Defence Research and Development Canada
Recherche et développement pour la défense Canada

CAN UNCLASSIFIED

DRDC | RDDC
technologysciencetechnologie

Standards for Evaluating Source Reliability and Information Credibility in Intelligence Production

Daniel Irwin
David Mandel
DRDC – Toronto Research Centre

NATO
Assessment and Communication of Uncertainty in Intelligence to Support Decision-Making
Chapter 7
AC/323 (SAS-114) TP/928
STO-TR-SAS-114
pp. 7-1–7-16

Date of Publication from Ext Publisher: May 2020

Defence Research and Development Canada

External Literature (N)
DRDC-RDDC-2020-N237
December 2020

CAN UNCLASSIFIED

Canada

IMPORTANT INFORMATIVE STATEMENTS

This document was reviewed for Controlled Goods by Defence Research and Development Canada using the Schedule to the *Defence Production Act*.

Disclaimer: This document is not published by the Editorial Office of Defence Research and Development Canada, an agency of the Department of National Defence of Canada but is to be catalogued in the Canadian Defence Information System (CANDIS), the national repository for Defence S&T documents. Her Majesty the Queen in Right of Canada (Department of National Defence) makes no representations or warranties, expressed or implied, of any kind whatsoever, and assumes no liability for the accuracy, reliability, completeness, currency or usefulness of any information, product, process or material included in this document. Nothing in this document should be interpreted as an endorsement for the specific use of any tool, technique or process examined in it. Any reliance on, or use of, any information, product, process or material included in this document is at the sole risk of the person so using it or relying on it. Canada does not assume any liability in respect of any damages or losses arising out of or in connection with the use of, or reliance on, any information, product, process or material included in this document.

Chapter 7 – STANDARDS FOR EVALUATING SOURCE RELIABILITY AND INFORMATION CREDIBILITY IN INTELLIGENCE PRODUCTION[1]

Daniel Irwin and David R. Mandel
Defence Research and Development Canada
CANADA

7.1 INTRODUCTION

Intelligence practitioners must regularly exploit information of uncertain quality to support decision making [1]. Whether information is obtained from a human source or an automated sensor, failure to assess and communicate its characteristics may contribute to intelligence failure [2], [3]. This is evident in the case of Curveball, the Iraqi informant who fabricated extensive testimony on Saddam Hussein's alleged Weapons of Mass Destruction (WMD) [4], [5]. Subjected to inadequate scrutiny, Curveball's false allegations underpinned the 2002 National Intelligence Estimate on Iraq's WMD programs, and may have influenced the ill-fated decision to invade Iraq in 2003 [4], [5].

Recognizing information evaluation as a key function within the intelligence process, some organizations provide standards for assessing and communicating relevant information characteristics. Despite their intent, however, many of these standards are inconsistent across organizations, and may be fundamentally flawed or otherwise ill-suited to the context of application. In certain situations, poorly formulated standards may actually inhibit collaboration, degrade the quality of analytic judgements, and impair decision making.

In order to develop evidence-based recommendations for future practice in the assessment and communication of information quality, SAS-114 collected standards in use across a variety of agencies and domains. The following chapter provides a critical examination of standards for evaluating source reliability and information credibility, and highlights avenues for future research and development.

7.2 OVERVIEW OF CURRENT STANDARDS

The information evaluation criteria presented in Allied intelligence doctrine is known as the Admiralty Code or NATO System [6]. Developed by the Royal Navy in the 1940s, the system has undergone little change since its inception, and forms the basis of standards used by several Alliance members, as well as organizations in other domains [7], [8]. Under the Admiralty Code, information is assessed on two dimensions: source reliability and information credibility. Users are instructed to consider these components independently and to rate them on two separate scales (Table 7-1). The resultant rating is expressed using the corresponding alphanumeric code (e.g., *probably true* information from a *usually reliable* source is rated B2). Both scales include an option to be used when there is an inability to assess ('F' for source reliability and '6' for information credibility). Thus, ratings 'F' and '6' are not part of the ordinal scales comprised of ratings A – E and 1 – 5, respectively.

The defunct NATO Standardization Agreement (STANAG) 2511 (superseded in terms of information evaluation standards by Ref. [9]) provides a more detailed version of the Admiralty Code, and is presented for historical reference in Table 7-2 and Table 7-3 [10]. In line with many of the standards examined, NATO STANAG 2511 includes a qualitative description for each reliability and credibility rating. Source reliability is conceptually linked to "confidence" in a given source, based on past performance, while information

[1] Funding support for this work was provided by the Canadian Safety and Security Program Project CSSP-2016-TI-2224 (Improving Intelligence Assessment Processes with Decision Science).

credibility reflects the extent to which new information conforms to previous reporting. It is also worth noting that NATO STANAG 2511 uses *confirmed by other sources* as its highest information credibility rating, where current Allied doctrine substitutes *completely credible.*

Table 7-1: NATO AJP 2.1 2016 Source Reliability and Information Credibility Scales [9].

Reliability of the Collection Capability		**Credibility of the Information**	
A	Completely reliable	**1**	Completely credible
B	Usually reliable	**2**	Probably true
C	Fairly reliable	**3**	Possibly true
D	Not usually reliable	**4**	Doubtful
E	Unreliable	**5**	Improbable
F	Reliability cannot be judged	**6**	Truth cannot be judged

Table 7-2: NATO STANAG 2511 Source Reliability Scale [10].

Reliability of Source		
A	Completely reliable	Refers to a tried and trusted source which can be depended upon with confidence.
B	Usually reliable	Refers to a source which has been successful in the past but for which there is still some element of doubt in a particular case.
C	Fairly reliable	Refers to a source which has occasionally been used in the past and upon which some degree of confidence can be based.
D	Not usually reliable	Refers to a source which has been used in the past but has proved more often than not unreliable.
E	Unreliable	Refers to a source which has been used in the past and has proved unworthy of any confidence.
F	Reliability cannot be judged	Refers to a source which has not been used in the past.

Critical examination of these standards and others collected by SAS-114 exposes a number of weaknesses and inconsistencies. Given the extensive influence of the Admiralty Code, and efforts by many Alliance members to conform to NATO doctrine, the issues outlined below are common across most of the standards examined.

7.2.1 Semantic Issues

Under the Admiralty Code, qualitative ratings of reliability and credibility form a demonstrably intuitive progression [11]. However, subjective interpretations of the boundaries between these ratings are likely to vary among users, as are interpretations of the relevant rating criteria [12]. For instance, in many versions of the Admiralty Code, a *reliable* ('A') source is said to have a "history of complete reliability," while a *usually reliable* ('B') source has a "history of valid information most of the time" [2], [13], [14], [15], [16], [17]. None of the standards examined associate these descriptions with numerical values (i.e., 'batting averages'), potentially leading to miscommunication. One analyst may assign *usually reliable* to sources that provide valid information > 70% of the time. An analyst receiving this rating may interpret it to mean valid information > 90% of the time, and place more confidence in the source than is warranted. Conversely, an analyst may assume *usually reliable* reflects valid information only > 50% of the time, and prematurely discount the source. Asked to assign absolute probability values to reliability and credibility ratings, US intelligence officers demonstrated considerable variation in their interpretations [11]. For example, probabilistic interpretations of *usually reliable* and *probably true* ranged from .55 to .90 and .53 to .90, respectively, while interpretations of *fairly reliable* and *possibly true* both ranged from .40 to .80 [11].

Table 7-3: NATO STANAG 2511 Information Credibility Scale [10].

Credibility of Information		
1	Confirmed by other sources	If it can be stated with certainty that the reported information originates from another source than the already existing information on the same subject, it is classified as "confirmed by other sources" and is rated "1".
2	Probably true	If the independence of the source of any item or information cannot be guaranteed, but if, from the quantity and quality of previous reports its likelihood is nevertheless regarded as sufficiently established, then the information should be classified as "probably true" and given a rating of "2".
3	Possibly true	If, despite there being insufficient confirmation to establish any higher degree of likelihood, a freshly reported item of information does not conflict with the previously reported behaviour pattern of the target, the item may be classified as "possibly true" and given a rating of "3".
4	Doubtful	An item of information which tends to conflict with the previously reported or established behaviour pattern of an intelligence target should be classified as "doubtful" and given a rating of "4".
5	Improbable	An item of information which positively contradicts previously reported information or conflicts with the established behaviour pattern of an intelligence target in a marked degree should be classified as "improbable" and given a rating of "5".
6	Truth cannot be judged	Any freshly reported item of information which provides no basis for comparison with any known behaviour pattern of a target must be classified as "truth cannot be judged" and given a rating of "6". Such a rating should be given only when the accurate use of higher rating is impossible.

Among the standards examined, *reliable* or *completely reliable* indicates maximum source reliability, while *confirmed*, *confirmed by other sources*, or *completely credible* marks the highest degree of information credibility. Despite these inconsistencies, most scales faithfully reproduce the Admiralty Code's A – F (reliability) / 1 – 6 (credibility) scoring scheme, and ratings are often communicated using only the appropriate alphanumeric code (e.g., A1). These terminological variations may therefore contribute to miscommunication between users familiar with different standards. For instance, under most US standards examined [13], [14], [16], [17] 'A' is defined as *reliable*, while UK Joint Doctrine 2-00 [18] defines 'A' as *completely reliable* (conforming to NATO doctrine). A US analyst who understands 'A' to mean *reliable* might transmit that rating to a UK counterpart, who interprets it as *completely reliable*. This translation is potentially problematic, given that an analyst or consumer may place more weight on a source labelled *completely reliable* than one labelled *reliable*. Alternatively, the translation from *completely reliable* to *reliable* could lead a recipient to undervalue a source.

Inter-standard miscommunication could also arise where scales use the term 'accuracy' as a synonym for information credibility (e.g., Refs. [14], [19]). While credibility often includes considerations of accuracy, it is likely a more multidimensional construct. Credibility generally incorporates criteria that can serve as cues to accuracy, but which are not equivalent to accuracy (e.g., triangulating evidence contributes to credibility, but does not require ground truth). Thus, this use of 'accuracy' by certain standards may further diversify interpretations of ratings, as well as the determinants considered during evaluation.

Another semantic issue relates to liberal use of terms conveying certainty (e.g., *confirmed*). In intelligence contexts where the information "is always incomplete... [and] frequently ambiguous," [20] these expressions could lead to overconfidence on the part of consumers. Compounding this issue is the observed tendency of analysts to confine their ratings to the high ends of the scales [21]. In their review of spot reports completed during a US Army field exercise, Baker, McKendry, and Mace [21] found that A1 and B2 represented 80% of all reliability/credibility ratings, with B2 alone comprising 74% of ratings. Allied intelligence doctrine explicitly discourages statements of certainty "given the nature of intelligence projecting forward in time" [9]. However, it remains unverified whether *completely credible* actually conveys less certainty than *confirmed by other sources*. A piece of information may be confirmed by some sources and simultaneously disconfirmed by others (an issue further explored in Section 7.2.3). Researchers could measure subjective interpretations of *completely credible* and compare them with interpretations of *confirmed by other sources*.

7.2.2 Source Reliability Determinants

To address miscommunication stemming from vague source history descriptors, this determinant could be quantified (e.g., source reliability = accurate information provided / total information provided). A quantitative method of tracking and updating source history could improve consistency and streamline the information evaluation process [22]. However, this would fail to address the Admiralty Code's implicit treatment of source reliability as constant across different contexts [12]. Regardless of past performance, source reliability may vary dramatically depending on the type of information provided, characteristics of the source(s), and the circumstances of collection. A Human Intelligence (HUMINT) source with a proven track record reporting on military operations may lack the expertise to reliably observe and report on economic developments. Beyond variable expertise, HUMINT source motivations, expectations, sensitivity, and recall ability may shift between situations, with major implications for information quality [23], [24]. Even the reliability of an 'objective source' (i.e., a sensor) is highly context dependent [25]. For example, inclement weather may compromise the quality of information provided by an optical sensor, despite a history of perfect reporting under ideal conditions. Future research could evaluate means of incorporating contextual information into ratings of source reliability, or into a more holistic measure of information quality.

Aside from source history, most of the standards examined highlight reliability determinants such as "authenticity," "competency," and "trustworthiness." The inclusion of these determinants is consistent with the broader literature on source reliability [23], [24]. However, the extant standards fail to formally define or operationalize these concepts. Their inclusion is therefore likely to increase subjectivity and further undermine the fidelity of reliability assessments. The standards examined also fail to operationalize the qualifiers used to describe each level. For instance, reliability ratings often incorporate whether an evaluator has "minor doubt," "doubt," or "significant doubt" about the source's authenticity. Aside from being vague, the use of modifiers ("minor," "significant") for some levels, and the unmodified term ("doubt") for another, is problematic because the unmodified term effectively subsumes the modified cases. Chang *et al.* [26] describe how a process designed to decompose and evaluate components of a problem (i.e., information characteristics) may amplify unreliability in assessments if that process is ambiguous and open to subjective interpretations. Given the ambiguity built into current standards, users are unlikely to retrieve every relevant determinant, let alone reliably and validly weigh every relevant determinant when arriving at an ordinal assessment.

Another issue with current source reliability standards is their failure to delineate procedures for evaluating ‘subjective sources’ vs. ‘objective sources’ (e.g., human sources vs. sensors) [27], or primary sources vs. secondary/relaying sources [28]. A determinant such as source motivation may be relevant when assessing HUMINT sources, but not sensors. Similarly, source expertise may be highly relevant for a primary source collecting technical information (e.g., a HUMINT asset gathering information on Iranian nuclear technology), but less so for an intermediary delivering this information to a collector. In cases where information passes through multiple sources, there are often several intervals where source reliability considerations are relevant [28]. For instance, when receiving second-hand information from a HUMINT source, one might consider the reliability of the primary source, the reliability of the secondary/relaying source(s), the reliability of the collector, as well as the reliability of any medium(s) used to transmit the information [28].

Following initial collection, Nobel [29] describes how information may undergo distortion at other stages of the intelligence process. Just like sources, intelligence practitioners will vary in terms of their ability to reliably assess and relay information. For instance, an economic subject matter expert may lack the expertise to accurately evaluate and transmit information on enemy troop movements. Beyond expertise, an intelligence practitioner’s assessment is also undoubtedly influenced by his/her personal characteristics (e.g., motivation, expectations, biases, recall ability) as well as various contextual factors [12], [23], [29]. When a finished intelligence product is edited and approved for dissemination, managers may inject additional distortion by adjusting analytic conclusions [29]. The many opportunities for distortion may warrant the formalization of information evaluation as an ongoing requirement throughout the intelligence process (see Section 7.4) [12]. At the very least, efforts should be made to ensure intelligence practitioners and consumers are cognizant of the mutability of information characteristics following the initial evaluation.

7.2.3 Information Credibility Determinants

Much like the source reliability standards examined, most of the information credibility scales suffer from an inherent lack of clarity. Information credibility generally incorporates confirmation “by other independent sources” as a key determinant. However, no guidance is provided as to how many independent sources must provide confirmation for information to be judged credible. Where one analyst considers confirmation by two sources sufficient for a *confirmed* rating, another might seek verification by three or more. Perceptions of how much corroboration is necessary may also vary depending on the information in question. For instance, an analyst may decide that a particularly consequential piece of information requires more corroboration than usual to be rated *confirmed*. This lack of consistency could lead analysts to misinterpret each other’s credibility ratings, and consider pieces of information more or less credible than intended.

The information credibility standards examined also lack instructions for grading pieces of information that are simultaneously confirmed and disconfirmed. Under the Admiralty Code, such information could be considered both *confirmed / completely credible* (‘1’) and *improbable* (‘5’) [25]. Without guidance, some analysts may base their assessments more heavily on confirmed information, while others focus on disconfirmed information, or pursue a balance between confirmed and disconfirmed. These three approaches could generate very different assessments, despite evaluating the same information.

Capet and Revault d’Allonnes [12] argue that confirmation does not, in itself, translate into information credibility, and that not all forms of confirmation should be weighted equally. Theoretically, a spurious rumour corroborated by many unreliable sources (e.g., tweets about a second shooter during a terrorist attack), and disconfirmed by a single reliable source (e.g., a police statement indicating a single attacker), could still be rated highly credible under current standards. Capet and Revault d'Allonnes [12] advocate identifying a threshold whereby information must be confirmed by a clear majority, and undermined by few or no sources, while accounting for source reliability. This would directly contravene the Admiralty Code's treatment of source reliability and information credibility as independent.

Lesot, Pichon, and Delavallade [30] note that current standards lack consideration of whether relationships of affinity, hostility, or independence[2] exist between corroborating sources. Corroboration from a source that has a 'friendly' relationship with the source under scrutiny should likely have less influence than corroboration from an independent or hostile source. For example, all else being equal, if Saudi Arabia corroborates information provided by Syria (with which it has a hostile relationship), that confirmation should carry more weight than identical confirmation provided by Russia (which has a relationship of affinity with Syria). As a general rule, friendly sources should be expected to corroborate each other [30].

Friedman and Zeckhauser [31] suggest that the current emphasis on consistency with existing evidence may encourage confirmation bias. "Biased attrition" is used to describe an information filtering process that systematically favours certain information types in a problematic way. Information that conflicts with prior beliefs and analysis may in fact be more valuable, as it can shift the views of analysts and consumers more significantly. Friedman and Zeckhauser [31] argue that credibility standards could reduce biased attrition by incorporating the extent to which information provides a new or original perspective on the intelligence requirement at hand. Capet and Revault d'Allonnes [12] also suggest that current standards be modified to gauge the extent to which information provides "meaningful" corroboration.

Along similar lines, Lemercier [28] notes that confirmation-based credibility standards do not account for the phenomenon of amplification, whereby analysts come to believe closely correlated sources are independently verifying a piece of information. In order to control for amplification, credibility evaluation could incorporate successive corroboration by the same source, corroboration by sources of the same type, as well as comparative corroboration from different collection disciplines [28].

The current emphasis placed on confirmation/consistency may also reinforce order effects, given that new information must conform to prior information to be deemed credible. All else being equal, if an analyst receives three new pieces of information, the first item received will typically face the fewest hurdles to being assessed as credible. Meanwhile, the second piece of information must conform to the first, and the third must conform to both the first and second. Under this system, an analyst may inadvertently underweight information that is in fact more accurate or consequential than information received earlier, potentially decreasing the quality of analysis. One option for dealing with order effects would be the formal inclusion of mechanisms to revaluate prior pieces of information as new information becomes available (a prospect that is further explored in Section 7.4). Two of the US standards examined [17], [19] advocate continuous analysis and re-evaluation of source reliability / information credibility as new information becomes available. However, neither document outlines a specific method for revaluation.

Beyond confirmation, most of the information credibility scales examined incorporate consideration of whether an item is "logical in itself." Current standards do not specify whether this simply refers to the extent that information conforms to the analyst's current assessment. Furthermore, the use of "not illogical" as a level between "logical in itself" and "illogical in itself" is nonsensical, as "not illogical" effectively means "logical" (in itself).

As noted with regards to source reliability, the Admiralty Code's one-size-fits-all approach to information credibility neglects important contextual considerations. Several US standards suggest that credibility determinants have more relevance depending on the collection discipline(s) utilized. For example,

[2] Several credibility standards do call for the independence of corroborating sources (e.g., Refs. [13], [15]), but none examined consider other types of relationships.

TC 2-91.8 [14] and ATP 2-22.9 [16] suggest that there is a greater risk of deception (an information credibility determinant) when utilizing Open-Source Intelligence (OSINT) than Captured Enemy Documents (CEDs). Similarly, ATTP 2-91.5 [19] refers to the Admiralty Code as the "HUMINT system," and recommends the development of separate rating systems to assess the three basic components of document and media exploitation (Document Exploitation [DOMEX], Media Exploitation [MEDEX], Cellphone Exploitation [CELLEX]).

Joseph and Corkill [32] stress that the Admiralty Code is a grading system rather than an evaluation methodology. Beyond what is outlined in the scales, evaluators may have a formal assessment procedure and/or a more exhaustive list of determinants to consider. Supplementary documents add some clarity to the standards examined, but also vary in terms of which determinants are identified and emphasized. Additionally, none of these extra determinants are defined or operationalized, and may further contribute to subjectivity. The following factors are highlighted in one or more of the documents examined:

Reliability:

- Circumstances under which information was obtained;
- Quality of source's bona fides; and
- Sensor capabilities.

Credibility:

- Internal and external consistency;
- Risk of denial and deception;
- Timeliness/recency; and
- Unusual absence of evidence.

Overlapping Factors[3]:

- Source access;
- Source expertise/authority; and
- Source motivation.

7.3 CONCEPTUALIZING INFORMATION QUALITY

As noted previously, the Admiralty Code is predicated on the independence of source reliability and information credibility. In developing a comprehensive, evidence-based means of information evaluation, an initial step would be to evaluate the independence of these constructs, and to assess whether they are unidimensional (e.g., credibility is understood as the probability that information is accurate) or multidimensional (e.g., credibility is understood as a profile of several determinants, including internal consistency, external consistency, timeliness, risk of deception, etc.) [12]. If the current terminology is found to be unidimensional, further experimentation could yield a list of qualitative terms with narrower and more consistent interpretations. Alternatively, if the meaning of the current terminology is multidimensional, this may warrant the creation of new scales to gauge factors comprising a comprehensive measure of information quality (e.g., information quality = a function of timeliness rated from 0 – 5; risk of deception rated from 0 – 5; source reporting history rated from 0 – 5, etc.). This latter approach would resemble the UK Defence Intelligence pilot approach to assessing analytic confidence.

[3] According to UK JDP 2-00 [18], these factors affect both source reliability and information credibility.

Several studies support the introduction of a single measure of perceived information quality (i.e., accuracy/truthfulness) incorporating all available information, including source reliability. Analysts are shown to pair reliability and credibility scores from the same level [8], [21] or to base decisions about accuracy more on credibility than reliability [22]. Nickerson and Feehrer [33] note that when no other information is available to gauge information credibility, analysts will logically base their rating on source reliability, given that reliable sources tend to produce credible information. To this point, Lemercier [28] posits that determining source reliability is not an end in itself, but rather a means of assessing information credibility, which he suggests is the ultimate goal of the evaluator.

A single measure of accuracy/truthfulness could address several challenges related to incongruent ratings and the lack of comparability between the two scales [12]. Samet [22] shows that analysts assign likely accuracy/truthfulness less reliably when basing their decision on separate reliability and credibility metrics, than a single measure. Similarly, in a preliminary analysis, Mandel, Dhami, Weaver, and Timms (cited in Ref. [34]) find that analysts show poor test-retest reliability when estimating the accuracy of information with incongruent reliability/credibility scores (e.g., A5, E1). Mandel *et al.* [34] also show that inter-analyst agreement plummets as source reliability and information credibility scores become less congruent. This suggests that while the two scales may be distinct in theory, in practice, users do not treat them as such. The ambiguity inherent in combining incongruent ratings may partially explain why evaluators often default to ratings from the same level [12].

Beyond issues stemming from incongruent ratings, currents standards also lack mechanisms for comparing multiple items of varying quality, which is a regular requirement for intelligence analysts [35]. For instance, it is unclear how analysts should weigh one piece of information rated B3 against another rated C2. The margin of interpretation may be increased by the use of two different scale types; credibility comprises a positive-negative scale (information is confirmed/invalidated), while reliability ranges from low/non-existent (the source has provided little/no credible information) to a maximum level (the source has a history of complete reliability) [12]. Without any sort of fusion methodology, this represents another sensitive process left to the subjective judgement of individual analysts [36]. The lack of comparability between scales also means that current measures of reliability and credibility are ill-suited for integration into an automated or semi-automated system for information evaluation [25]. A single, comprehensive measure of information quality would likely be more conducive to the collation of information of varying quality.

If quantified, such a measure could also enable users to grade information with finer discrimination. While Samet [11] finds that users can make quantitative distinctions between the five levels in each scale, the average size of the difference between the mean probabilities assigned by users to adjacent levels indicates there is room for greater precision. These findings are similar those of Friedman *et al.* [37] in the context of qualitative probability assessments. They find that analysts can assign probabilities more precisely than conventional wisdom supposes, and argue that the imprecision built into current standards sacrifices predictive accuracy. The inclusion of empirically grounded numerical values could also mitigate language barriers and some of the inter-standard semantic issues identified, while improving collaboration and analyst accountability [22].

In developing a comprehensive measure of information quality, it would be necessary to consider the hierarchy of relevant determinants, possible interactions or tradeoffs between determinants, as well as their importance depending on context and end-user information requirements [27], [38]. For example, confirmation may be less important than (or even conflict with) considerations of timeliness, where the pursuit of confirmation translates into unacceptable decision latency [38]. As noted previously, certain determinants (e.g., motivation) may be completely irrelevant depending on the information under scrutiny. In general, a determinant of information quality can be deemed relevant if a change of its value impacts the hypotheses under consideration; the levels of belief assigned to those hypotheses; or the utility values assigned to a set of potential courses of action [27].

Rogova [27] provides ontologies of quality of information content (Figure 7-1) and quality of information sources (Figure 7-2) that were derived from the broader literature on information fusion for decision support. Detailed explanations of each concept within these ontologies and the evidential basis for the concepts are beyond the scope of this chapter but can be found in Ref. [27]. Our purpose in presenting these ontologies is simply to highlight that a comprehensive measure of information quality developed for intelligence analysis might incorporate many of the determinants identified. These models also depict the interconnectedness of information characteristics, which current standards fail to address. Researchers formulating an intelligence-oriented model of information quality could evaluate the prospects of combining these ontologies, and the utility of other quality measures identified in the information fusion scholarship.

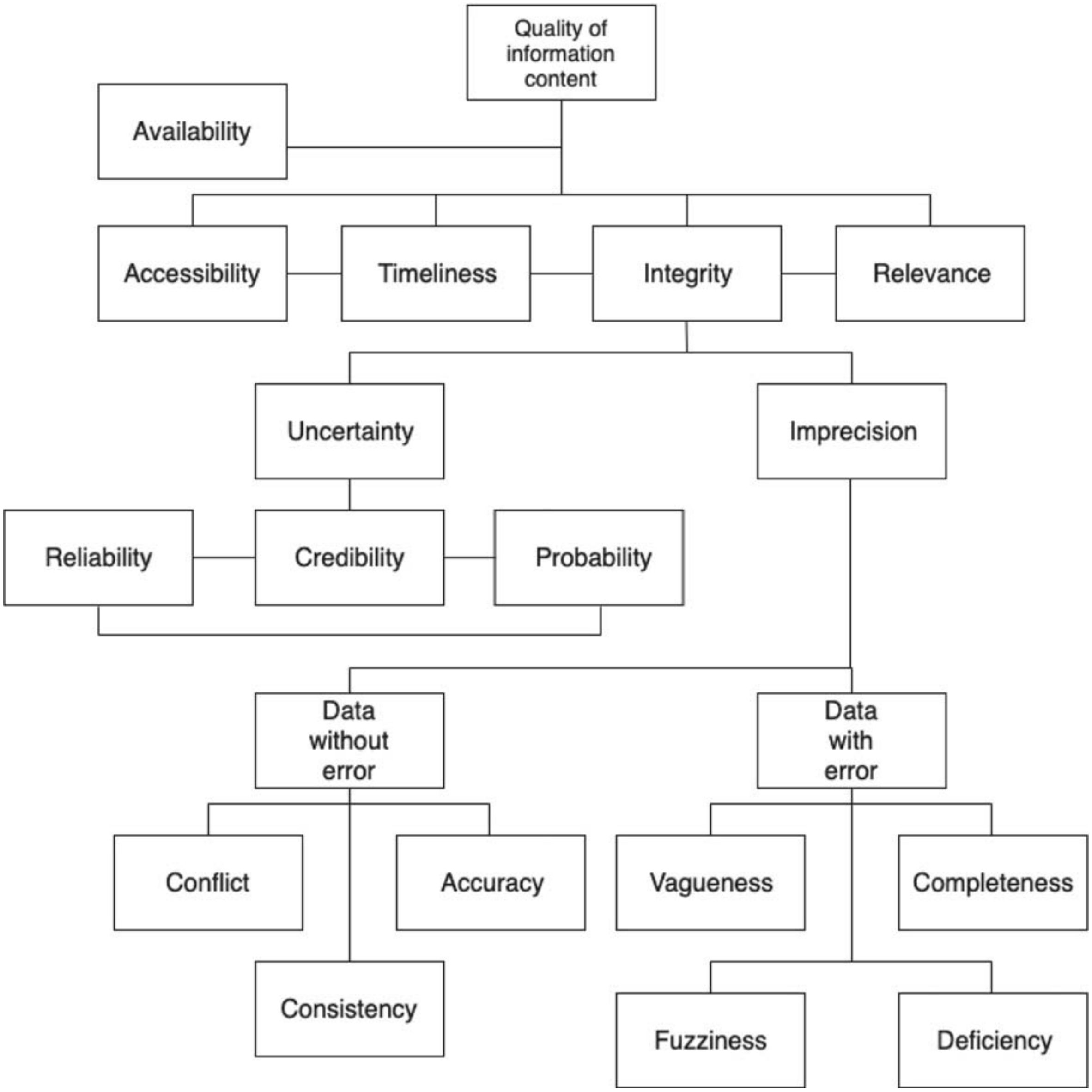

Figure 7-1: Rogova's Ontology of Quality of Information Content [27].

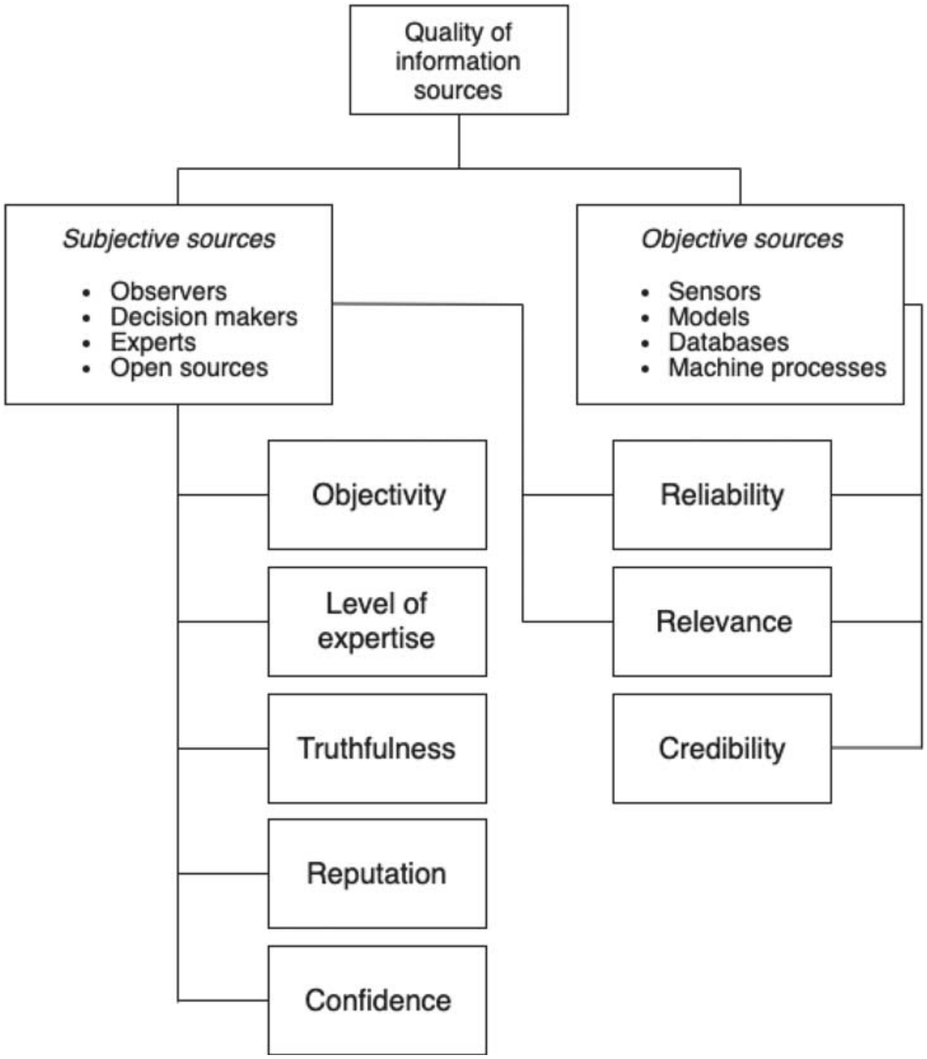

Figure 7-2: Rogova's Ontology of Quality of Information Sources [27].

7.4 ALTERNATIVE APPROACHES TO INFORMATION EVALUATION

After determining the relationship between source reliability and information credibility, and formulating a comprehensive measure (or measures) of information quality, the next task for researchers would be to identify possible models for evaluation. A wide range of information evaluation systems have been proposed, but few have undergone rigorous empirical evaluation, particularly in intelligence contexts [4], [35], [39], [40]. Once promising models are identified, researchers could evaluate them using intelligence practitioners and realistic intelligence problems.

Before reviewing alternative approaches to information evaluation, it is worth noting the variation among current standards in terms where information evaluation is situated within the intelligence process. For instance, NATO intelligence doctrine [9] embeds evaluation procedures within the processing stage, thus emphasizing the analyst's role in gauging information characteristics. UK JDP 2-00 [18] outlines a joint role for analysts and collectors, whereby collectors pre-rate information characteristics before analysts weigh in with their own (potentially broader) understanding of the subject. Refs. [3], [24], and [29] stress the primary collector’s role in assessing reliability, particularly in contexts where access to a clandestine source is restricted to the agent handler. This inconsistency is significant given the noted mutability of information characteristics over time, across contexts, and at different stages of the intelligence process itself [23], [24], [25], [29]. Whether information is assessed upon initial collection, by an analyst during processing, or by several practitioners throughout the intelligence process could have a substantial impact on its reliability/credibility evaluation. Consequently, the extent to which information is deemed fit to use will largely determine its influence (or lack thereof) on analytic judgements. In other words, the timing of source and information evaluation within the intelligence process could add additional inter-analytic unreliability to such meta-informational ratings and, by proxy, to analytic judgements themselves. For instance, an intriguing question is whether information evaluation is given more weight in intelligence analysis when the evaluation step is also conducted by the analyst rather than by the collector. Moreover, do individual differences in analyst characteristics play a role in how the source of meta-information is treated in subsequent analysis. Perhaps analysts who have a disposition of high self-confidence place more weight in such evaluations when they rendered them, whereas analysts who are perennial self-doubters might give more weight to evaluations that come from other sources. These hypotheses could be tractably tested in future research.

A compounding issue is the absence of mechanisms for revaluation when new information becomes available and determinants, such as a source’s reliability rating, are updated [28]. For example, under current standards, it is unclear how users should treat information provided by a source long considered *completely reliable*, but suddenly discovered to be *unreliable*. This is particularly complicated when information ratings form an interdependent chain (e.g., Info A’s rating is tied to Info B’s rating; Info B’s rating is tied to the rating of Source X; Source X has just been exposed as a double agent). Together, these issues may warrant the implementation of information evaluation as an iterative function throughout the intelligence process. This approach could be applied to the evaluation of individual pieces of information [4], as well as the marshalling of evidence when forming analytic judgements [35]. As noted, certain US standards [17], [19] advocate continuous revaluation of information quality, but none of them provide formalized mechanisms for doing so.

Bayesian networks could represent one means of capturing complicated interactions between items of evidence, which may influence each other’s reliability/credibility [1]. As new information becomes available, Bayesian networks can be updated coherently; that is, respecting the axioms of probability theory, such as unitarity, additivity, and non-negativity [41]. This process may reduce the systematic errors exhibited by individuals estimating the impact of less than totally reliable information [1]. Integrating Bayesian methods (or probabilistic approaches, more generally) into information evaluation could also improve the fidelity of assessments where sources communicate information using probabilistic language. Current standards provide no guidance for incorporating probabilistic expressions into a broader evaluation of information quality determinants (e.g., if a *usually reliable* source reports that she *probably* saw two helicopters). Given the subjectivity inherent in interpreting probability phrases commonly used in intelligence production (for discussion, see Ref. [42]), this could be another source of miscommunication embedded in current standards.

McNaught and Sutovsky [35] propose using a Bayesian network as a computer-assisted framework to facilitate evidence marshalling, and the fusion of information of varying quality. While they suggest that such networks may help analysts explore uncertain situations and overcome cognitive biases, they warn that routine (as opposed to supplemental) use of these models could generate inaccuracy due to the challenges of estimating certain input parameters (e.g., quantifying the reliability of a HUMINT source, especially under conditions of anonymity). To this point, Rogova [27] argues that *a priori* domain knowledge is often essential when determining many of the input parameters in a system for assessing information quality. McNaught and Sutovsky [35] only advocate the use of Bayesian networks for evidence marshalling where the input parameters are known with a “reasonable degree” of accuracy. Simply put, a coherent integration of “garbage in”, which Bayesian approaches should ensure, will still yield “garbage out.”

The complete automation of information evaluation may be undesirable, given the requirement for analysts to easily understand and modify their inputs as new information becomes available [30]. To this end, Lesot, Pichon, and Delavallade [30] propose a semi-automated method for evaluating information derived from textual documents, which is based on a possibility framework for managing uncertainty resembling the structured analytic technique known as Analysis of Competing Hypotheses (ACH) [43]. Their method first identifies pieces of information relevant to the requirement at hand, and then attaches an independent level of confidence to each piece of information. These ratings are then combined to calculate an overall degree of confidence, which the analyst can attach to judgements derived from all available information. Lesot, Pichon, and Delavallade [30] suggest that their method automates a large portion of information evaluation, enabling the processing of large volumes of data, while giving analysts control over each stage of the process. Simulating a situation where high-quality information is provided by relatively reliable sources, they demonstrate the ability of the proposed method to identify an optimal aggregation operator with an information fusion function. They note that this process requires further evaluation under different conditions, as well as a review by domain specialists.

An alternative model is the computational Method for Assessing the Credibility of Evidence (MACE), designed by Schum and Morris [4] for application in HUMINT contexts. Incorporating both Baconian and Bayesian methods, MACE draws on procedures from the Anglo-American legal tradition for gauging the competence and credibility of witnesses. MACE first guides users through a Baconian analysis to assess how much evidence is available about a particular source, and how completely source competence and credibility (i.e., reliability) can be evaluated. Users answer 25 sequential questions related to source competence, veracity (the extent to which a source believes the information being relayed is true), objectivity, and observational sensitivity. The system tracks inputs for each question, as well as any questions that remain unanswered. While subjective, the final output (which also resembles an ACH matrix), is evidence-based and can be easily updated as more information becomes available. Referencing these scores, the second stage of MACE guides users as they generate three pairs of likelihood estimates related to source characteristics, which are plotted on a two-dimensional probability space. MACE then applies Bayesian probability methods to estimate the strength of evidence about a particular HUMINT source. The process yields an assessment of posterior odds favouring the believability of the source. In this second stage, MACE again enables users to easily update and modify their assessments. While MACE focuses on HUMINT sources, the question set could be extended to evaluate other types of information [44].

The use of such approaches could also be supplemented with training designed to improve collectors' and analysts' applied statistical skill. For example, Mandel [45] designed a brief (approximately 30-minute) training protocol on Bayesian belief revision and hypothesis testing with probabilistic information. Intelligence analysts were assessed on the accuracy and probabilistic coherence before and after receiving the training. Mandel [45] found statistically significant improvement after training on both accuracy and coherence of analysts' probability estimates, suggesting that intelligence professionals can reap quick wins in learning that might enable them to better understand the kinds of probabilistic models noted in the aforementioned examples. Similar encouraging results have been reported elsewhere [46], [47]. This type of training is not only important for understanding such models, however. People routinely violate logical constraints on probability assessments (e.g., see Refs. [41] and [48], [49], [50], [51]), and there is no good reason to believe that analysts are exempt. Indeed, the findings of Ref. [45] show that they are not.

The aforementioned evaluation systems are by no means exhaustive. Other mechanisms for information evaluation and the challenges they present can be found in Refs. [35], [39], [40]. Each system requires further experimentation. Furthermore, each system responds to a different analytic challenge: either the need to integrate information quality into analytic judgements [35]; to collate information of varying quality [30]; or to evaluate the quality of information *about* a source, as well as the quality of the source itself [4]. Despite these differences, each system is iterative, easily updated, and designed to be more reliable, transparent, and comprehensive than current evaluation standards. Researchers and intelligence practitioners ought to examine these and other systems as they pursue more effective methods of information evaluation in intelligence production. More generally, the intelligence community would be well served by taking a more evidence-based approach to verifying the effectiveness of its current methods and improving upon those where possible [52], [53], [54]. For far too long it has relied on developing analytic tradecraft methods that merely have to pass the test of apparent plausibility and effectiveness. Going forward, it should proactively leverage relevant information – theory, findings, and methods – from judgement and decision science.

7.5 REFERENCES

[1] Johnson, E.M., Cavanagh, R.C., Spooner, R.L., and Samet, M.G. (1973). Utilization of reliability measurements in Bayesian inference: Models and human performance. *IEEE Transactions on Reliability* 22 (3):176-182.

[2] United Nations Office on Drugs and Crime. (2011). *Criminal Intelligence Manual for Analysts*. Vienna, Austria.

[3] Carter, D.L. (2009). *Law Enforcement Intelligence: A Guide for State, Local, and Tribal Law Enforcement Agencies*, (2nd ed.). Washington DC: Office of Community Oriented Policing Services, US Department of Justice.

[4] Schum, D.A., and Morris, J.R. (2007). Assessing the competence and credibility of human sources of intelligence evidence: Contributions from law and probability. *Law, Probability and Risk,* 6(1-4):247-274.

[5] Betts, R.K. (2008). Two faces of intelligence failure: September 11 and Iraq's missing WMD. *Political Science Quarterly* 122 (4):585-606.

[6] Hanson, J.M. (2015). The admiralty code: A cognitive tool for self-directed learning. *International Journal of Learning, Teaching and Educational Research* 14 (1):97-115.

[7] United States Department of the Army. (1951). *Field Manual FM 30-5, Combat Intelligence*. Washington DC.

[8] Miron, M.S., Patten, S.M., and Halpin, S.M. (1978). *The Structure of Combat Intelligence Ratings*. Technical Paper 286. Arlington, VA: US Army Research Institute for Behavioral and Social Sciences.

[9] North Atlantic Treaty Organization (2016). *Allied Joint Doctrine for Intelligence Procedures AJP-2.1*. Brussels, Belgium.

[10] NATO Standardization Office. (2003). *STANAG 2511 – Intelligence Reports*, (1st ed.) Brussels, Belgium.

[11] Samet, M.G. (1975). *Subjective Interpretation of Reliability and Accuracy Sales for Evaluating Military Intelligence*. Technical Paper 260. Arlington, VA: US Army Research Institute for Behavioral and Social Sciences.

[12] Capet, P., and Revault d'Allonnes, A. (2014). Information evaluation in the military domain: Doctrines, practices, and shortcomings. In: *Information Evaluation*, Capet, P., and Delavallade, T. (Eds.), 103-125. Hoboken, NJ: Wiley-ISTE.

[13] United States Department of the Army. (2006). *Field Manual FM 2-22.3, Human Intelligence Collector Operations*. Washington DC.

[14] United States Department of the Army. (2010). *Training Circular TC 2-91.8, Document and Media Exploitation*. Washington DC.

[15] Department of National Defence. (2011). *Canadian Forces Joint Publication CFJP 2-0, Intelligence*. Ottawa, ON.

[16] United States Department of the Army. (2012). *Army Techniques Publication ATP 2-22.9, Open-Source Intelligence*. Washington DC.

[17] United States Department of the Army. (2012). *Army Techniques Publication ATP 3-39.20 Police Intelligence Operations*. Washington DC.

[18] United Kingdom Ministry of Defence. (2011). *Joint Doctrine Publication JDP 2-00, Understanding and Intelligence Support to Joint Operations*, (3rd ed.) Swindon, UK. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/311572/20110830_jdp2_00_ed3_with_change1.pdf.

[19] United States Department of the Army. (2010). *Document and Media Exploitation Tactics, Techniques, and Procedures ATTP 2-91.5* – Final Draft. Washington DC.

[20] Tecuci, G., Boicu, M., Schum, D., and Marcur, D. (2010). *Coping with the Complexity of Intelligence Analysis: Cognitive Assistants for Evidence-Based Reasoning*. Research Report #7, Learning Agents Center. Fairfax, VA: George Mason University.

[21] Baker, J.D., McKendry, J.M., and Mace, D.J. (1968). *Certitude Judgements in an Operational Environment.* Technical Research Note 200. Arlington, VA: US Army Research Institute for Behavioral and Social Sciences.

[22] Samet, M.G. (1975). Quantitative interpretation of two qualitative scales used to rate military intelligence. *Human Factors* 17 (2):192-202.

[23] Schum, D.A. (1987). *Evidence and Inference for the Intelligence Analyst*. Lanham, MD: University Press of America.

[24] Pechan, B.L. (1995). The collector's role in evaluation. In: *Inside CIA's Private World: Declassified Articles from the Agency's Internal Journal*, Westerfield, H.B. (Ed.), 99-107. New Haven, CT: Yale University Press.

[25] Cholvy, L., and Nimier, V. (2003). Information evaluation: Discussion about STANAG 2022 recommendations. In: *Proceedings of the NATO-IST Symposium on Military Data and Information Fusion.* Prague, Czech Republic.

[26] Chang, W., Berdini, E., Mandel, D.R. and Tetlock, P.E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security* 33 (3):337-356.

[27] Rogova, G.L. (2016). Information quality in information fusion and decision making with applications to crisis management. In: *Fusion Methodologies in Crisis Management, Higher Level Fusion and Decision Making*, Rogova, G.L., and Scott, P. (Eds.), 65-86. Cham, Switzerland: Springer International Publishing.

[28] Lemercier, P. (2014). The fundamentals of intelligence. In: *Information Evaluation*, Capet, P., and Delavallade, T. (Eds.), 55-100. Hoboken, NJ: Wiley-ISTE.

[29] Noble, G.P., Jr. (2009). Diagnosing distortion in source reporting: Lessons for HUMINT reliability from other fields. Master's thesis. Erie, PA: Mercyhurst College.

[30] Lesot, M., Pichon, F., and Delavallade, T. (2014). Quantitative information evaluation: Modeling and experimental evaluation. In: *Information Evaluation*, Capet, P., and Delavallade, T. (Eds.), 187-228. Hoboken, NJ: Wiley-ISTE.

[31] Friedman, J.A., and Zeckhauser, R. (2012). Assessing uncertainty in intelligence. *Intelligence and National Security* 27 (6):824-847.

[32] Joseph, J., and Corkill, J. (2011). Information evaluation: How one group of intelligence analysts go about the task. In: *Fourth Australian Security and Intelligence Conference*. Perth, Australia.

[33] Nickerson, R.S., and Feehrer, C.E. (1975). *Decision Making and Training: A review of Theoretical and Empirical Studies of Decision Making and Their Implications for the Training of Decision Makers*. Technical Report NAVTRAEQUIPCEN 73-C-0128-1. Cambridge, MA: Bolt, Beranek and Newman, Inc.

[34] Mandel, D.R. (2018). *Proceedings of SAS-114 Workshop on Communicating Uncertainty, Assessing Information Quality and Risk, and Using Structured Techniques in Intelligence Analysis*. NATO Meeting Proceedings. Brussels, Belgium: NATO STO.

[35] McNaught, K. and Sutovsky, P. (2012). Representing variable source credibility in intelligence analysis with Bayesian networks. In: *Fifth Australian Security and Intelligence Conference*, 44-51. Perth, Australia.

[36] Cholvy, L. (2004). Information evaluation in fusion: A case study. In: *Proceedings of the International Conference on Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2004*. Perugia, Italy.

[37] Friedman, J.A., Baker, J.D., Mellers, B.A., Tetlock, P.E., and Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly* 62 (2):410-422.

[38] Rogova, G., Hadzagic, M., St-Hillaire, M., Florea, M., and Valin, P. (2013). Context-based information quality for sequential decision making. In: *Proceedings of the 2013 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. San Diego, CA.

[39] Capet, P., and Delavallade, T. (2014). *Information Evaluation*. Hoboken, NJ: Wiley-ISTE.

[40] Rogova, G., and Scott, P. (2016). *Fusion Methodologies in Crisis Management Higher Level Fusion and Decision Making*. Cham, Switzerland: Springer International Publishing.

[41] Karvetski, C.W., Olson, K.C., Mandel, D.R., and Twardy, C.R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis,* 10 (4):305-326.

[42] Irwin, D., and Mandel, D.R. (2018). *Methods for Communicating Estimative Probability in Intelligence to Decision-Makers: An Annotated Collection*. DRDC Scientific Letter DRDC-RDDC-2018-L017. Toronto, ON: DRDC.

[43] Heuer, R.J., Jr. (1999). The *Psychology of Intelligence Analysis*. Washington DC: Central Intelligence Agency, Center for the Study of Intelligence.

[44] Tecuci, G., Schum, D.A., Marcu, D., and Boicu, M. (2016). *Intelligence Analysis as Discovery of Evidence, Hypotheses, and Arguments: Connecting the Dots*. New York, NY: Cambridge University Press.

[45] Mandel, D.R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology*, 6:387.

[46] Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General* ,130 (3):380-400.

[47] Chang, W., Chen, E., Mellers, B., and Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making,* 11(5):509-526.

[48] Tversky, A., and Koehler, D.J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review,* 101(4):547-567.

[49] Villejoubert, G., and Mandel, D.R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition*,30(2):171-178.

[50] Mandel, D.R. (2005). Are risk assessments of a terrorist attack coherent? *Journal of Experimental Psychology: Applied*, 11(4):277-288.

[51] Mandel, D.R. (2008). Violations of coherence in subjective probability: A representational and assessment processes account. *Cognition*, 106(1):130-156.

[52] Pool, R. (2010). *Field Evaluation in the Intelligence and Counterintelligence Context: Workshop Summary*. Washington DC: The National Academies Press.

[53] Dhami, M.K., Mandel, D.R., Mellers, B.A., and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science,* 10(6):753-757.

[54] Mandel, D.R. (2019). Can decision science improve intelligence analysis? In: *Researching National Security Intelligence: National Security Intelligence: Multidisciplinary Approaches*. Coulthart, S., Landon-Murray, M., and Van Puyvelde, D. (Eds.), 117-140. Washington, DC: Georgetown University Press.

DOCUMENT CONTROL DATA

*Security markings for the title, authors, abstract and keywords must be entered when the document is sensitive

1. ORIGINATOR (Name and address of the organization preparing the document. A DRDC Centre sponsoring a contractor's report, or tasking agency, is entered in Section 8.) North Atlantic Treaty Organization BP 25 F-92201 Neuilly-sur-Seine Cedex France	2a. SECURITY MARKING (Overall security marking of the document including special supplemental markings if applicable.) CAN UNCLASSIFIED 2b. CONTROLLED GOODS NON-CONTROLLED GOODS DMC A

3. TITLE (The document title and sub-title as indicated on the title page.)

Standards for Evaluating Source Reliability and Information Credibility in Intelligence Production

4. AUTHORS (Last name, followed by initials – ranks, titles, etc., not to be used)

Irwin, D.; Mandel, D.

5. DATE OF PUBLICATION (Month and year of publication of document.) May 2020	6a. NO. OF PAGES (Total pages, including Annexes, excluding DCD, covering and verso pages.) 16	6b. NO. OF REFS (Total references cited.) 54

7. DOCUMENT CATEGORY (e.g., Scientific Report, Contract Report, Scientific Letter.)

External Literature (N)

8. SPONSORING CENTRE (The name and address of the department project office or laboratory sponsoring the research and development.)

DRDC – Toronto Research Centre
Defence Research and Development Canada
1133 Sheppard Avenue West
Toronto, Ontario M3K 2C9
Canada

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.) 05da - Joint Intelligence Collection and Analysis Capability (JICAC)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)
10a. DRDC PUBLICATION NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.) DRDC-RDDC-2020-N237	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.) CSSP-2016-TI-2224

11a. FUTURE DISTRIBUTION WITHIN CANADA (Approval for further dissemination of the document. Security classification must also be considered.)

Public release

11b. FUTURE DISTRIBUTION OUTSIDE CANADA (Approval for further dissemination of the document. Security classification must also be considered.)

NATO

12. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Use semi-colon as a delimiter.)

Intelligence Analysis; Intelligence; Source Reliability; Information Credibility

13. ABSTRACT/RÉSUMÉ (When available in the document, the French version of the abstract must be included here.)