was preregistered on the *Open Science Framework*.

[59] van Gelder and de Rozario, "Pursuing Fundamental Advances".

[60] https://www.goodjudgment.com/.

[61] We did not have access to the aggregating method used for the *Rating Scale* in the practice of AIS.

[62] Lowenthal, "Towards a Reasonable Standard", 307.

[63] See Jonsson and Svingby, "The Use of Scoring Rubrics"; Judd et al., *Being Confident about Results*; Reddy and Andrade, "A Review of Rubric Use"; RiCharde "The Humanity versus Interrater Reliability", Turbow and Evener, "Norming a VALUE Rubric".

[64] Due to no-shows only 10 out of the 13 reports used in Experiment 2 were evaluated by 3 raters.

[65] The same selection of reports in Experiment 1 generated very low agreement for both the Equal Weights and the Weighted scoring system, ICC ≈ 0 (95% CI [-1.750,0.632]) and ICC= ≈ 0 (95% CI [-1.989,0.600]), respectively.

[66] Immerman, "Transforming Analysis" 168.