Received: 17 May 2022 | Revised: 12 October 2022 | Accepted: 17 October 2022

DOI: 10.1002/bdm.2307

RESEARCH ARTICLE

WILEY

Meta-informational cue inconsistency and judgment of information accuracy: Spotlight on intelligence analysis

David R. Mandel[1] | Daniel Irwin[2] | Mandeep K. Dhami[3] | David V. Budescu[4]

[1]Intelligence, Influence, and Collaboration Section, Defence Research and Development Canada, Toronto, Ontario, Canada

[2]Government of Canada, Ottawa, Ontario, Canada

[3]Department of Psychology, Middlesex University, London, UK

[4]Department of Psychology, Fordham University, New York City, New York, USA

Correspondence
David R. Mandel, Intelligence, Influence, and Collaboration Section, Defence Research and Development Canada, 1133 Sheppard Ave W., Toronto. ON, M3K 2C9, Canada.
Email: drmandel66@gmail.com

Funding information
Canadian Safety and Security Program, Grant/Award Numbers: CSSP-2016-TI-2224, CSSP-2018-TI-2394

Abstract

Meta-information is information about information that can be used as cues to guide judgments and decisions. Three types of meta-information that are routinely used in intelligence analysis are source reliability, information credibility, and classification level. The first two cues are intended to speak to information quality (in particular, the probability that the information is accurate), and classification level is intended to describe the information's security sensitivity. Two experiments involving professional intelligence analysts ($N = 25$ and 27, respectively) manipulated meta-information in a 6 (source reliability) $\times$ 6 (information credibility) $\times$ 2 (classification) repeated-measures design. Ten additional items were retested to measure intra-individual reliability. Analysts judged the probability of information accuracy based on its meta-informational profile. In both experiments, the judged probability of information accuracy was sensitive to ordinal position on the scales and the directionality of linguistic terms used to anchor the levels of the two scales. Directionality led analysts to group the first three levels of each scale in a positive group and the fourth and fifth levels in a negative group, with the neutral term "cannot be judged" falling between these groups. Critically, as reliability and credibility cue inconsistency increased, there was a corresponding decrease in intra-analyst reliability, interanalyst agreement, and effective cue utilization. Neither experiment found a significant effect of classification on probability judgments.

KEYWORDS
inconsistency, information accuracy, information credibility, intelligence analysis, meta-information, secrecy, source reliability

1 | INTRODUCTION

Uncertainty is ubiquitous and highly consequential in several walks of life, and national security decision-making is no exception (Friedman, 2019; Mandel & Irwin, 2021b). Such decision-making critically relies on intelligence analysis to assess and communicate information and uncertainties accurately and effectively (Dhami & Mandel, 2021; Fingar, 2011; Friedman & Zeckhauser, 2012; Kent, 1964; Mandel & Barnes, 2018). In supporting decision-making, intelligence analysts must routinely evaluate the accuracy of information in the form of raw intelligence that they receive. They must also be aware of the security sensitivity of such information and guard it accordingly. To assist analysts with information evaluation, they are often provided with meta-information—namely, cues that provide

TABLE 1 The admiralty code as presented in NATO intelligence doctrine (North Atlantic Treaty Organization, 2016)

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

information about the quality and security sensitivity of the information they are evaluating.

Three meta-informational cues routinely used in intelligence production are source reliability, information credibility, and classification level. The first two are determinants of information quality, especially the probability that the information is accurate. These cues are typically given as ratings using the Admiralty Code, a standard developed by the British Admiralty's Naval Intelligence Division during World War II (McLachlan, 1968) and promulgated by the North Atlantic Treaty Organization (2016) along with other defense, intelligence, and law enforcement organizations (see overview in Irwin & Mandel, 2019). Indeed, the Admiralty Code is described in NATO joint intelligence doctrine, and the evaluation of source reliability and information credibility formally constitutes the information evaluation step in the broader intelligence analysis stage of the intelligence cycle (North Atlantic Treaty Organization, 2016).

Under the Admiralty Code, users are instructed to assess source reliability and information credibility independently and to rate them using two separate scales (Table 1). While specific evaluation criteria and procedures vary between organizations (Irwin & Mandel, 2019), information credibility usually incorporates the extent to which a piece of information is consistent with other available information, whereas source reliability is often linked to confidence in a given source based on the estimated accuracy of past reporting (McDowell, 2009; Samet, 1975).[1] Once information has been evaluated, the resultant ratings are jointly communicated using an alphanumeric code. For instance, intelligence based on "probably true" information from a source deemed "usually reliable" would be marked B2. Ratings "F" and "6" are applied when it is not possible to assess source reliability and information credibility, respectively. Therefore, the ordinal scales comprising ratings A–E and 1–5 exclude ratings "F" and "6."

Past research shows that as source reliability and information credibility ratings on the Admiralty scales increase in quality (i.e., go from E to A for source reliability and 5 to 1 for information credibility), so too do intelligence practitioners' subjective probability estimates that the information is accurate (Miron et al., 1978; Samet, 1975).

[1]The use of two measures reflects the view that information credibility and source reliability may be individually insufficient as cues to accuracy. For instance, United States Marine Corps (2018, p. 3-13) intelligence doctrine instructs analysts that "A completely reliable agency may report information obtained from a completely reliable source which, on the basis of other information, is judged to be improbable and rated as A-5." Conversely, "A source known to be unreliable may provide raw information that is confirmed by reliable sources, accepted as credible information, and rated as E-1."

TABLE 2 UK security classification levels

OFFICIAL[a]	SECRET	TOP SECRET
The majority of information that is created or processed by the public sector. This includes routine business operations and services, some of which could have damaging consequences if lost, stolen, or published in the media but are not subject to a heightened threat profile.	Very sensitive information that justifies heightened protective measures to defend against determined and highly capable threat actors. For example, where compromise could seriously damage military capabilities, international relations, or the investigation of serious organized crime.	HMG's most sensitive information requiring the highest levels of protection from the most serious threats. For example, where compromise could cause widespread loss of life or else threaten the security or economic well-being of the country or friendly nations.

[a]UK official is equivalent to "confidential" in Canada and the United States.

However, due to the semantic vagueness of these scales, such ratings elicit a wide range of interpretations (Samet, 1975; Baker & Mace, 1973 as cited in Samet, 1975). Practitioners often restrict the assignment of Admiralty scales to those alphanumeric pairs exhibiting a high degree of consistency, despite being instructed to consider their assignment independently (Baker et al., 1968; Miron et al., 1978; Samet, 1975). For example, reviewing source reliability and information credibility ratings assigned during a US Army field exercise, Baker et al. (1968) found that 87% of military operators' ratings fell along the diagonal A1, B2, C3, and so on, with B2 comprising 74% of all ratings. Intelligence officers also appear to weigh information credibility more heavily than source reliability when judging information accuracy (Samet, 1975). In a similar vein, Miron et al. (1978) found that trained analysts often treat source reliability as a cue to information credibility but not vice versa.

Research also indicates that source reliability and information credibility cues influence decision-making. For instance, military decision-makers who receive highly rated information are less inclined to seek additional information before making a decision (Levine &

Samet, 1973). Halpin et al. (1978) suggest that practitioners routinely draw on these ratings when collating information from multiple sources and when tasking collection assets. These findings cohere with literature showing that when confronted with conflicting information, meta-information such as source reliability significantly influences decision-making (Carroll & Sanchez, 2021).

Whereas source reliability and information credibility cues speak to the quality of information, classification level is intended to indicate the security sensitivity of information (i.e., potential damage to national security resulting from its loss or unauthorized disclosure) and appropriate security controls (Quist, 1993). Within the UK government, for instance, information is classified using the levels in Table 2 (United Kingdom Cabinet Office, 2018). Such classifications are widely known by allies and adversaries alike.

While the classification level partially reflects the sources and methods used to obtain information, it is at best an indirect cue to information accuracy. Nevertheless, past studies show that experts and nonexperts alike exhibit a *secrecy heuristic* in which they assign more weight to classified information than to unclassified information. Travers et al. (2014) found that nonexperts judged intelligence assessments labeled "secret" to be of higher quality (defined as a composite of accuracy, impact, and soundness of reasoning) than identical assessments labeled "public." Nonexperts weighed secret information more heavily than identical public information when making hypothetical foreign policy recommendations. Moreover, they responded more favorably to foreign policy decisions based on secret information than identical public information.

Pedersen and Jansen (2019) partially replicated the secrecy heuristic in a sample of intelligence practitioners. They found that participants judged intelligence assessments labeled "secret" to be of higher quality than identical assessments labeled "open-source" but only when the assessment constituted a complex problem characterized by a high degree of uncertainty. Moreover, participants exhibited greater confidence in their own assessments when they processed intelligence labeled secret versus identical intelligence labeled open-source. Beyond the intelligence domain, Sample et al. (2020) suggest that fake-news stories routinely exploit the secrecy heuristic to appear more credible by purporting to reveal leaked classified information.

2 | THE PRESENT RESEARCH

We report two experiments using the UK and Canadian samples of professional intelligence analysts who are routinely required to assess the accuracy of information they receive and who, as NATO members, are expected to be familiar with the Admiralty Code and classification levels. In both experiments, we systematically manipulated meta-informational cues by crossing all levels of source reliability, information credibility, and two levels of classification (TOP SECRET vs. OFFICIAL; see Table 2). These experiments build on prior research investigating the influence of meta-information used in intelligence analysis in several respects. First, whereas previous studies examined the effects of source reliability (Levine & Samet, 1973), source reliability, and information credibility (Samet, 1975) or classification level (Pedersen & Jansen, 2019; Travers et al., 2014), we examined the effects of all three cues on individuals' judgments about information accuracy. This is an important consideration given that intelligence is often encoded with all three types of meta-information.[2] Second, past research used meta-information to characterize hypothetical (Karvetski & Mandel, 2020; Levine & Samet, 1973; Pedersen & Jansen, 2019; Samet, 1975) or historical intelligence assessments (Travers et al., 2014). While these studies independently manipulated informational and meta-informational content to improve internal validity, we took the additional precaution of presenting "information" that was stripped of all content except for the meta-information itself—a sacrifice of mundane realism for tighter experimental control. Finally, in contrast to previous studies that have focused only on the ordinal portions of the source reliability and information credibility scales, we examined the full 6×6 matrix of the Admiralty Code. Karvetski and Mandel (2020) found that, on average, nonexperts judged information with an F3 profile (i.e., reliability cannot be judged, possibly true) to have a .44 probability of information accuracy, whereas a C3 profile (i.e., fairly reliable and possibly true) was assigned a probability of .59. This suggests that "cannot be judged" codes tend to be interpreted nonpositively. However, to the best of our knowledge, no study has assessed precisely where such codes fall in relation to the ordinal parts of the two scales. In the present research, we examined how the two "cannot be judged" cue values affect the judged probability of information accuracy.

Given previous findings (e.g., Samet, 1975), we hypothesized that judged probability of information accuracy would be positively related to higher levels of source reliability and information credibility. That is, we predicted a linear effect of each cue on judged accuracy (Hypothesis 1). However, we also predicted that the precise shape of the function would entail a discontinuity in the linear effect across levels of the scales: As Table 1 shows, the linguistic labels for the first three ordinal positions (i.e., "completely reliable," "usually reliable," "fairly reliable," and "completely credible," "probably true," "possibly true") on both Admiralty scales are directionally positive, whereas they are directionally negative for the fourth and fifth positions (i.e., "not usually reliable," "unreliable," "doubtful," and "improbable"). In other words, levels A–C for source reliability convey that the source is to varying degrees *reliable*, whereas levels D and E convey that the source is to varying degrees *unreliable*. The same applies to the credibility scale. Directionality is a pragmatic characteristic of linguistic quantifiers (Budescu et al., 2003; Teigen & Brun, 1995, 1999) that has implications for inferred recommendations and decisions (Collins et al., 2022) and for perceived trust and confidence in advisors (Collins & Mandel, 2019; Jenkins et al., 2018). In the present context, we hypothesized that directionality would impose a grouping of scale

[2]Intelligence may be encoded with additional meta-informational markers, which are beyond the scope of this study. Examples include markers indicating the originating agency, customer(s), the source of the information being reported, dates when the report was produced or last updated, and report precedence or level of importance. Such markers usually serve a more bureaucratic function, vary from organization to organization, and are not of theoretical or practical interest, at least insofar as evaluations of information accuracy are concerned.

levels such that levels A–C and levels 1–3 belong to a positive group of ratings and levels D and E and levels 4 and 5 belong to a negative group of ratings. Therefore, we predicted that the largest gap in judged accuracy between adjacent levels of both scales will lie between the third and fourth levels (Hypothesis 2). Moreover, we expected the directionally neutral "cannot be judged" codes (i.e., F and 6) to elicit judged accuracy levels that would fall between the two directional sets, namely, between the third and fourth ordinal positions (Hypothesis 3). In fact, these codes are not intended to convey neutrality. Rather, they are meant to convey in the case of source reliability that there is no prior information on the source on which to base an assessment, and in the case of information credibility, a score of 6 reflects "any freshly reported item of information which provides no basis for comparison with any known behaviour pattern of a target" (North Atlantic Treaty Organization, 2003, p. A-2).

We further hypothesized that participants would weigh information credibility more heavily than source reliability when judging accuracy (Hypothesis 4). Compared with source reliability, information credibility is a proximal cue of information accuracy because it directly characterizes the information under consideration, whereas source reliability refers to information indirectly or reflects the veracity of past information from a source (Lemercier, 2014). Consequently, "completely credible" information from an "unreliable" source can conceivably be assigned a high probability of accuracy, whereas "improbable" information from a "reliable" source cannot.

Given prior evidence that source reliability and information credibility ratings tend to be assigned to highly consistent cue values that fall on or close to the diagonal of the 6×6 matrix (Baker et al., 1968; Miron et al., 1978; Samet, 1975), we also examined how the (in)consistency between the source reliability and information credibility ratings affected the reliability of analysts' judgments and the agreement among analysts. Samet (1975) found that the numeric probabilities assigned to the likelihood of an event reported with the maximally consistent rating A1 ranged from .88 to 1.00, whereas the probabilities assigned to an event reported with the minimally consistent rating E1 ranged much more widely from .35 to 1.00. Samet (1975) further observed that participants exhibited poor reliability when responding to the same ratings across separate tasks and that responses between participants diverged more as source reliability and information credibility decreased. However, the separate tasks that Samet (1975) used differed from one another in multiple attributes. Seeking to expand on these findings, we examined the intra-analyst reliability of judgments made for pairings of source reliability and information credibility that ranged in cue consistency from low (e.g., A5) to high (e.g., A1) within a single invariant task. Likewise, we examined interindividual agreement as a function of scale consistency. We hypothesized that (intra-analyst) reliability and (inter-analyst) agreement would be directly proportional to the consistency of levels on the source reliability and information credibility scales (Hypotheses 5 and 6, respectively). These predictions follow not only from earlier research on the Admiralty scales but also from research on the effect of cue inconsistency on judgment (Anderson & Jacobson, 1965; Lynch & Ofir, 1989; Slovic, 1966; Wyer, 1970). For instance, Slovic (1966) observed that if two diagnostic cues agreed in their cue values, participants tended to use both, but if the cues had inconsistent values, participants tended to use only one. In the present research, we examined this phenomenon by isolating consistent and inconsistent subsets of trials and then determining whether each participant used neither, one, or both of the Admiralty scales. We hypothesized that a significantly greater proportion of participants will use both cues in the consistent set than in the inconsistent set (Hypothesis 7). We also predicted that there would be a greater proportion of participants for whom neither cue was significant in the inconsistent set than

TABLE 3 Summary of hypotheses and hypothesis-testing results in Experiments 1 and 2

Hypothesis	Experiment 1	Experiment 2
1. The judged probability of information accuracy will be positively related to higher levels of source reliability and information credibility.	Supported	Supported
2. The largest gap in judged accuracy will occur between the third and fourth ordinal positions of both scales (D and E and 4 and 5, respectively).	Supported	Supported
3. Directionally neutral codes (F and 6) will elicit judged accuracy levels that, on average, fall between the third and fourth ordinal positions.	Supported	Supported
4. Participants will weigh information credibility more heavily than source reliability when judging accuracy.	Not supported	Not supported
5. Intra-individual reliability will be directly proportional to the consistency of levels on the source reliability and information credibility scales.	Supported	Supported
6. Interindividual reliability will be directly proportional to the consistency of levels on the source reliability and information credibility scales.	Supported	Supported
7. When judging accuracy, a significantly greater proportion of participants will use both source reliability and information credibility cues in the consistent set than in the inconsistent set.	Supported	Supported
8. When judging accuracy, a significantly greater proportion of participants will use neither source reliability nor information credibility cues in the inconsistent set than in the consistent set.	Supported	Supported
9. Participants will exhibit a secrecy heuristic by assigning a higher average probability of accuracy to information described as of higher classification.	Not supported	Not supported

in the consistent set (Hypothesis 8)—a result that would indicate that participants not only do not integrate inconsistent cue information but that their cue prioritizations are unreliable.

Finally, we tested the hypothesis (Hypothesis 9) that analysts would use a secrecy heuristic and assign a higher average probability of information accuracy to information labeled with a higher classification (i.e., TOP SECRET rather than merely OFFICIAL), even though classification is not intended to be a cue to information accuracy. Our interest in testing Hypothesis 9 is to verify whether earlier tests of this hypothesis by Travers et al. (2014) and Pedersen and Jansen (2019) are conceptually replicable. Table 3 summarizes all the hypotheses tested.

3 | EXPERIMENT 1

3.1 | Materials and methods

3.1.1 | Participants

We recruited 25 UK intelligence analysts who were attending regular training at a UK training facility. The sample size was based on the capacity of the training course during the period that data was collected, and no a priori power analysis was conducted.[3] Participation was voluntary and not remunerated. Most participants (72%) were male. The sample was aged 23 to 53 ($M = 35.36$, $SD = 9.91$). The sample was 50% civilian and 50% military (three participants did not respond), and mean experience in the operational community was 7.99 years ($SD = 9.10$). Roughly half of the sample (48%) indicated having been formally trained to use the Admiralty Code.

3.1.2 | Procedure and materials[4]

Middlesex University's Department of Psychology Ethics Committee approved Experiment 1, which was administered in a paper-pencil format. After providing informed consent, participants were shown the current NATO standards for evaluating source reliability and information credibility (Table 1) and definitions for the classification levels "OFFICIAL" and "TOP SECRET." Then, participants were asked to judge the accuracy of 82 pieces of information. The order of these items was randomized per participant, with 11–12 items presented on a given page. Seventy-two of the items represented all possible combinations of these three variables, namely, 6 (Source Reliability) × 6 (Information Credibility) × 2 (Classification). Ten additional items (all rated TOP SECRET) were resampled, providing a basis for measuring intra-individual reliability. For the retest items, cue consistency between the source reliability and information credibility scales was low (A5, E1), medium (A3, C1, C5, E3), or high (A1, C3, E5, F6).

For each item, participants were asked, "What is the probability that this piece of information is accurate rather than inaccurate?" Responses were provided on a 0% (*certainly inaccurate*) to 100% (*certainly accurate*) scale with 50% labeled as *completely uncertain*. The scale values were in increments of 5% with four dashes representing the intermediate values. The information item was nondescriptive and simply labeled by an alphabetic code (e.g., "Information J: OFFICIAL; Source reliability: Unreliable; Information credibility: Doubtful"). Participants could refer back to the Admiralty scales at any point during the task. Following the experimental task, participants answered demographic questions about their age, sex, and professional experience and indicated whether they had been trained to use the Admiralty scales. Participants were subsequently debriefed on the aims of the experiment.

3.2 | Results and discussion

3.2.1 | Preliminary analyses

There were 19 missing accuracy judgments out of 2050 (i.e., across the 82 trials × 25 participants). Missing responses were replaced using mean substitution (i.e., the mean of other responses for the same trial was used as a proxy value). Most participants circled a single marked value (i.e., multiple of 5) and in a few instances, participants circled the space between two marked values. In these cases, we used the midpoint. For instance, if a participant circled the region between 5% and 10%, they were assigned a value of 7.5.

3.2.2 | Effect of meta-informational cues on judgments of information accuracy

We first examined the effect of source reliability, information credibility, and classification on judged probability of information accuracy in a 6 (Source Reliability) × 6 (Information Credibility) × 2 (Classification) × 2 (Training) mixed analysis of variance (ANOVA). All factors except training were repeated measures. Training referred to whether or not analysts reported having been formally trained in the Admiralty Code. Reliability check items were excluded from this analysis. There were significant main effects of source reliability ($F[5, 19] = 21.65$, $p < .001$, $\eta_p^2 = .851$) and information credibility ($F[5, 19] = 45.22$, $p < .001$, $\eta_p^2 = .922$), whereas the main effects of classification, training, and assessable interaction effects were not significant (i.e., there were insufficient degrees of freedom to assess the reliability × credibility interaction or higher-order interactions involving both of these factors). Thus, we did not find support for the secrecy heuristic hypothesis (Hypothesis 9). The mean judgment of accuracy in the TOP SECRET condition was 49.07 ($SE = 1.86$), and in the OFFICIAL condition, it was 48.59 ($SE = 1.82$).

[3]To compensate for the expected small sample, we used a repeated-measures design with a large number of trials (i.e., 82).

[4]Materials and data for Experiments 1 and 2 are available online (https://osf.io/zrqc8/.). The experiments were not preregistered.

Providing support for Hypothesis 1, Figure 1 shows that mean probability of information accuracy is well approximated by the linear effects of the source reliability and information credibility cues excluding the "cannot be judged" levels (i.e., F on the source reliability scale and 6 on the information credibility scale) from the analysis: for source reliability, $F_{linear}(1,24) = 108.67$, $p < .001$, $\eta_p^2 = .819$; for information credibility, $F_{linear}(1,24) = 192.54$, $p < .001$, $\eta_p^2 = .889$. Supporting Hypothesis 2 (i.e., the directionality of verbal labels on the scales imposes a grouping effect), there was a significantly steeper drop in mean probability from the third to fourth levels of the source reliability scale (mean difference = 16.55, 95% confidence interval (CI) [12.76, 20.35]) than from the next largest difference between adjacent scale values on the ordinal part of the scale (i.e., the mean difference between values 1 and 2 = 9.11, 95% CI = 6.86, 11.36). Likewise, for the information credibility scale, the mean difference in probability between adjacent scale values on the ordinal part of the scale was greatest for the values 3 and 4 (mean = 21.50, 95% CI [17.78, 25.22]), which was significantly greater than the next largest difference for values 1 and 2 (mean = 9.82, 95% CI [7.58, 12.06]). For both scales, as can be verified in Figure 1 and supporting Hypothesis 3, the "cannot be judged" options have mean probabilities that lie between those observed for third and fourth levels on the relevant scale.

To provide an additional test of Hypothesis 3, we calculated the average source reliability rating for each level of that scale, and we did the same thing for the information credibility scale. Then we computed the frequency of participants whose mean probability judgment for the "cannot be assessed" code fell between the means of the individual scale values {1–3} and {4–5} for each of the two scales. We computed 1000 bias-corrected and accelerated bootstrap samples to generate 95% confidence intervals on the percentage of cases that conformed to Hypothesis 3. For the source reliability scale, 80% [68%, 92%] of participants placed "cannot be judged" mean probabilities between the mean probabilities of the directionally positive and the directionally negative sets of scale levels. For the information credibility scale, 84% [72%, 96%] of participants placed "cannot be judged" mean probabilities between the mean probabilities of the directionally positive and the directionally negative sets of scale levels. These percentages far exceed the 1/6 probability (16.7% relative frequency) expected by chance by a Binomial test.

To test Hypothesis 4 (i.e., differential cue weight), we examined the comparative strength of the linear relation between the judged probability of information accuracy and the two scales to test the hypothesis that information credibility would be more strongly correlated with probability judgments than would source reliability. We eliminated retest items and all "cannot be judged" items. Then, for each of the 25 participants, we correlated their probability judgments with the source reliability and information credibility cue values that characterized the remaining 50 items. The resulting correlations were Fisher transformed to *z*-scores and finally subjected to a paired *t*-test. This revealed no statistically significant difference between the two cues, $t(24) = 1.60$, one-sided $p = .062$, and Cohen's $d = .32$. The mean Fisher correlation was −.66 ($SD = .36$) for source reliability and −.88 ($SD = .36$) for information credibility. The magnitude of the correlation with information credibility was greater than that with source reliability for 12/25 judges, and it was less than the magnitude of the correlation with source reliability for 12/25 (the correlations were equal for the remaining participant). Therefore, Hypothesis 4 was not supported.

3.2.3 | Effect of cue consistency on reliability of analysts' judgments

To test Hypothesis 5 (i.e., intra-analyst reliability of judgments would be directly proportional to the consistency of source reliability and information credibility cues), for each of the 10 test–retest pairs, we calculated the standardized absolute error (SAE) as follows:

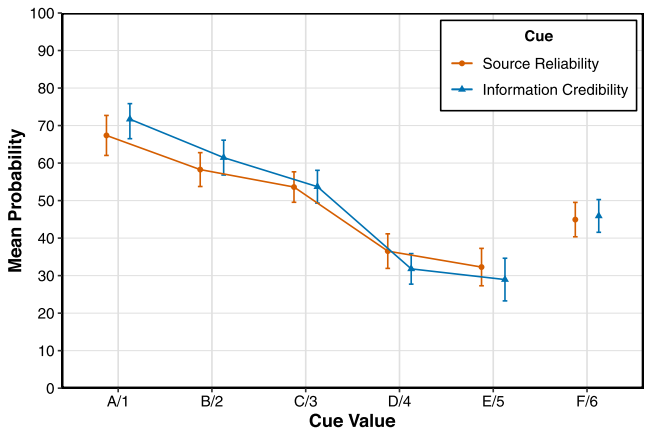

FIGURE 1 Mean probability of information accuracy (with 95% confidence intervals) by source reliability and information credibility cues in Experiment 1. Cue value labels are provided in Table 1.

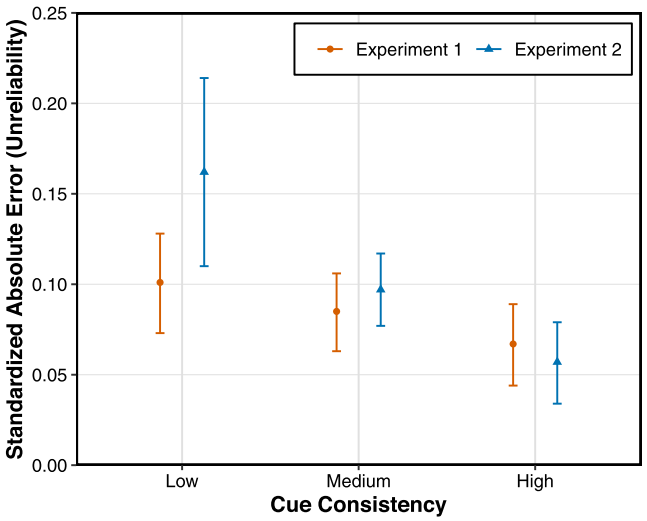

FIGURE 2 Standardized absolute error (i.e., unreliability; with 95% confidence intervals) by cue consistency in Experiments 1 and 2

$$SAE = |p_1 - p_2| / Max(p_1, p_2, 1 - p_1, 1 - p_2),$$

where p_1 and p_2 refer to the probabilities elicited on test and retest, respectively. The term in the denominator corrects for variation in the extremity of probability judgments (since more extreme probabilities can also have the potential for greater error), and it is consistent with similar usage in other studies (Fan et al., 2019; Mandel, Dhami, et al., 2021). SAE was averaged for low, medium, and high cue consistency sets, as defined earlier. To test Hypothesis 5, we conducted a one-way (consistency: low, medium, high), repeated-measures ANOVA on SAE. The main effect of consistency was significant, $F(2, 23) = 3.49$, one-sided $p = .024$, $\eta_p^2 = .233$. As hypothesized, the planned linear contrast was significant, $F(1, 24) = 6.66$, $p = .016$, $\eta_p^2 = .217$. Supporting Hypothesis 5, as cue consistency increased, unreliability (i.e., SAE) decreased (see Figure 2 in red).

3.2.4 | Effect of cue consistency on agreement among analysts

Next, we tested Hypothesis 6 (i.e., interanalyst agreement in probability judgments of information accuracy is directly proportional to the consistency between the source reliability and information credibility cues). For each of the 50 trials excluding retest and "cannot be judged" cases, we calculated the standard deviation as a measure of agreement. Since mean substitutions for missing responses reduce variability, we computed the standard deviations on data with missing cases excluded per trial. These standard deviations were then subjected to a one-way (consistency) ANOVA where the independent variable was the ordinal degree of scale consistency (i.e., ranging over a 5-point scale; e.g., A1 or E5 would be perfectly consistent and have a score of 0 and A5 or E1 would be maximally inconsistent and have a score of 4). Figure 3 (in red) shows that the mean standard deviations decreased with increasing cue consistency. The main effect of consistency was significant, $F(4, 45) = 19.62$, $p < .001$, $\eta_p^2 = .636$. Polynomial contrasts further revealed that only the linear effect was significant, K-matrix contrast estimate $= 7.51$, $p < .001$. Therefore, Hypothesis 6 was supported.

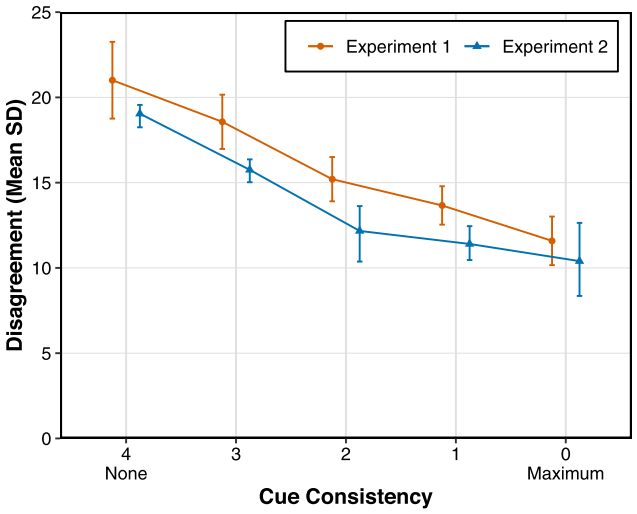

FIGURE 3 Disagreement (mean standard deviation) by cue consistency in Experiments 1 and 2. Error bars are 95% confidence intervals from 1000 bias-corrected and accelerated bootstrap samples.

3.2.5 | Effect of cue consistency on cue use

Finally, to test Hypothesis 7 (i.e., a greater proportion of analysts would use both cues when the Admiralty scales were consistent than when they were inconsistent) and Hypothesis 8 (i.e., a greater proportion of analysts would use neither cue when the scales were inconsistent than when they were consistent), we regressed each participant's probability judgments on the source reliability and information credibility cue values. Given that the factorial experimental design ensures a zero correlation between the two cues over the 50 trials analyzed earlier, we restricted this analysis to two sets of 16 cases in which information was coded as either A1, A2, B1, B2, D4, D5, E4, and E5 (the consistent set) or A4, A5, B4, B5, D1, D2, E1, and E2 (the inconsistent set). The correlation between the two cues in the consistent set is .90 and it is −.90 in the inconsistent set. Table 4 shows the percentage of analysts with zero, one, or two of the cues as significant predictors. Supporting Hypothesis 7, the percentage of analysts with two significant predictors was significantly greater in the consistent set (76%) than in the inconsistent set (20%), $p = .001$ by two-sided McNemar test. Furthermore, supporting Hypothesis 8, the percentage of analysts with neither of the cues as significant predictors was significantly greater in the inconsistent set (44%) than in the consistent set (4%), $p = .006$ by two-sided McNemar test.

In summary, Experiment 1 found support for Hypotheses 1, 2, 3, 5, 6, 7, and 8 and did not support Hypotheses 4 and 9. Despite the expert status of the participants, given the small sample size, the results observed in this experiment should be replicated in a new sample of intelligence analysts, which is the aim of the next experiment.

4 | EXPERIMENT 2

Experiment 2 sought to replicate and extend the findings of Experiment 1 in a second sample of intelligence analysts and using an online

TABLE 4 Percentage of analysts by significant predictors and set in Experiment 1

	Set	
Significant predictors	**Consistent**	**Inconsistent**
Both cues	76.0	20.0
Source reliability only	8.0	8.0
Information credibility only	12.0	28.0
Neither cue	4.0	44.0

Note: $N = 25$.

survey tool. Experiment 1 lacked experimental control over the degree to which analysts reviewed the Admiralty Code during the judgment task. In Experiment 2, in contrast, participants were shown the source reliability, information credibility, and classification scales either throughout the core experimental task (i.e., on each screen presenting a unique item) or only at the beginning of the task. This permitted an examination of whether scale access on a case-by-case basis improved reliability. In addition, participants in Experiment 2 were prevented from viewing their previous responses, which could not be controlled in Experiment 1 and could have influenced intra-analyst reliability. Also, in Experiment 1, we did not test whether analysts could correctly rank order the levels of the scales after completing the task. In Experiment 2, we added objective measures of source reliability and information credibility scales and classification knowledge, allowing us to screen out participants with an inadequate understanding of the Admiralty Code and/or the levels of classification. Finally, in Experiment 2, we used a Canadian sample of intelligence analysts, which permitted a test of the generalizability of our earlier findings to a distinct national sample that is nevertheless part of NATO and other intelligence-sharing multilateral organizations (e.g., the "Five-Eyes" intelligence community).

4.1 | Materials and methods

4.1.1 | Participants

Thirty-seven Canadian intelligence analysts participated remotely using a Qualtrics survey link distributed by their managers, who did not have access to the study data and who did not discuss the research hypotheses with participants prior to the study. Participation was voluntary and was not remunerated. As in Experiment 1, sample size was based solely on the availability of analysts. After excluding participants who could not correctly rank order the ordinal levels of source reliability, information credibility, and classification after completing the primary task, the final sample was reduced to 27 participants who were 67% male and aged 25 to 72 years ($M = 40.52$, $SD = 11.07$). The final sample was 89% civilian and 11% military, and mean experience in the operational community was 10.67 years ($SD = 7.89$).

4.1.2 | Design

Defence Research and Development Canada Human Research Ethics Committee approved Experiment 2. Participants were randomly assigned to two experimental conditions based on Scales Presentation modality. Participants were shown the source reliability, information credibility, and classification scales either only once at the beginning (at-start-only condition; $n = 17$) or both at the start and then again for each trial (at-start-and-per-trial condition; $n = 10$). Source reliability, information credibility, and classification level were manipulated as repeated measures as in Experiment 1.

4.1.3 | Procedure and materials

After providing consent, participants were told that they would be asked to judge the accuracy of 82 pieces of information coded in terms of source reliability, information credibility, and classification level. They were shown the source reliability, information credibility, and classification scales and instructed to carefully review them. Next, participants were presented with an annotated sample question demonstrating how to identify the three meta-informational markings on a given piece of information and input a response.

During the core experimental task, participants responded to the same 82 items from Experiment 1 presented in random order for each participant. Participants indicated the accuracy of each item using a percentage-chance slider ranging from 0 to 100 (default position = 0). As the participant adjusted the slider, the unit value was shown onscreen. Each item appeared on a separate screen, and participants were unable to view or modify previous responses. Participants were required to provide a response in order to proceed to the next item, and this prevented missing responses. As in Experiment 1, the meta-informational cues were not contaminated by genuine information content or variations in such content.

Following the core experimental task, participants answered demographic questions (i.e., age, sex, and professional experience) and indicated their familiarity with the rating scales. Next, they completed the second set of questions designed to test their understanding of the rating system. Specifically, they were asked to rank the levels of the source reliability (i.e., A–F) and information credibility (i.e., 1–6) scales from most to least reliable/credible and to indicate whether OFFICIAL or TOP SECRET information is more sensitive. These questions were presented one at a time, in random order. Following the experiment, participants were debriefed on the research aims.

4.2 | Results and discussion

4.2.1 | Effect of meta-informational cues on judgments of information accuracy

We first examined the effect of the three meta-informational cues on judged probability of information accuracy in a 6 (Source Reliability) × 6 (Information Credibility) × 2 (Classification Level) × 2 (Scales Presentation) mixed ANOVA. There were significant main effects of source reliability ($F[5,21] = 57.44$, $p < .001$, $\eta_p^2 = .932$) and information credibility ($F[5,21] = 61.44$, $p < .001$, $\eta_p^2 = .936$). No other effects in the model were statistically significant. Therefore, Hypothesis 9 was not supported. As Figure 4 shows, and supporting Hypothesis 1, consistent with the results of Experiment 1, mean probability is approximately linearly related to the source reliability and information credibility levels (excluding the "cannot be judged" levels from the analysis): for source reliability, $F_{linear}(1,26) = 378.78$, $p < .001$, $\eta_p^2 = .936$; for information credibility, $F_{linear}(1,26) = 382.38$, $p < .001$, $\eta_p^2 = .936$. Consistent with Experiment 1 and supporting Hypothesis 2, there was a significantly steeper drop in mean

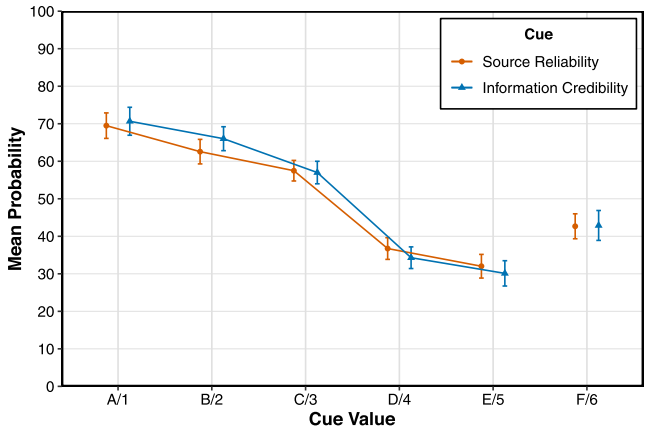

FIGURE 4 Mean probability of information accuracy (with 95% confidence intervals) by source reliability and information credibility cues in Experiment 2. Cue value labels are provided in Table 1.

probability from the third to fourth levels of the source reliability scale (mean difference = 20.44, 95% CI [17.22, 23.65]) than from the next largest difference between adjacent scale values on the ordinal part of the scale (i.e., the mean difference between values 1 and 2 = 7.28, 95% CI = 5.84, 8.71). Likewise, for the information credibility scale, the mean difference in probability between adjacent scale values on the ordinal part of the scale was greatest for the values 3 and 4 (mean = 22.03, 95% CI [18.36, 25.71]), which was significantly greater than the next largest difference for values 2 and 3 (mean = 8.74, 95% CI [6.90, 10.59]). Moreover, for both scales, as can be verified in Figure 4 and supporting Hypothesis 3, the "cannot be judged" items have mean probabilities that lie between those observed for third (weakest positive) and fourth (weakest negative) levels on the relevant scale.

As in Experiment 1, we tested Hypothesis 3 using the same bootstrap procedure used in Experiment 1. For both scales, 81.5% [70.4%, 92.6%] of participants had "cannot be judged" mean probabilities that fell between the mean probabilities of the directionally positive and the directionally negative sets of scale levels. As in Experiment 1, these percentages far exceed the 1/6 probability (16.7% relative frequency) expected by chance and provide additional support for Hypothesis 3.

We next examined the comparative strength of the linear relation between the judged probability of information accuracy and the source reliability and information credibility cues. Consistent with Experiment 1, a paired *t*-test did not reveal a significant difference between the two cues, $t(26) = 1.03$, one-sided $p = .158$, and Cohen's $d = .202$. The mean Fisher correlation was −.69 ($SD = .23$) for source reliability and −.78 ($SD = .25$) for information credibility. The correlation with information credibility was greater than the correlation with source reliability for 15/27 (56%) participants, which is not significantly different from ½ (at $\alpha = .05$ by a one-sided sign test.) Therefore, as in Experiment 1, Hypothesis 4 was not supported.

4.2.2 | Effect of cue consistency on reliability of analysts' judgments

To test Hypothesis 5, that the reliability of analysts' judgments will be directly proportional to the consistency between the source reliability and information credibility cues, we first conducted a 3 (Consistency) × 2 (Scales Presentation) mixed ANOVA on the SAE. Only the main effect of consistency was significant, $F(2, 25) = 7.46$, one-sided $p = .002$, $\eta_p^2 = .374$. Moreover, the planned linear contrast was significant, $F(1, 26) = 14.89$, one-sided $p < .001$, $\eta_p^2 = .364$. As Figure 2 (in blue) shows, consistent with the results of Experiment 1 and supporting Hypothesis 5, reliability increased as a function of cue consistency.

4.2.3 | Effect of cue consistency on agreement among analysts

We tested Hypothesis 6 that agreement in probability judgments of information accuracy is directly proportional to the consistency between the source reliability and information credibility cues, using a one-way (Consistency) ANOVA on standard deviations. The effect of consistency was significant, $F(4, 45) = 14.34$, $p < .001$, $\eta_p^2 = .560$, and polynomial contrasts further revealed that the linear effect was significant, *K*-matrix contrast estimate = 6.84, $p < .001$. A weaker quadratic effect was also significant, *K*-matrix contrast estimate = 4.95, $p < .029$. Figure 3 (in blue) shows that, supporting Hypothesis 6 and consistent with Experiment 1, the mean standard deviations decreased with increasing cue consistency.

4.2.4 | Effect of cue consistency on cue use

We used the same procedure as in Experiment 1 to test Hypotheses 7 and 8 regarding cue use under conditions of consistent and inconsistent Admiralty scale values and obtained results compatible with Experiment 1. As Table 5 shows, supporting Hypothesis 7, the percentage of analysts with two significant predictors was significantly greater in the consistent set (74.1%) than in the inconsistent set (7.4%), $p < .001$ by two-sided McNemar test. Furthermore,

TABLE 5 Percentage of analysts by significant predictors and set in Experiment 2

	Set	
Significant predictors	**Consistent**	**Inconsistent**
Both cues	74.1	7.4
Source reliability only	11.1	11.1
Information credibility only	11.1	14.8
Neither cue	3.7	66.7

Note: $N = 27$.

supporting Hypothesis 8, the percentage of analysts with no significant predictors was significantly greater in the inconsistent set (66.7%) than in the consistent set (3.7%), $p < .001$ by two-sided McNemar test.

In summary, Experiment 2 replicated the key results of Experiment 1 using a sample of analysts from a different population (Canadian) and employing a different data collection procedure. As in Experiment 1, Hypotheses 1, 2, 3, 5, 6, 7, and 8 were supported, and Hypotheses 4 and 9 were not supported. None of the observed effects depended on whether the scales were presented only at the start of the task or on each trial.

5 | META-ANALYSIS OF THE EFFECT OF CLASSIFICATION ON INFORMATION ACCURACY

The effect of security classification on information accuracy was nonsignificant in both experiments and, therefore, did not support Hypothesis 9 (i.e., the findings did not indicate that analysts were using the secrecy heuristic). However, both experiments have small sample sizes that making it difficult to detect small statistical effects. Moreover, in both experiments, the probability of information accuracy was judged to be higher, on average, in the TOP SECRET condition than in the OFFICIAL condition, as predicted by the secrecy heuristic. Therefore, we estimated the meta-analytic effect size by calculating the difference score (i.e., mean of TOP SECRET trials minus mean of OFFICIAL trials) and obtaining Cohen's *d* in each experiment. These values were 0.19 in Experiment 1 and 0.40 in Experiment 2. Submitting these values to the ESCI Excel macro[5] for combining single-group effect sizes yielded a combined effect of $d = 0.30$ with a 95% confidence interval of [0.02, 0.57]. This represents a small to medium-sized combined effect with a wide error margin.

6 | GENERAL DISCUSSION

The evaluation and communication of meta-informational cues are important steps in intelligence production (Irwin & Mandel, 2019), and they play a vital role in efforts to monitor and tackle current and emerging threats and risks to national and international security. The present research sheds light on the most common methods for encoding raw intelligence with meta-information—the Admiralty Code, which specifies source reliability and information credibility as an alphanumeric pair along with security classification level. Despite the fact that our samples included professional intelligence analysts from two NATO countries, who on average had several years of experience, roughly half of the UK analysts sampled indicated that they had not been trained to use the Admiralty Code and just over one-quarter of Canadian analysts were excluded for incorrectly rank ordering the verbal labels of at least one of these scales. These findings suggest that intelligence organizations might require their analysts to undergo additional training to ensure adequate awareness of joint intelligence doctrine for NATO operations, in general, and methods for information evaluation, in particular.

Notwithstanding the variance in awareness of the Admiralty Code, we found strong consistency in our findings between Experiment 1 in which UK analysts who indicated unfamiliarity with the Admiralty Code were not excluded and Experiment 2 where the minority of Canadian analysts who failed a competence-based screening test were excluded from further analyses. The high degree of finding consistency is unsurprising. In some cases, such as how mean probability varies as a function of the scale levels, the consistency likely reflects the tendency for people, in general, to order the verbal terms in each scale as they are ordered in the Admiralty Code. Thus, we would expect Hypotheses 1 and 2 to be supported even if this experiment were conducted in a nonexpert sample. Similarly, as we discuss later, we view the low test–retest reliability and inter-rater agreement as a consequence of cognitive processes encountered when confronting cues that vary in their consistency.

Our findings also cohere with those of earlier studies on US military personnel (e.g., Baker et al., 1968; Samet, 1975) and reveal that the judged probability of information accuracy varied as a linear function of both source reliability and information credibility cues (see summary of findings in Table 3). Given that source reliability is an indirect cue to information accuracy, whereas information credibility is a direct cue, we expected information credibility to be a stronger correlate of judged probability of information accuracy than source reliability. However, this hypothesis was not supported in either experiment, and we did not replicate Samet's (1975) finding. One possibility is that the "information-absent" stimuli used in the experimental task caused the meta-information to be weighted equally, whereas they might be weighted unequally in contexts where they are attached to specific sources (e.g., different informants) and pieces of evidence. More generally, given the paucity of research on this issue, variations in task characteristics could be profitably explored in future research.

The findings from both experiments supported our hypotheses that were motivated by the notion of linguistic directionality effects on probability judgment. On average, analysts treated "cannot be judged" as somewhere between "fairly reliable" and "not usually reliable" for the source reliability cue and as somewhere between "possibly true" and "doubtful" for information credibility cue. In both cases, this location reflects an inflection point where the initial run of directionally positive terms switches to a shorter run of directionally negative terms. This inflection point in directionality can further explain why there is a steep drop-off in the judged probability of information accuracy. Indeed, in both studies, the mean difference in judged probability between the third and fourth levels of each scale was significantly larger than the next largest distance between adjacent levels on the ordinal scales (i.e., from A to E and from 1 to 5). Taken together, these characteristics support the idea that analysts are sensitized to the directionality of linguistic terms used to anchor the levels of the scales. In the present context, directionality appears to impose a grouping of scale levels such that A–C and 1–3 belong to a

[5]Retrieved from: https://thenewstatistics.com/itns/esci/

positive group and D and E and 4 and 5 belong to a negative group with "cannot be judged" representing a directionally neutral zone between these groups. Since negatively directional terms are often interpreted as "recommendations against" (Collins et al., 2022), labeling sources of information with negatively directional terms may prompt the receiver to infer that the sender is recommending not to use that source or that information. This potential implication ought to be addressed in future research.

Note that the mean locations of F and 6 (i.e., the "cannot be judged" codes) close to the middle of the ordinal scale might also reflect a wide dispersion of probability judgments. That is, if one has no clear meta-information on which to base one's information evaluation, then one might expect the guesstimates to be quite variable. However, it is evident from Figures 1 and 4 that this was not the case. In fact, the confidence intervals for these ratings are comparable with the confidence intervals for the other ordinal scale values. Yet, a closer examination of their distributions reveals that they are also highly asymmetric with a modal probability of .50. To examine this more precisely, we computed the average probability across the three elicitations involving an F6 pattern (i.e., two with an OFFICIAL designation and one with a TOP SECRET designation) in both experiments. The skewness of the distributions of probabilities was negative: -1.41 ($SE = 0.46$) in Experiment 1 and -1.23 ($SE = 0.45$) in Experiment 2. The fraction of participants having averaged judgments between .49 and .51 was 60.0% in Experiment 1 and 44.4% in Experiment 2. These results suggest that the location of the F/6 ratings might have had more to do with the "fifty-fifty blip" response pattern (Fischhoff & Bruine de Bruin, 1999) than with the perception of neutral directionality, as we originally hypothesized. It has been well-documented that when individuals are at an epistemic loss for how to judge the probability of an event, they tend to default to the .50 response, reflecting as best they could their maximally unsure "fifty-fifty" judgment (Bruine de Bruin et al., 2000; Fischhoff & Bruine de Bruin, 1999; Mandel & Irwin, 2021a). In practice, analysts would not be compelled to assign a specific probability of information accuracy to an F/6 code. Future research could examine analysts' responses when given the opportunity to provide a numeric probability range or to select a "probability cannot be assessed" option, which has been used in prior research to mitigate the fifty-fifty blip (Mandel, 2005).

Another important set of findings concerned the effect of cue consistency in the Admiralty scales on intra-analyst reliability, inter-analyst agreement, and cue utilization. As cue consistency declined, so did reliability, agreement, and cue use. These effects were observed in both experiments, and they were quite large. Compatible with earlier studies on cue inconsistency (e.g., Slovic, 1966), these findings indicate that analysts are unable to fuse inconsistent admiralty ratings in a manner that is consistent within and across individuals. Indeed, the fact that both meta-informational cues were not significant predictors of analysts' judgments when the cue values were inconsistent suggest not only that cue discounting occurs (Anderson & Jacobson, 1965) but also that it manifests itself unreliably (i.e., as noise). To the extent that, in practice, such cues are used in tandem, it might be helpful to provide explicit guidelines for mapping the full set of possibilities onto probability levels, much as risk matrices that cross threat likelihood and consequence severity often map each cell onto specific risk levels. However, these risk assessment/communication methods are not without limitations and a serious concern is that the mappings are arbitrary (e.g., Dowie, 1999; Kaplan, 1997; Mandel, 2007). Likewise, if the Admiralty scales were mapped onto probability of information accuracy levels (or some other dependent measure such as trustworthiness or information utility), it is unclear what normative status it would have. In all likelihood, as with many organizational scales for communicating uncertainty, it would be decided by fiat (Mandel, Wallsten, & Budescu, 2021) and may be misapplied (Ho et al., 2015).

Alternatively, the considerations that underlie the evaluation of each scale could be used in a structured approach to arrive at a numeric probability judgment (e.g., Halpin et al., 1978; Miron et al., 1978; Phelps et al., 1980; Samet, 1975). For instance, Irwin and Mandel (2019) outlined (as examples) 24 questions that intelligence analysts could consider when evaluating information to better assess the probability that it is accurate. The numeric probabilities subsequently assigned could be precise values or probability ranges that, in turn, convey the degree of confidence (or credible interval) the analyst has in the judgment.

At present, however, NATO intelligence doctrine instructs analysts to treat reliability and credibility assessments independently. As formal analyses of the Admiralty scales indicate (Icard, 2019), this may in any case be unsound advice. Our findings suggest that it is psychologically implausible to implement even if its normative status was sound. Such instruction parallels guidance to analysts to treat the assessment of event probabilities and analytic confidence independently. However, Irwin and Mandel (2022) found that manipulations of confidence levels (i.e., whether they were *low*, *medium*, or *high*) had a greater effect on analysts' inferred event probabilities (which confidence ratings should not affect) than on the width of their inferred numeric confidence intervals (which confidence rating should affect). These examples of ratings that are stipulated to be independent, but which are correlated, appear to be instances of the more general "halo effect" tendency (Thorndike, 1920) and which reflects the workings of an associative reasoning system in human cognition (Kahneman, 2011).

Finally, we found weak support for the secrecy heuristic among intelligence analysts. The effect of classification was not statistically significant in either experiment. However, there was a small meta-analytic effect size in the predicted direction. Given that the classification level of intelligence is not intended as a cue to information accuracy, our findings in conjunction with those of Travers et al. (2014) and Pedersen and Jansen (2019) suggest that analysts might benefit from training aimed at mitigating use of the secrecy heuristic. However, the culture of the US intelligence community prizes access to secret information (Johnston, 2005; Lieberthal, 2009; Treverton, 2001). Likewise, an interview study of intelligence managers in Canada found that several respondents identified overattention to secret information as an important challenge for the intelligence community (Derbentseva et al., 2010). It may prove difficult to overcome the attitude that, all else being equal, classified

information is intrinsically better than information that is unclassified unless the intelligence community undergoes a congruent cultural change. We should also caution that, in practice, it is difficult to gauge whether classification level might inadvertently provide valid cues to information accuracy. Although the intention of classification is not to convey information accuracy, it may very well be the case that classified information is more likely to be accurate than unclassified sources. Thus, caution is warranted in interpreting the results of this and other studies on the secrecy heuristic.

6.1 | Future research

Beyond the directions for future research already mentioned, future research could examine how the meta-informational cues used in intelligence analysis influence judgments about a wider set of measures. Information quality is a multidimensional construct (Lee et al., 2002), and real or perceived accuracy is just one qualitative factor by which to evaluate intelligence (Miron et al., 1978; Thompson et al., 1989). It would be informative to examine in what manner the Admiralty Code cues influence perceived informativeness, trustworthiness, or the likelihood that analysts or other end-users will use intelligence. Similarly, the response to the Admiralty Code could also be varied beyond a measure of probability. For instance, we predict that evaluation of the "cannot be judged" codes to be much less susceptible to the fifty-fifty blip response pattern if the dependent measure is not recorded on a probability scale. As Schneider et al. (2022) found, pandemic-related information whose quality was ambiguous (akin to the F/6 codes in the Admiralty Code) was rated in terms of trustworthiness as similar to information described as being of low quality. This may be due to the fact that trustworthiness ratings are unlikely to exhibit a fifty-fifty blip.

A broadening of the dependent measures collected would also allow researchers to examine the extent to which intra-analyst reliability and interanalyst agreement of judgments are contingent on cue consistency in the Admiralty Code. One prediction that could be tested is whether cue consistency has a weaker influence on reliability and agreement for measures in which people tend to weigh one meta-informational cue more heavily than the other. For example, if analysts judge informativeness mainly on the basis of information credibility, we would expect cue inconsistency to have at most a weak effect on the reliability of informativeness judgments.

Future research could also test the viability of alternative methods for conveying meta-information. If analysts were sensitized to the directionality of the linguistic terms used to anchor the levels of each scale, one could examine the effect of relabeling the lower scale levels (i.e., D and E and 4 and 5) such that all levels are directionally positive. Such a change may mitigate the discontinuity in ratings observed in our studies. Recall that in the present research, meta-information did not have to compete or interact with the raw information it describes, but in practice, meta-information would not be presented without context. We intentionally designed the present research to isolate the effects of the three meta-informational sources we studied. However, studies examining the interaction between attributes of information and of meta-information would be of value.

Certain versions of the Admiralty Code provide short descriptions for each scale item. For instance, in the now defunct NATO Standardization Agreement 2511 (North Atlantic Treaty Organization, 2003), an information credibility rating of "improbable" describes a piece of information "which positively contradicts previously reported information or conflicts with the established behaviour pattern of an intelligence target in a marked degree." Meanwhile, a source reliability rating of "not usually reliable" describes a source "which has been used in the past but has proved more often than not unreliable." Future research could examine the effect of these scale descriptions on information accuracy and other judgments. For instance, if analysts are reminded of the associated definition of each label, they may rely less on the directionality cues and focus on the ordinal properties of the scales. If so, we might expect that information credibility ratings would have more weight on judgments of information accuracy than source reliability ratings.

Although intelligence doctrines may stipulate the meaning of the levels of these scales (see Irwin & Mandel, 2019), it is unlikely that analysts or other end-users always keep them in mind. Considerable research on the stipulated numeric range interpretations for verbal probability terms used by various organizations, including intelligence agencies, shows that people lose track of the stipulated meanings provided in official lexicons and default quickly to their own interpretations of vague terminology when encountering the probability terms in assessment statements (Budescu et al., 2014; Ho et al., 2015; Mandel & Irwin, 2021a; Wintle et al., 2019). Given that the definitions of the Admiralty scale levels are much longer and involve multiple concepts (unlike a simple numeric range equivalent), we expect that when these are out of sight (i.e., buried in dense intelligence doctrine, as is usually the case), they will also be out of mind—or worse: Given their vagueness and ambiguity, they may be prone to variable interpretations by analysts. If so, much as has been argued for in the case of structured analytic techniques used by the intelligence community (Chang et al., 2018), these definitions may amplify rather than tamp down on noise in analysts' judgments.

ACKNOWLEDGMENTS

Funding support for this work was provided by the Canadian Safety and Security Program Projects CSSP-2016-TI-2224 and CSSP-2018-TI-2394 under the direction of the first author. We thank Brenda Fraser, Sarah Gibbon, William Kozey, and Mark Timms for their research assistance. This work was initiated under the North Atlantic Treaty Organization's Systems Analysis and Studies Panel Research Technical Group on Assessment and Communication of Uncertainty in Intelligence to Support Decision-Making (SAS-114), and we thank its members for their feedback on this research.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the Open Science Framework (https://osf.io/zrqc8/).

ORCID

David R. Mandel https://orcid.org/0000-0003-1036-2286
Mandeep K. Dhami https://orcid.org/0000-0001-6157-3142
David V. Budescu https://orcid.org/0000-0001-9613-0317

REFERENCES

Anderson, N. H., & Jacobson, A. (1965). Effect of stimulus inconsistency and discounting instructions in personality impression formation. *Journal of Personality and Social Psychology*, *2*(4), 531–539. https://doi.org/10.1037/h0022484

Baker, J. D., McKendry, J. M., & Mace, D. J. (1968). *Certitude judgments in an operational environment* (Technical Research Note 200). US Army Research Institute for Behavioral and Social Sciences.

Bruine de Bruin, W., Fischhoff, B., Millstein, S. G., & Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty-fifty chance.". *Organizational Behavior and Human Decision Processes*, *81*(1), 115–131. https://doi.org/10.1006/obhd.1999.2868

Budescu, D. V., Karelitz, T. M., & Wallsten, T. S. (2003). Predicting the directionality of probability words from their membership functions. *Journal of Behavioral Decision Making*, *16*(3), 159–180. https://doi.org/10.1002/bdm.440

Budescu, D. V., Por, H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, *4*(6), 508–512. https://doi.org/10.1038/nclimate2194

Carroll, M. B., & Sanchez, P. L. (2021). Decision making with conflicting information: Influencing factors and best practice guidelines. *Theoretical Issues in Ergonomics Science*, *22*(3), 296–316. https://doi.org/10.1080/1463922X.2020.1764660

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

Collins, R. N., & Mandel, D. R. (2019). Cultivating credibility with probability words and numbers. *Judgment and Decision Making*, *14*(6), 683–695.

Collins, R. N., Mandel, D. R., & MacLeod, B. (2022). *Verbal and numeric probability information differentially shapes decisions*. https://doi.org/10.31234/osf.io/ad7gw

Derbentseva, N., McLellan, L., & Mandel, D. R. (2010). *Issues in intelligence production: Summary of interviews with Canadian intelligence managers* (DRDC Toronto Technical Report 2010–144). Defence Research and Development Canada.

Dhami, M. K., & Mandel, D. R. (2021). Words or numbers? Communicating probability in intelligence analysis. *The American Psychologist*, *76*(3), 549–560. https://doi.org/10.1037/amp0000637

Dowie, J. (1999). Against risk. *Risk, Decision and Policy*, *4*(1), 57–73. https://doi.org/10.1080/135753099348102

Fan, Y., Budescu, D. V., Mandel, D., & Himmelstein, M. (2019). Improving accuracy by coherence weighting of direct and ratio probability judgments. *Decision Analysis*, *16*(3), 197–217. https://doi.org/10.1287/deca.2018.0388

Fingar, T. (2011). *Reducing uncertainty: Intelligence analysis and national security*. Stanford Security Studies. https://doi.org/10.1515/9780804781657

Fischhoff, B., & Bruine de Bruin, W. (1999). Fifty-fifty = 50%? *Journal of Behavioral Decision Making*, *12*(2), 149–163. https://doi.org/10.1002/(SICI)1099-0771(199906)12:2<149::AID-BDM314>3.0.CO;2-J

Friedman, J. A. (2019). *War and chance: Assessing uncertainty in international politics*. Oxford University Press. https://doi.org/10.1093/oso/9780190938024.001.0001

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

Halpin, S. M., Moses, F. L., & Johnson, E. M. (1978). *A validation of the structure of combat intelligence ratings* (Technical Paper 302). US Army Research Institute for Behavioral and Social Sciences.

Ho, E. H., Budescu, D. V., Dhami, M. K., & Mandel, D. R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy*, *1*(2), 43–55. https://doi.org/10.1353/bsp.2015.0015

Icard, B. (2019). Lying, deception and strategic omission: Definition and evaluation. Doctoral dissertation. Retrieved from https://tel.archives-ouvertes.fr/tel-02170022

Irwin, D., & Mandel, D. R. (2019). Improving information evaluation for intelligence production. *Intelligence and National Security*, *34*(4), 503–525. https://doi.org/10.1080/02684527.2019.1569343

Irwin, D., & Mandel, D. R. (2022). Communicating uncertainty in national security intelligence: Expert and non-expert interpretations of and preferences for verbal and numeric formats. *Risk Analysis*. Advance online publication. https://doi.org/10.1111/risa.14009

Jenkins, S. C., Harris, A. J. L., & Lark, R. M. (2018). When unlikely outcomes occur: The role of communication format in maintaining communicator credibility. *Journal of Risk Research*, *22*(5), 537–554. https://doi.org/10.1080/13669877.2018.1440415

Johnston, R. (2005). *Analytic culture in the US intelligence community, an ethnographic study*. Central Intelligence Agency.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kaplan, S. (1997). The words of risk analysis. *Risk Analysis*, *17*(4), 407–417. https://doi.org/10.1111/j.1539-6924.1997.tb00881.x

Karvetski, C. W., & Mandel, D. R. (2020). Coherence of probability judgments from uncertain evidence: Does ACH help? *Judgment and Decision Making*, *15*(6), 939–958.

Kent, S. (1964). Words of estimative probability. *Studies in Intelligence*, *8*(4), 49–65.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management*, *40*(2), 133–146. https://doi.org/10.1016/S0378-7206(02)00043-5

Lemercier, P. (2014). The fundamentals of intelligence. In P. Capet & T. Delavallade (Eds.), *Information evaluation* (pp. 55–100). Wiley-ISTE. https://doi.org/10.1002/9781118899151.ch3

Levine, J. M., & Samet, M. G. (1973). Information seeking with multiple sources of conflicting and unreliable information. *Human Factors*, *15*(4), 407–419. https://doi.org/10.1177/001872087301500412

Lieberthal, K. (2009). *The U.S. intelligence community and foreign policy: Getting analysis right*. The Brookings Institution.

Lynch, J. G., & Ofir, C. (1989). Effects of cue consistency and value on base-rate utilization. *Journal of Personality and Social Psychology*, *56*(2), 170–181. https://doi.org/10.1037/0022-3514.56.2.170

Mandel, D. R. (2005). Are risk assessments of a terrorist attack coherent? *Journal of Experimental Psychology: Applied*, *11*(4), 277–288. https://doi.org/10.1037/1076-898X.11.4.277

Mandel, D. R. (2007). *Toward a concept of risk for effective military decision making* [Technical Report 2007–124]. Defence Research and Development Canada.

Mandel, D. R., & Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making*, *31*(1), 127–137. https://doi.org/10.1002/bdm.2055

Mandel, D. R., Dhami, M. K., Tran, S., & Irwin, D. (2021). Arithmetic computation with probability words and numbers. *Journal of Behavioral Decision Making*, *34*(4), 593–608. https://doi.org/10.1002/bdm.2232

Mandel, D. R., & Irwin, D. (2021a). Facilitating sender-receiver agreement in communicated probabilities: Is it best to use words, numbers or both? *Judgment and Decision Making*, *16*(2), 363–393.

Mandel, D. R., & Irwin, D. (2021b). Uncertainty, intelligence, and national security decisionmaking. *International Journal of Intelligence and Counterintelligence*, *34*(3), 558–582. https://doi.org/10.1080/08850607.2020.1809056

Mandel, D. R., Wallsten, T. S., & Budescu, D. V. (2021). Numerically bounded linguistic probability schemes are unlikely to communicate uncertainty effectively. *Earth's Future*, *9*, e2020EF001526. https://doi.org/10.1029/2020EF001526

McDowell, D. (2009). *Strategic intelligence: A handbook for practitioners, managers, and users* (rev. ed.). The Scarecrow Press.

McLachlan, D. (1968). *Room 39: Naval intelligence in action 1939–45*. Weidenfeld & Nicolson.

Miron, M. S., Patten, S. M., & Halpin, S. M. (1978). *The structure of combat intelligence ratings (technical paper 286)*. US Army Research Institute for Behavioral and Social Sciences.

North Atlantic Treaty Organization. (2003). *Standardization agreement 2511—Intelligence reports* (NATO STANAG 2511 (1st ed.). NATO Standardization Agency.

North Atlantic Treaty Organization. (2016). *Allied joint doctrine for intelligence procedures* (NATO AJP-2.1, edition B, version 1). NATO Standardization Office.

Pedersen, T., & Jansen, P. T. (2019). Seduced by secrecy - perplexed by complexity: Effects of secret vs open-source on intelligence credibility and analytic confidence. *Intelligence and National Security*, *34*(6), 881–898. https://doi.org/10.1080/02684527.2019.1628453

Phelps, R. H., Halpin, S. M., Johnson, E. M., & Moses, F. L. (1980). *Implementation of subjective probability estimates in army intelligence procedures: A critical review of research findings* (Research Report 1242). US Army Research Institute for Behavioral and Social Sciences.

Quist, A. (1993). *Security classification of information*. Oak Ridge National Laboratory.

Samet, M. G. (1975). Quantitative interpretation of two qualitative scales used to rate military intelligence. *Human Factors*, *17*(2), 192–202. https://doi.org/10.1177/001872087501700210

Sample, C., Jensen, M. J., Scott, K., McAlaney, J., Fitchpatrick, S., Brockinton, A., Ormrod, D., & Ormrod, A. (2020). Interdisciplinary lessons learned while researching fake news. *Frontiers in Psychology*, *11*, 537612–537612. https://doi.org/10.3389/fpsyg.2020.537612

Schneider, C. R., Freeman, A. L. J., Spiegelhalter, D., & van der Linden, S. (2022). The effects of communicating scientific uncertainty on trust and decision making in a public health context. *Judgment and Decision Making*, *17*(4), 849–882.

Slovic, P. (1966). Cue-consistency and cue-utilization in judgment. *The American Journal of Psychology*, *79*(3), 427–434. https://doi.org/10.2307/1420883

Teigen, K. H., & Brun, W. (1995). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. *Acta Psychologica*, *88*, 233–258. https://doi.org/10.1016/0001-6918(93)E0071-9

Teigen, K. H., & Brun, W. (1999). The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes*, *80*, 155–190. https://doi.org/10.1006/obhd.1999.2857

Thompson, J., Landee-Thompson, B., Fichtl, T., & Adelman, L. (1989). *Measurement and evaluation of military intelligence performance* (Technical Report 833). US Army Research Institute for Behavioral and Social Sciences.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29. https://doi.org/10.1037/h0071663

Travers, M., Van Boven, L., & Judd, C. (2014). The secrecy heuristic: Inferring quality from secrecy in foreign policy contexts. *Political Psychology*, *35*(1), 97–111. https://doi.org/10.1111/pops.12042

Treverton, G. F. (2001). *Reshaping national intelligence for an age of information*. Cambridge University Press. https://doi.org/10.1017/CBO9780511754470

United Kingdom Cabinet Office. (2018). *Government security classifications: May 2018* (Version 1.1). The author. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/715778/May-2018_Government-Security-Classifications-2.pdf

United States Marine Corps. (2018). *MCTP 2-10B, MAGTF Intelligence production and analysis*. The author. Retrieved from https://www.marines.mil/portals/1/Publications/MCTP%202-10B%20GN.pdf?ver=2019-01-31-111956-437

Wintle, B. C., Fraser, H., Wills, B. C., Nicholson, A. E., & Fidler, F. (2019). Verbal probabilities: Very likely to be somewhat more confusing than numbers. *PLoS ONE*, *14*(4), e0213522. https://doi.org/10.1371/journal.pone.0213522

Wyer, R. S. Jr. (1970). Information redundancy, inconsistency, and novelty and their role in impression formation. *Journal of Experimental Social Psychology*, *6*(1), 111–127. https://doi.org/10.1016/0022-1031(70)90079-X

AUTHOR BIOGRAPHIES

David R. Mandel, a Senior Defence Scientist, studies human judgment and decision-making. Mandel received the 2020 NATO SAS Panel Excellence Award for an international research activity he led. He serves on the editorial boards of *Decision*, *Futures and Foresight Science*, *Intelligence and National Security*, *and Judgment and Decision Making*.

Daniel Irwin holds an M.S. in Applied Intelligence from Mercyhurst University and works for the Government of Canada. Irwin's research focuses on the assessment and communication of uncertainty, especially in the domain of intelligence analysis. Irwin is a recipient of a 2020 NATO scientific excellence award.

Mandeep K. Dhami is a Professor of Decision Psychology, Middlesex University London, UK. She has worked as a Ministry of Defence Principal Scientist. Her expertise is judgment and decision-making, risk perception, and uncertainty communication. She received the 2020 NATO SAS Panel Excellence Award. Mandeep is a co-Editor of *Judgment and Decision Making*.

David V. Budescu is the Anne Anastasi Professor of Quantitative Psychology at Fordham University. His research is in the areas of human judgment, individual and group decision-making under uncertainty and information aggregation. He is the Editor of *Decision*, and past president of the Society for Judgment and Decision Making.

How to cite this article: Mandel, D. R., Irwin, D., Dhami, M. K., & Budescu, D. V. (2023). Meta-informational cue inconsistency and judgment of information accuracy: Spotlight on intelligence analysis. *Journal of Behavioral Decision Making*, *36*(3), e2307. https://doi.org/10.1002/bdm.2307