Judgment and Decision Making (2025), **vol. 20**:e36 1–21
doi:10.1017/jdm.2025.10007

EMPIRICAL ARTICLE

The effect of source reliability and information credibility on judgments of information quality in intelligence analysis

Megan O. Kelly [1,2,3], David V. Budescu[4], Mandeep Dhami[5] and David R. Mandel[1]

[1]Defence Research and Development Canada, Toronto, ON, Canada; [2]University of Waterloo, Waterloo, ON, Canada; [3]Princeton University, Princeton, NJ, USA; [4]Fordham University, New York City, NY, USA and [5]Middlesex University, London, UK

Corresponding authors: Megan O. Kelly and David R. Mandel; Emails: megan.kelly@princeton.edu; david.mandel@forces.gc.ca

Received: 30 April 2024; **Revised:** 13 June 2025; **Accepted:** 14 June 2025

Keywords: information quality; information credibility; source reliability; intelligence analysis; information evaluation; decision-making under uncertainty

Abstract

The quality of information that informs decisions in expert domains such as law enforcement and national security often requires assessment based on meta-informational attributes such as source reliability and information credibility. Across 2 experiments with intelligence analysts (n = 74) and nonexperts (n = 175), participants rated the accuracy, informativeness, trustworthiness, and usefulness of information varying in source reliability and information credibility. The latter 2 attributes were communicated using ratings from the Admiralty Code, an information-evaluation system widely used in the defence and security domain since the 1940s. Ratings of accuracy, informativeness, and likelihood of use were elicited as repeated measures to examine intraindividual reliability. Across experiments, intraindividual reliability was best when levels of source reliability and information credibility were moderately consistent compared to when they were maximally inconsistent (i.e., one low and one high) or maximally consistent (both high or low). As well, trustworthiness ratings depended more on source reliability than on information credibility. Finally, the likelihood of using information was consistently predicted by accuracy ratings and not by judged informativeness or trustworthiness. The current findings offer insights into the ability of experts and novices to reliably use information-evaluation systems for structuring human judgments about intelligence.

Across virtually every domain, individuals must integrate and evaluate information to support effective judgment and decision-making (e.g., eye-witness testimony, Brewer and Burke, 2002; navigation, Nardini et al., 2008; sensation and perception, Knill and Saunders, 2003; Landy et al., 2011; Zaki, 2013; emotion recognition, Juslin 2000, Leitzke and Pollak, 2016; weather forecasting, Lusk, 1993; consumer behavior, Miyazaki et al., 2005; Wall et al., 1991; auditing and accounting, Lambert and Peytcheva, 2020; clinical diagnoses, Hoffman et al., 1968; intelligence and national security, Mandel et al., 2023; Samet, 1975). Effective information evaluation requires access to useful, accurate, informative, and trustworthy information. However, the quality of information is often uncertain and may vary in terms of such criteria. This is particularly important to understand in high-stakes contexts such as defence and national security where intelligence professionals routinely evaluate information quality to inform critical and highly influential decisions.

The evaluation of information quality is an important precursor to making intelligence assessments that support defence and national security decision-making (Carter, 2009; Irwin and Mandel, 2019; NATO, 2016; Samet, 1975). To facilitate effective information evaluation, raw intelligence (e.g., communications from informants or intercepted electronic signals) is often accompanied by explicit *meta-informational* attributes that provide information *about* the information or raw intelligence—for example, the reliability of the intelligence source. Despite the routine occurrence of judging information quality, it is unclear how information-quality judgments are influenced by the consistency (or inconsistency) among the available meta-informational attributes. In the present work, we aim to better understand the various assessments made in response to 2 key meta-informational attributes, *source reliability* and *information credibility*, based on a widely implemented information-evaluation system used in intelligence analysis.

1. Information-evaluation processes

Formal information evaluation in the intelligence and national security domain often employs the *Admiralty Code* or *NATO System*, which was developed by the British Admiralty's Naval Intelligence Division during World War II (Hanson, 2015; Irwin and Mandel, 2019). The North Atlantic Treaty Organization (NATO) joint intelligence doctrine (among other organizations; Hanson, 2015; Miron et al., 1978; United States Department of the Army, 1951) uses the Admiralty Code for evaluating the quality of raw intelligence, which is often unverified, incomplete, and/or prone to inaccuracies or biases. The Admiralty Code uses a 6 × 6 alphanumeric rating system, which requires the evaluation of *source reliability* and *information credibility* independently from one another. See Table 1 for standards (updated NATO AJP-2.1, NATO, 2016).

In applying the Admiralty Code, source reliability has been linked to the confidence in the information source provided and past performance levels (Irwin and Mandel, 2019; McDowell, 2009). For additional context, an older standard for the Admiralty Code detailed that source reliability could range from a source that 'has proved unworthy of any confidence' ('E': unreliable) to a source that is 'tried and trusted' and 'depended upon with confidence' ('A': completely reliable; NATO STANAG 2511; NATO, 2003). The standards imply that source reliability can only be judged for sources that have been used in the past. Thus, if a source is completely novel, it is designated as 'reliability cannot be judged' ('F'). In general, source reliability can be thought of as the confidence associated with a source's long-term success in providing quality information (Irwin and Mandel, 2019; McDowell, 2009; NATO, 2003; Samet, 1975).

In contrast to source reliability, the information credibility scale of the Admiralty Code refers to the probability that the information being evaluated is true. The earlier standards outlined that information credibility can range from information that 'positively contradicts previously reported information' or 'conflicts with the established pattern of an intelligence target in a marked degree' ('5': improbable) to information that 'originates from another source than the already existing information on the same

Table 1. *NATO AJP-2.1 source reliability and information credibility scales.*

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

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

subject' ('1': confirmed by other sources; NATO STANAG 2511; NATO, 2003). Notably, completely novel information that 'provides no basis for comparison with any known behavior pattern of a target' is classified as 'truth cannot be judged' ('6'; NATO STANAG 2511; NATO, 2003). Generally, then, information credibility is based on the degree to which the information seems true, with a key factor in determining truth being confirmation from other independent sources (e.g., Irwin and Mandel, 2019; NATO, 2003; Samet, 1975).

Once the information has been 'encoded' with the alphanumeric designations for source reliability and information credibility independently, the information and codes are communicated for subsequent interpretation ('*decoding*'), often by an intelligence analyst who is distinct from the *encoding* process. For example, if information from a source judged as 'Completely Reliable' is perceived by the encoding party as 'Possibly True' in information credibility, the information would be communicated with the alphanumeric designation of 'A3' for subsequent decoding.

2. Challenges with applying the Admiralty Code

Despite the wide practice of the Admiralty Code, challenges with using the Admiralty Code to assess and communicate information quality are evident. For example, when assigning (i.e., encoding) the alphanumeric codes, analysts or operators most often assign codes that fall along the diagonal such that source reliability and information credibility align perfectly in the ordinal position (i.e., assigned codes of A1, B2, C3, etc.; Baker et al., 1968). That is, encoders do not tend to assign inconsistent meta-informational attributes such as high source reliability and low information credibility, or vice versa. Relatedly, encoders may use source reliability as a cue for information credibility and not vice versa (Miron et al., 1978).

These findings indicate that analysts may not be assigning source reliability and information credibility classifications independently from one another despite the instruction within the Admiralty Code to do so (Mandel et al., 2023; Samet, 1975). Nevertheless, national intelligence communities commonly advise that intelligence assessments, which do share dependencies, be given independently from one another. For instance, the Office of Director of National Intelligence advises that analysts assign probability and confidence assessments independently, even though probabilities do, in fact, put boundaries on confidence levels. For instance, an event that is judged to have a 99% chance of occurring can have at most an upper bound of 100% on a confidence interval around the estimate. Such a small room for error would be inconsistent with expressing 'low confidence'. Therefore, the guidance from the intelligence community that such a probability could be coupled with low confidence is ill advised, and recent research shows that analysts and nonexperts alike do treat probability and confidence as related constructs (Irwin and Mandel, 2023). Similarly, it is unclear whether information evaluators are capable of treating source reliability and information credibility as fully independent, even if such guidance were prescriptively valid.

The research discussed thus far has focused on the unreliability of the *encoding* processes of the Admiralty Code. However, further ambiguity in applying the Admiralty Code may be evident during the *decoding* process in which analysts use such meta-information to weigh the raw intelligence they receive. This issue is the focus of the current work. In past research, Samet (1975) compared the consistency of Admiralty Code interpretations across 2 separate tasks in US army captains. In the first task, participants selected which of 2 camps were more likely to be attacked provided the camps differed in source reliability and/or information credibility (e.g., Camp X: A3 vs. Camp Y: C1). In the second task, participants assigned the probability of event likelihood for each of the alphanumeric codes involving A–E source reliability and 1–5 information credibility. If A3 was judged as *more likely* than C1 in the first task, and if probabilities x and y were assigned to A3 and C3, respectively, in the second task, then responses between tasks were considered *consistent* if $x > y$, *ambiguous* if $x = y$, and *inconsistent* if $x < y$. Samet (1975) found that across 2 different tasks, response consistency was lower than expected: consistency was ambiguous or inconsistent for one-third of the cases, and the vast majority of participants were inconsistent at least some of the time.

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

These findings suggest that, at an intraindividual level, users of the Admiralty Code interpret the same meta-information inconsistently. This poses issues for downstream intelligence analysis (i.e., use of that information in making intelligence assessments) and is compatible with the idea that intelligence processes that aim to support analysts' judgments may inadvertently amplify noise (or unreliability) due to undetected ambiguities or algorithmic imprecision in how those processes should be executed (Chang et al., 2018; Mandel, 2020).

3. Consistency among meta-information

Multiple intelligence processes intended to facilitate formal structure in analysis procedures and increase intelligence accountability have exhibited unreliability (Chang et al., 2018; Karvetski and Mandel, 2020; Marcoci et al., 2019a, 2019b). For instance, in the Admiralty Code, the consistency or reliability within individuals' interpretations of the alphanumeric codes could depend on whether the meta-informational attributes of source reliability and information credibility are consistent with one another. We sought to examine how the reliability within individuals' assessments might be influenced by the agreement among source reliability and information credibility (e.g., both high source reliability and high information credibility vs. high source reliability and low information credibility).

While the influence of consistency between the source reliability and information credibility on information assessment is underexplored in the context of the Admiralty Code, work in other domains suggests *attribute consistency* is an important factor in effective or reliable use of attributes. For example, Slovic (1966) examined how the use of attributes in judgments depended on attribute consistency and found that both attributes contributed to judgments if both of them agreed, but that only one of these attributes were used when they disagreed. In a similar vein, Miyazaki et al. (2005) tested the effect of consistency in 2 product attributes on the price–quality relationship for a product and found that 2 positive attributes (e.g., strong brand, strong warranty) led to a clear price–quality relation for the product. When a product had both a positive attribute *and* a negative attribute, consumers tended to focus on the negative attribute during evaluation. Thus, individuals may opt to favor one attribute over the other in the case of disagreement. Individuals may also resort to 'averaging' the attributes when receiving inconsistent information (e.g., Lichtenstein et al., 1975). More generally, these studies support an *attribute consistency hypothesis* in which intrajudge reliability is expected to increase with increases in attribute consistency.

In the current context of the Admiralty Code, Mandel et al. (2023) examined whether *consistency* in the levels of source reliability and information credibility affected intraindividual reliability (i.e., test–retest) in ratings of information accuracy. In support of the attribute consistency hypothesis, they found that intraindividual reliability was directly related to attribute consistency. If attribute consistency was high (as in A1 or E5 ratings), reliability was significantly better than if attribute consistency was low (as in A5 or E1 ratings). These findings suggest that individuals may experience increased difficulty in assigning reliable judgments when presented with inconsistent attributes compared to consistent ones.

4. The present research

4.1. Effect of attribute consistency on intraindividual reliability

A central aim of the current work was to further examine the relation between intraindividual reliability in information quality ratings and the level of meta-informational attribute consistency. Critically, we aimed to do so with improved statistical power by collecting larger samples than those in earlier work. Across 2 experiments, we elicited information quality ratings in response to varying levels of source reliability and information credibility. Crossing levels of source reliability and information credibility allowed for the key manipulation of attribute consistency such that information presented to individuals was either of low, medium, or high attribute consistency.

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

Mandel et al.'s (2023) research provides initial evidence that analysts have difficulty integrating source-reliability and information-credibility information when the 2 attributes indicate disparate levels of information quality. However, their experiments were conducted on small samples of intelligence analysts and focused on a single dependent measure, namely, the probability that the information received was accurate. Although information accuracy undoubtedly contributes to information quality, information quality is a multifaceted construct. Other indicators include the *informativeness* of information (Batini and Scannapieco, 2016; Eppler and Wittig, 2000; Stvilia et al., 2007), its *trustworthiness* (Batini and Scannapieco, 2016; Nurse et al., 2011; Stvilia et al., 2007), and the receiver's intention to *use* the information (Batini and Scannapieco, 2016; Huang et al., 1999; Madnick et al., 2009). In the present work, we studied intraindividual reliability across 3 measures (i.e., accuracy, informativeness, and likelihood of use), which were assessed in 2 blocks in a repeated measures. In addition to studying reliability across more dependent measures, we examined how results generalized to nonexperts unfamiliar with the Admiralty Code (Experiment 2). This extension to nonexperts in Experiment 2 is of interest when considering that previous work has focused on sampling analysts specifically.

Another motivation for the current work was to ensure the effects reported by Mandel et al. (2023) were not an artifact of the scoring rule they used. Mandel et al. (2023) indexed intraindividual reliability by computing a standardized absolute error (SAE) measure between a first and second probability judgment for items asked twice, wherein p_1 and p_2 refer to the ratings elicited at test and retest, respectively:

$$\mathrm{SAE} = |p_1 - p_2| / \mathrm{Max}[p_1, p_2, 100 - p_1, 100 - p_2]$$

With SAE, however, error is expected to be inversely related to the extremity of the mean of the 2 repeated judgments. As judgments of highly congruent Admiralty Codes tend to have the most extreme mean values, Mandel et al.'s (2023) results may be due to this confounding influence. In the present research, we examine the reliability of judgments as a function of attribute congruence using a novel scoring rule (described later) that, contrary to SAE, mildly penalizes extreme scores.

4.2. Effect of meta-information on information quality

While we expect general positive effects of source reliability and information credibility across indices of information quality as found by Mandel et al. (2023), another aim of this work was to examine how information-quality measures might be differentially affected by source reliability and information credibility given previous mixed findings. Mandel et al. (2023) proposed that information accuracy and information credibility both directly refer to characteristics of the information being evaluated, whereas source reliability indirectly describes that information because it refers instead to the source (see also Lemercier, 2014). This proposal was indirectly supported by Samet (1975), who found greater weighting of information credibility than source reliability on likelihood judgments. However, Mandel et al. (2023) did not find evidence of unequal weighting of these 2 factors on judgments of information accuracy. In contrast, Samet (1975) found that participants' event likelihood estimates depended more strongly on information credibility than source reliability. That said, estimates were not a direct measure of information quality in Samet (1975) as they were in Mandel et al. (2023), and this difference might account for the apparent inconsistency. It remains plausible that other quality indices encourage differential weighting. For example, given the well-documented relation between source and trustworthiness (e.g., Bertino and Lim, 2010; Dai et al., 2008; Gil and Artz, 2007; Nurse et al., 2011), an assessor might weigh source reliability heavier than information credibility when assessing the trustworthiness of the information. We extended the previous work to test for differential effects of source reliability and information credibility on 4 indices of information quality (i.e., accuracy, informativeness, trustworthiness, and usefulness).

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

4.3. Predictors of likelihood of information use

Finally, the intent to use information represents an ultimately consequential aspect of information quality. Evaluating information quality will affect whether the analyst or consumer uses the information. Therefore, a final aim of this research was to examine whether the rated probability of using information characterized by a particular meta-informational profile was influenced by the other information-quality measures (i.e., accuracy, informativeness, and trustworthiness). We assessed the predictive effect of those measures across differing meta-informational profiles in an exploratory manner.

5. Experiment 1

5.1. Method

5.1.1. Participants

Participants were 74 intelligence analysts (n = 52 Canadian; n = 22 UK) who participated remotely using a Qualtrics survey link distributed by their managers. Participation was voluntary, anonymous, and there was no remuneration. Sixty-two percent of the sample indicated being familiar with the source-reliability and information-credibility scales. Demographic information was available from 45.9% of participants, (the demographics for some analysts were not reported to protect their identities). Of those reporting this information, 23.5% reported as female and the age range was 21 to 60 years (M = 35.06, SD = 9.45). The reported highest level of education attained was high school or equivalent for 5.9%, trade school (nonmilitary) for 11.8%, undergraduate for 52.9%, and graduate for 29.4%. In all, 53% stated that they were civilian and 47% reported being military (all participants responded).

5.1.2. Design

The experiment was a mixed design with one between-subjects factor (item set) and 3 within-subjects factors (information-quality measure, source reliability, and information credibility). Specifically, participants were randomly assigned to 1 of 2 experimental stimuli conditions (item set 1, item set 2; explained further below). Each participant provided ratings for each information-quality measure (i.e., accuracy, informativeness, likelihood of use, and trustworthiness) as a function of both source reliability (low, medium, high), and information credibility (low, medium, high). In addition, given the varying levels of source reliability and information credibility, the attribute consistency between source reliability and information credibility also varied within-subjects. Specifically, the consistency in the levels of source reliability and information credibility could be low consistency (e.g., low source reliability and high information credibility), medium consistency (e.g., low source reliability, medium information credibility), or high consistency (e.g., low source reliability, low information credibility). As shown in Table 2, both item sets included the same low- and high-attribute consistency items (A5, E1; A1, E5), but different medium-attribute consistency items (Set 1: A3, C1; Set 2: C5, E3).

5.1.3. Materials and procedure

Experiment 1 was approved by the Defence Research and Development Canada Human Research Ethics Committee. At the start of the experiment, participants were informed that they would be judging pieces of information classified as TOP SECRET along different dimensions, including accuracy, likelihood of use, informativeness, trustworthiness, and prevalence. Next, they were shown the source-reliability and information-credibility scales, followed by an annotated sample question (see the Supplementary Material).

Before advancing to the experimental task, participants had to pass 2 screening questions. First, they were asked to select the highest source reliability rating from a set of 4: ‘Reliability cannot be judged (F)’, ‘Not Usually Reliable (D)’, ‘Fairly Reliable (C)’, and ‘Usually Reliable (B)’. Participants who provided an incorrect response were prompted to review the scales and try again. Those who failed on their second attempt were automatically screened out of the experiment. This process was repeated

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

Table 2. *Item characteristics by set in Experiments 1 and 2.*

Item rated from code	Set	Source reliability (SR): A–E	Information credibility (IC): 1–5	SR–IC attribute consistency
A1	1 and 2	High (A)	High (1)	High
E5	1 and 2	Low (E)	Low (5)	High
A5	1 and 2	High (A)	Low (5)	Low
E1	1 and 2	Low (E)	High (1)	Low
A3	1 only	High (A)	Medium (3)	Medium
C1	1 only	Medium (C)	High (1)	Medium
C5	2 only	Medium (C)	Low (5)	Medium
E3	2 only	Low (E)	Medium (3)	Medium

with a question asking participants to select the lowest information credibility rating from a set of 4: 'Doubtful (4)', 'Possibly True (3)', 'Improbable (5)', and 'Completely Credible (1)'.

Eliciting judgment ratings. The experimental task comprised 4 blocks of questions about information quality. Within each block, participants were shown the 6 pieces of information corresponding to the condition they were randomly allocated to (i.e., Set 1 or Set 2), one at a time and in random order. For each item, participants responded to 2 of 5 questions about information quality below (determined by the block) using 100-point sliders with a starting position of 50:

1. Use: 'How likely would you be to *use* a piece of information with this combination of reliability and credibility ratings in your analysis?'[1] [0 = Extremely unlikely, 100 = Extremely likely]
2. Accuracy: 'How likely is a piece of information with this combination of reliability and credibility ratings to be *accurate*?' [0 = Extremely unlikely, 100 = Extremely likely]
3. Informativeness: 'How *informative* would a piece of information with this combination of reliability and credibility ratings be?' [0 = Not at all informative, 100 = Extremely informative]
4. Trustworthiness: 'How *trustworthy* do you think a piece of information with this combination of reliability and credibility ratings would be?' [0 = Not at all trustworthy, 100 = Extremely trustworthy]
5. Prevalence: 'How likely is it to *encounter* a piece of information with this combination of reliability and credibility ratings?' [0 = Extremely unlikely, 100 = Extremely likely][2]

The information quality questions were posed to participants as follows: *Block 1* elicited *accuracy* and *likelihood of use* judgments; *Block 2* elicited *trustworthiness* and *informativeness* judgments; *Block 3* elicited *prevalence* and, for a *second* time, *likelihood of use* judgments; finally, *Block 4* elicited the *second* judgments for *informativeness* and *accuracy*. Hence, likelihood of use, informativeness, and accuracy judgments were each elicited *twice* in total, while trustworthiness and prevalence judgments were each elicited *once*. Critically, because judgments of accuracy, informativeness, and likelihood of use were elicited twice—that is, as *test* and *retest*—this allowed for a measure of intraindividual reliability.

For half of the participants, question order was reversed within blocks (e.g., Block 1: likelihood of use, then accuracy; Block 2: informativeness, then trustworthiness). Between blocks, a buffer printed in capital letters and italics stated, '*Note: the questions you will be asked are changing. Please read*

[1]For the participants in the condition that had item set 1 and question order 1, the item was actually stated unintentionally as 'How likely would you be to include a piece of information with this combination of reliability and credibility ratings in your analysis?' instead.

[2]Although we collected prevalence ratings (see question wordings below), we opted to exclude prevalence as a dependent measure of interest because, upon reflection, we did not consider it to be an indicator of *information quality*—a key focus in our investigation. Therefore, prevalence ratings results are not discussed, although they are available in the Supplementary Material for the relevant key main analyses (the results were qualitatively the same).

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

them carefully!' Participants provided 48 responses in total (i.e., 4 blocks × 6 items × 2 questions). Following the experimental task, participants completed brief tasks unrelated to the current research. Finally, all participants indicated their familiarity with the rating scales, provided basic demographic information, and were debriefed.

5.1.4. Analyses

Analyses were conducted using R version 4.3.1. For the relevant within-participants variables for analysis of variance or covariance (ANOVA or ANCOVA, respectively) ('*ez*'; Lawrence, 2016), Greenhouse-Geisser corrected degrees of freedom, and p values are reported where applicable (i.e., whenever Mauchly's test for sphericity was significant). For any regression analyses conducted ('*stats*' package; R Core Team, 2023), multicollinearity, heteroscedasticity, and normality of residuals were all checked using the *vif* ('*car*' package; Fox and Weisberg, 2019), *bptest* ('*lmtest*' package; Zeileis and Hothorn, 2002), and *qqnorm* ('*stats*' package; R Core Team, 2023) functions, respectively. Effect sizes from ANOVAs are generalized eta squared, $\eta_G{}^2$, which is more comparable across within- and between-participant designs and are expected to be smaller than partial eta squared in repeated-measures designs with multiple factors (Bakeman, 2005; Olejnik and Algina, 2003).

5.2. Results and discussion

5.2.1. Effect of meta-information on information quality

We examined the influences of source reliability and information credibility on ratings using a 2 (source reliability: low, high) × 2 (information credibility: low, high) × 2 (set: 1, 2) × 4 (measure: use, accuracy, informativeness, trustworthiness) mixed ANOVA, wherein item set was the only between-participants factor. There were main effects of source reliability [$F(1, 72) = 562.06$, $p < .001$, $\eta_G^2 = .51$] and information credibility [$F(1, 72) = 490.91$, $p < .001$, $\eta_G^2 = .48$] such that ratings were higher when source reliability and information credibility were higher. There was also a main effect of information-quality measure [$F(2.60, 187.33) = 4.43$, $p = .007$, $\eta_G^2 = .01$] such that ratings were lowest for the accuracy measure and highest for informativeness. Source reliability and measure significantly interacted [$F(2.58, 186.03) = 23.64$, $p < .001$, $\eta_G^2 = .04$] such that the positive effect of high source reliability (vs. low source reliability) on ratings was largest for trustworthiness and smallest for informativeness. Information credibility and measure significantly interacted [$F(2.68, 192.91) = 15.38$, $p < .001$, $\eta_G^2 = .03$] such that the positive effect of high information credibility on ratings (vs. low information credibility) was largest for accuracy and smallest for trustworthiness. Finally, source reliability and information credibility also significantly interacted [$F(1, 72) = 30.43$, $p < .001$, $\eta_G^2 = .03$] such that the positive effect of information credibility was significantly larger for a high reliability source than a low one. No other main or interaction effect was significant [Fs < 3.57, ps > .060]. Importantly, as Figure 1 shows, the interaction effects reflect variations in the degree of the observed effects but not in the direction of those effects. In all cases, higher source reliability and information credibility were associated with higher information quality ratings.[3]

To explore whether source reliability and information credibility differed from one another in their influence on each information-quality measure, we compared 3 pairs of items that were complementary in terms of high/low source reliability and low/high information credibility. For example, A5 (highest source reliability and lowest information credibility) was compared to E1 (lowest source reliability and highest information credibility). The same was performed for A3 versus C1 (observations from item set 1), and C5 versus E3 (observations from item set 2). Table 3 presents the results of these paired comparisons for each dependent variable. Comparing A5 and E1, ratings for information accuracy were significantly higher for E1 than A5, indicating that information credibility was weighted more heavily than source reliability [$t(73) = 2.30$, $p = .024$, $d = 0.27$]. However, the comparisons between A3 and C1 and C5 and E3 showed no such significant effect for accuracy ratings [ts < 1, ps > .599]. In fact,

[3]The Supplementary Material presents figures analogous to Figure 1 that are disaggregated by dependent measure.

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

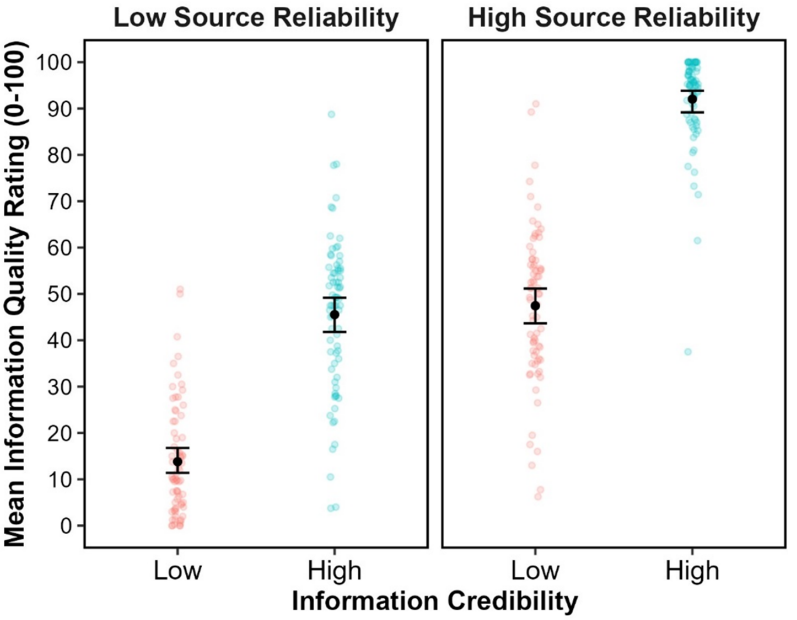

Figure 1. *Mean information quality rating (i.e., averaged across accuracy, informativeness, trustworthiness, and likelihood of use) as a function of source reliability and information credibility in Experiment 1. Error bars represent 95% bootstrap bias corrected and accelerated CIs using 10,000 samples.*

when comparing within item pairs (A5 vs. E1, A3 vs. C1, C5 vs. E3) in their ratings for probability of using information and degree of informativeness, there was no evidence of a significant difference. The exception to this was trustworthiness ratings: in all 3 comparisons, the item with higher source reliability had higher ratings compared to the item with higher information credibility. That is, A5 was rated significantly more trustworthy than E1, A3 was rated significantly more trustworthy than C1, and C5 significantly more trustworthy than E3. Recall that participants were asked to judge the trustworthiness of the information rather than the source. These results suggest that when information quality is framed in terms of trustworthiness, greater weight will be given to the source even if directly information-relevant meta-information is provided.

5.2.2. Effect of attribute consistency on intraindividual reliability

As noted earlier, an aim of our research was to examine whether the attribute consistency hypothesis would be supported when reliability was scored using a rule that did not yield higher error scores because the 2 repeated judgments were more extreme. Letting p_1 and p_2 refer to the judgments elicited at test and retest, respectively, the maximum absolute normalized error (MANE) is defined as follows:

If $p_2 = p_1$:

$$\text{MANE} = 0 \ \ (\text{perfect reliability})$$

If $[100 - \text{Min}(p_1, p_2)] > \text{Max}(p_1, p_2)$:

$$\text{MANE} = |p_1 - p_2| \,/ \text{Max}\,(p_1, p_2)\,,$$

Else:

$$\text{MANE} = |p_1 - p_2| \,/ [100 - \text{Min}\,(p_1, p_2)]\,.$$

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

Table 3. *Effect of source-reliability/information-credibility trade-offs on mean information-quality measures in Experiment 1.*

Stimuli pair and measure	Mean (*SD*)	*t*	*df*	*p*	Cohen's *d*
A5–E1					
Accuracy	M_{A5} = 39.11 (24.15)	2.30*	73	**.024**	0.27
	M_{E1} = 49.07 (23.76)				
Use	M_{A5} = 42.55 (24.77)	0.79	73	.430	0.09
	M_{E1} = 45.42 (24.75)				
Informativeness	M_{A5} = 47.36 (23.92)	1.38	73	.172	0.16
	M_{E1} = 52.12 (21.86)				
Trustworthiness	M_{A5} = 60.76 (23.85)	6.17***	73	**<.001**	0.72
	M_{E1} = 35.53 (24.04)				
A3–C1					
Accuracy	M_{A3} = 67.51 (15.73)	0.42	34	.677	0.07
	M_{C1} = 66.49 (15.52)				
Use	M_{A3} = 71.83 (16.71)	1.61	34	.117	0.27
	M_{C1} = 67.77 (17.46)				
Informativeness	M_{A3} = 70.26 (11.62)	0.54	34	.596	0.09
	M_{C1} = 68.91 (10.04)				
Trustworthiness	M_{A3} = 77.11 (11.55)	3.77**	34	**.001**	0.64
	M_{C1} = 65.77 (14.44)				
C5–E3					
Accuracy	M_{C5} = 32.15 (17.18)	0.27	38	.786	0.04
	M_{E3} = 33.05 (13.61)				
Use	M_{C5} = 38.03 (21.21)	1.09	38	.282	0.17
	M_{E3} = 34.38 (21.68)				
Informativeness	M_{C5} = 40.00 (19.15)	0.66	38	.511	0.11
	M_{E3} = 42.21 (20.84)				
Trustworthiness	M_{C5} = 41.72 (16.40)	4.19**	38	**<.001**	0.67
	M_{E3} = 27.59 (17.34)				

Note: Bold *p*-values are statistically significant.

Thus, MANE calculates 2 absolute normalized error scores and selects the one that maximizes error.[4] The exception is where scores perfectly agree and error is set to 0 by default.

We conducted a 3 (consistency: low, medium, high) × 3 (measure: accuracy, informativeness, use) × 2 (set: 1, 2) mixed ANOVA on MANE. There was a significant main effect of consistency [$F(1.78, 127.82) = 12.47$, $p < .001$, $\eta_G^2 = .04$] such that the medium consistency led to significantly lower MANE than both low consistency [$t(73) = 2.42$, $p = .018$, $d = 0.28$] and high consistency [$t(73) = 5.56$, $p < .001$, $d = 0.65$], with MANE lower for low consistency than for high consistency [$t(73) = 2.33$, $p = .023$, $d = 0.27$]. No other main or interaction effects were statistically significant [Fs < 2.33, ps > .100]. Figure 2 presents mean MANE and 95% bootstrap confidence intervals according to each level of attribute consistency within each item set for Experiment 1.[5] These findings indicate that the earlier attribute consistency effect reported by Mandel et al. (2023) may have been due to the choice

[4]We thank Jon Baron for suggesting this rule. We adapted it slightly by setting cases where the two judgments were identical to MANE = 0 to avoid missing cases.

[5]The analogous means and 95% CIs as a function of dependent measure are presented in the Supplementary Material.

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

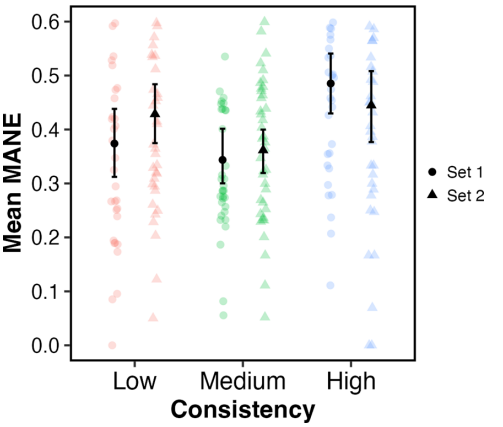

Figure 2. *Mean maximum normalized error (MANE) by consistency and set in Experiment 1. Error bars are bias corrected and accelerated 95% bootstrap CIs using 10,000 samples.*

of scoring rule, which penalizes judgment pairs with less extreme mean values. In the present study, contrary to the consistency effect, highly consistent attributes were the least reliable.

To examine the relation between an individual's mean MANE across levels of consistency, we conducted 3 separate correlations: (i) the mean MANE at low consistency with the mean MANE at medium consistency, (ii) the mean MANE at medium consistency with the mean MANE at high consistency, and (iii) the mean MANE at low consistency with the mean MANE at high consistency. Mean MANE correlated positively between low and medium levels of consistency [$t(72) = 4.04$, $r = .43$, $p < .001$], correlated positively between medium and high levels of consistency [$t(72) = 5.01$, $r = .51$, $p < .001$], and correlated positively between low and high levels of consistency [$t(72) = 2.43$, $r = .28$, $p = .017$]. Therefore, it appears there are stable individual differences in reliability across levels of attribute consistency.

5.2.3. Predictors of likelihood of information use

We regressed participants' ratings of likelihood of information use on their ratings of perceived information accuracy, informativeness, and trustworthiness for each of the 4 items common to Sets 1 and 2 (i.e., A1, A5, E1, and E5). Each of the 4 models was significant. Table 4 shows the overall model statistics for the 4 models and the statistics for each parameter in the 4 models.[6] Across stimuli, accuracy ratings significantly predicted likelihood of use ratings. Additionally, when source reliability was low and information credibility was high, informativeness ratings significantly predicted likelihood of use ratings. No other relations were significant. The findings suggest that judged information accuracy is the principal determinant of intention to use information.

6. Experiment 2

In Experiment 2, we sought to examine whether the findings observed in Experiment 1 are specific to professional intelligence analysts who may have received training in information-evaluation methods or whether those with little or no prior experience in information-evaluation methods would show similar findings. While we do not presume that the Admiralty Code itself is relevant for use in the general population, the generalizability of findings across expert and nonexpert samples would indicate that

[6]Results were qualitatively the same when excluding potentially influential observations and participants with potentially influential observations (identified with *influence_plot()* from the *regclass* package; Petrie, 2020), see analogous results for these analyses in the Supplementary Material.

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

Table 4. *Models fitted for each stimulus and predictor of use likelihood ratings for Experiment 1.*

Stimulus, model fit, and predictor	*b*	*SE*	*β*	*t*	*p*
A1 (high SR, high IC):					
$F(3, 70) = 23.10$, $p < .001$, $R^2_{\text{Adj.}} = .48$					
Accuracy	0.80	0.12	0.78	6.77	**<.001**
Informativeness	–0.13	0.10	–0.11	1.27	.208
Trustworthiness	0.13	0.22	0.13	0.57	.573
A5 (high SR, low IC):					
$F(3, 70) = 13.98$, $p < .001$, $R^2_{Adj.} = .35$					
Accuracy	0.50	0.11	0.48	4.59	**<.001**
Informativeness	0.15	0.10	0.13	1.45	.151
Trustworthiness	0.17	0.11	0.17	1.60	.113
E1 (low SR, high IC):					
$F(3, 70) = 27.63$, $p < .001$, $R^2_{\text{Adj.}} = .52$					
Accuracy	0.61	0.09	0.59	6.76	**<.001**
Informativeness	0.33	0.10	0.30	3.30	**.002**
Trustworthiness	–0.01	0.09	–0.01	0.14	.892
E5 (low SR, low IC):					
$F(3, 70) = 6.05$, $p = .001$, $R^2_{\text{Adj.}} = .17$					
Accuracy	0.44	0.15	0.43	2.96	**.004**
Informativeness	0.01	0.10	0.01	0.10	.923
Trustworthiness	0.21	0.15	0.21	1.46	.150

Note: SR, source reliability; IC, information credibility. Bold *p*-values are statistically significant.

general cognitive processes underlie the results. In contrast, if such findings from expert populations do not generalize to nonexpert samples, it might suggest that responding in these information systems is driven by analyst training. Finally, testing the generalizability of the results from expert analysts to nonexperts could be relevant for better understanding the application of systems analogous to the Admiralty Code to other areas of specialization, such as education and training, given the broad utility of properly assessing information when provided meta-informational attributes such as source reliability or information credibility (e.g., Hanson, 2015).

We also extended Experiment 1 (and earlier work by Mandel et al. 2023) by exploring how potentially relevant individual differences in cognition may influence the reliability of individuals' responses. Specifically, individuals who are more likely to answer intuitively—or less likely to answer accurately—might be less reliable or reflective in their responses. We provide a test of this general idea by considering both the intuitive response rate and the overall accuracy on the cognitive reflection task (CRT; Frederick, 2005) as proximate indicators of less reflective responding (Erceg and Bubić, 2017; Piazza and Sousa, 2014; Shenhav et al., 2012 Pennycook et al., 2016). This permits the inclusion of the intuitive response rate and accuracy rate each in an analysis of intraindividual reliability in quality measures, provided such measures relate to reliability.

6.1. Method

6.1.1. Participants

Experiment 2 was approved by the Defence Research and Development Canada Human Research Ethics Committee. Data was collected from 175 participants using Qualtrics Panels. Participants were sampled from Canada and the United States and were required to have English as their first language. They were prohibited from completing the experiment on a smartphone. The sample was 62.9%

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

female and sample age ranged from 18 to 60 years (M = 40.26, SD = 12.84). The highest level of education attained reported was high school or equivalent for 36.0%, trade school (non-military) for 21.1%, undergraduate for 37.7%, and graduate-masters for 4.6% with 1 participant not reporting. All participants stated that they were 'civilian', and 13.1% reported prior familiarity with the source reliability or information credibility scales.[7]

6.2. Design, materials, and procedure

The design was the same as in Experiment 1 and the materials and procedure were also the same as Experiment 1 with one exception. Specifically, all participants also completed the CRT after the core task of the experiment before responding to basic demographic questions and debriefing. Two participants did not complete the CRT and were excluded from analyses involving the CRT scores.

6.2.1. Analyses

The intuitiveness and accuracy scorings of the CRT items were performed based on earlier work such that the intuitive responses were coded with '1' with other response types coded as '0' and such that, for accuracy, correct responses were coded with '1' and incorrect responses with '0' (Erceg and Bubić, 2017; Piazza and Sousa, 2014; Shenhav et al., 2012; alternatively, see Pennycook et al., 2016). Analyses were conducted in R and Greenhouse-Geisser corrected degrees of freedom and p values are reported where applicable for the within-participant factors. Analyses were not preregistered.

6.3. Results and discussion

6.3.1. Effect of meta-information on information quality

We examined the influences of source reliability and information credibility on ratings using a 2 (source reliability) × 2 (information credibility) × 2 (set) × 4 (measure) mixed ANOVA with set as the only between-participants factor. There were significant main effects of source reliability [$F(1, 173) = 393.52$, $p < .001$, $\eta_G^2 = .27$] and information credibility [$F(1, 173) = 398.41$, $p < .001$, $\eta_G^2 = .26$], such that ratings were higher when source reliability and information credibility were higher. As in Experiment 1, there was also a significant interaction between source reliability and measure [$F(2.53, 437.74) = 6.84$, $p < .001$, $\eta_G^2 < .01$] such that the positive effect of high source reliability (vs. low source reliability) was the largest for trustworthiness ratings and the smallest for informativeness ratings. Information credibility and measure also significantly interacted [$F(2.82, 487.33) = 3.73$, $p = .014$, $\eta_G^2 < .01$] such that the positive effect of high (vs. low) information credibility was largest for accuracy ratings and smallest for trustworthiness ratings, which was also seen in Experiment 1. An interaction between source reliability and information credibility [$F(1, 173) = 32.72$, $p < .001$, $\eta_G^2 = .01$] revealed that the positive effect of information credibility was larger for a high-reliability source than a low one, like in Experiment 1. The interaction effects reflect variations in the degree of the observed effects but not in the direction of those effects such that higher source reliability and information credibility was associated with higher information quality ratings. No other effects or interactions were statistically significant [Fs < 1.97, ps > .110]. Figure 3 presents the mean values across the information-quality measures (i.e., accuracy, informativeness, trustworthiness, and likelihood of use) as a function of source reliability and information credibility (see the Supplementary Material for the analogous figures disaggregated by measure).

To examine the differential influences of source reliability and information credibility on the separate dependent measures, we compared 3 pairs of items that were complementary in terms of high/low source reliability and low/high information credibility as in Experiment 1 (i.e., A5-E1, A3–C1, and C5–E3). Table 5 presents the results of these paired comparisons for each dependent variable. Comparing

[7]The rate of prior familiarity might be inflated if participants interpreted prior familiarity as experience with similar constructs in other studies (i.e., not prior familiarity with the NATO Admiralty Code specifically).

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

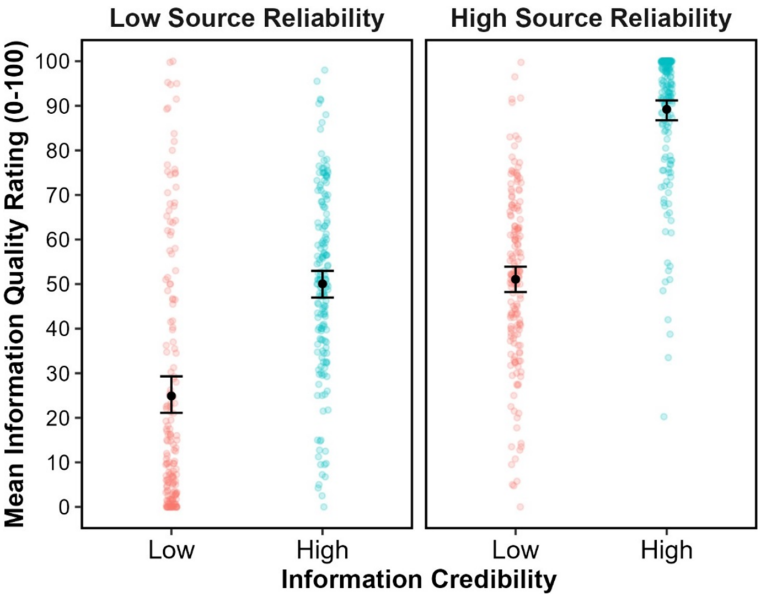

Figure 3. *Mean information quality rating (i.e., averaged across accuracy, informativeness, trustworthiness, and likelihood of use) as a function of source reliability and information credibility in Experiment 2. Error bars are bias corrected and accelerated 95% bootstrap CIs using 10,000 samples.*

within the item pairs (A5 vs. E1, A3 vs. C1, C5 vs. E3), there was no evidence of any significant differences in ratings of accuracy, informativeness, or likelihood of use (ts ≤ 1.67, ps ≥ .096, ds ≤ 0.13). Consistent with the findings from Experiment 1, the exception to this was trustworthiness ratings such that, in 2 out of 3 comparisons (A5–E1 and C5–E3), the item with higher source reliability had higher ratings compared to the item with higher information credibility. However, unlike in Experiment 1, A3 was *not* rated significantly more trustworthy than C1, though it is in the numerically consistent direction as the other comparisons, with A5 having the numerically higher trustworthiness rating.

6.3.2. Effect of attribute consistency on intraindividual reliability

Participants' mean MANE scores did not significantly correlate with the proportion of intuitive responding on the CRT [$t(171) = 1.61$, $r = .12$, $p = .110$] or the proportion of accurate responding on the CRT [$t(171) = 1.22$, $r = -.09$, $p = .225$]. Therefore, these CRT measures were not included as covariates when analyzing intraindividual reliability.

As in Experiment 1, we examined the effect of attribute consistency on intraindividual reliability as a function of set and consistency. We conducted a 3 (consistency: low, medium, high) × 3 (measure: accuracy, informativeness, use) × 2 (set: 1, 2) mixed ANOVA on MANE with set as a between-participants factor. Contrary to the attribute consistency hypothesis, we found a significant main effect of attribute consistency [$F(1.67, 288.60) = 3.27$, $p = .048$, $\eta_G^2 = .01$] such that medium consistency was significantly lower in error than low consistency [$t(174) = 3.24$, $p = .001$, $d = 0.24$]. There was no significant difference between low and high consistency [$t(174) = 0.55$, $p = .583$, $d = .0.04$] or between medium and high consistency [$t(174) = 1.71$, $p = .088$, $d = 0.13$]. The interaction between item set and measure was also significant [$F(1.70, 294.34) = 3.78$, $p = .025$, $\eta_G^2 < .01$] such that use and accuracy errors were greater in set 2 than in set 1, while error in informativeness judgments was greater in set 1 than in set 2 (see Figure 4).[8] No other main or interaction effects were statistically significant [Fs < 1.94, ps > .140].

[8]The analogous means and 95% CIs as a function of dependent measure are presented in the Supplementary Material.

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

Table 5. *Effect of source-reliability/information-credibility trade-offs on mean information-quality measures in Experiment 2.*

Stimuli pair and measure	Mean (*SD*)	*t*	*df*	*p*	Cohen's *d*
A5–E1					
Accuracy	M_{A5} = 50.64 (27.89)	0.69	174	.490	0.05
	M_{E1} = 52.62 (27.81)				
Use	M_{A5} = 47.98 (27.65)	0.20	174	.845	0.01
	M_{E1} = 48.47 (28.46)				
Informativeness	M_{A5} = 48.62 (28.29)	1.67	174	.096	0.13
	M_{E1} = 53.18 (28.92)				
Trustworthiness	M_{A5} = 57.04 (28.75)	3.71	174	**<.001**	0.28
	M_{E1} = 45.95 (29.09)				
A3–C1					
Accuracy	M_{A3} = 68.45 (18.07)	0.73	87	.469	0.08
	M_{C1} = 66.70 (18.07)				
Use	M_{A3} = 68.08 (18.90)	1.06	87	.292	0.11
	M_{C1} = 65.36 (18.65)				
Informativeness	M_{A3} = 69.01 (19.34)	0.27	87	.791	0.03
	M_{C1} = 69.64 (18.52)				
Trustworthiness	M_{A3} = 69.76 (20.15)	1.26	87	.210	0.13
	M_{C1} = 66.44 (18.99)				
C5–E3					
Accuracy	M_{C5} = 42.29 (25.08)	0.28	86	.781	0.03
	M_{E3} = 41.53 (24.78)				
Use	M_{C5} = 38.94 (25.92)	0.41	86	.682	0.04
	M_{E3} = 37.83 (25.81)				
Informativeness	M_{C5} = 43.40 (26.89)	0.63	86	.531	0.07
	M_{E3} = 44.94 (27.51)				
Trustworthiness	M_{C5} = 47.23 (24.21)	4.31	86	**<.001**	0.46
	M_{E3} = 35.15 (27.38)				

Note: Bold *p*-values are statistically significant.

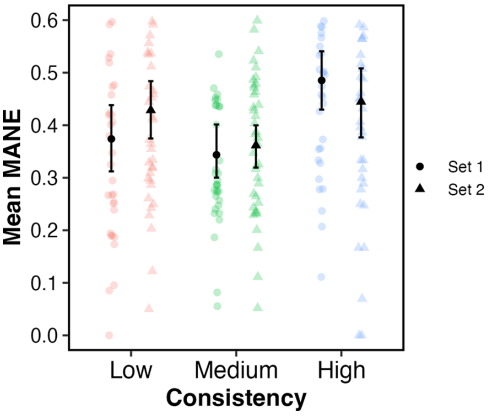

Figure 4. *Mean maximum absolute normalized error (MANE) as a function of consistency and set in Experiment 2. Error bars are bias corrected and accelerated 95% bootstrap CIs using 10,000 samples.*

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

Table 6. *Models fitted for each stimulus and predictor of use likelihood ratings in Experiment 2.*

Stimulus, model fit, and predictor	*B*	*SE*	*β*	*t*	*p*
A1 (high SR, high IC):					
$F(3, 171) = 161.5, p < .001, R^2_{Adj.} = .73$					
Accuracy	0.78	0.05	0.77	16.53	**<.001**
Informativeness	0.04	0.06	0.04	0.74	.463
Trustworthiness	0.05	0.07	0.05	0.66	.510
A5 (high SR, low IC):					
$F(3, 171) = 23.9, p < .001, R^2_{Adj.} = .28$					
Accuracy	0.35	0.07	0.34	4.78	**<.001**
Informativeness	0.31	0.07	0.31	4.82	**<.001**
Trustworthiness	0.03	0.07	0.03	0.41	.681
E1 (low SR, high IC):					
$F(3, 171) = 16.8, p < .001, R^2_{Adj.} = .21$					
Accuracy	0.29	0.07	0.29	3.96	**<.001**
Informativeness	0.24	0.07	0.24	3.29	**.001**
Trustworthiness	0.10	0.08	0.10	1.26	.210
E5 (low SR, low IC):					
$F(3, 171) = 154.9, p < .001, R^2_{Adj.} = .73$					
Accuracy	0.67	0.05	0.65	12.62	**<.001**
Informativeness	–0.18	0.08	–0.18	2.36	**.020**
Trustworthiness	0.41	0.08	0.41	5.21	**<.001**

Note: SR, source reliability; IC, information credibility. Bold *p*-values are statistically significant.

As in Experiment 1, we examined the correlations between mean MANE at different levels of consistency. As in Experiment 1, mean errors correlated positively between low and medium levels of consistency [$t(173) = 8.60, r = .55, p < .001$], but the correlations between medium and high levels of consistency [$t(173) = 1.56, r = .12, p = .120$] and low and high levels of consistency were not significant [$t(173) = 1.75, r = .13, p = .081$]. Thus, there is weaker evidence of stable individual differences in reliability in the nonexpert sample.

6.4. Predictors of likelihood of information use

We regressed participants' ratings of likelihood of information use on their ratings of perceived information accuracy, informativeness, and trustworthiness for each of the 4 items common to sets 1 and 2 (i.e., A1, A5, E1, and E5). Each of the 4 models was significant. Table 6 shows the overall model statistics for the 4 models and the statistics for each parameter in the 4 models.[9] In each model, accuracy ratings significantly predicted likelihood of information use ratings, and informativeness ratings predicted likelihood of use ratings in all cases except when both source reliability and information credibility were high. The fit of the models was much higher in the 2 reliable cases. Finally, when source reliability and information credibility were both low, trustworthiness ratings also predicted likelihood of use ratings. Thus, compared to analysts, the nonexperts were influenced by a wider range of considerations.

[9]Results were qualitatively the same when excluding potentially influential observations and participants with potentially influential observations (identified with *influence_plot()* from the *regclass* package; Petrie, 2020), see analogous results for these analyses in the Supplementary Material.

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

6.5. Reanalysis of intraindividual reliability in Mandel et al. (2023)

Given the lack of support for the attribute consistency hypothesis in these studies, we reanalyzed data from Mandel et al. (2023) using the MANE scoring rule to determine whether the attribute consistency effect would likewise be eliminated. We combined the data from Experiments 1 and 2 because the samples of each experiment were small (combined $N = 57$) and the methods were the same in all critical respects. As in the present analyses, MANE was computed for judgments for which a given Admiralty Code was presented twice. The critical codes that were subject to reliability tests were A1, A3, A5, C1, C3, C5, E1, E3, and E5. Thus, there were 3 high-consistency values (A1, C3, E5), 4 moderate-consistency values (A3, C1, C5, E3), and 2 low-consistency values (A5, E1). The MANE scores for items in each level of consistency were averaged and these mean MANE values were analyzed using one-way (consistency) repeated-measures ANOVA. The main effect of consistency was significant, $F(1.37, 69.71) = 5.35$, $p = .006$, $\eta_G^2 = .06$. Medium consistency ($M = 0.20$, $SD = 0.10$) was significantly lower in error than high consistency ($M = 0.30$, $SD = 0.24$), [$t(51) = 3.20$, $p = .002$, $d = 0.44$]. There were no significant differences between low ($M = 0.23$ $SD = 0.17$) and high consistency, [$t(51) = 1.83$, $p = .073$, $d = 0.25$] or between medium and low consistency [$t(51) = 1.30$, $p = .199$, $d = 0.18$]. Thus, as in the present experiments, MANE is lowest in the medium-consistency level rather than the high-consistency level predicted on the basis of the attribute consistency hypothesis.

The data from Mandel et al. (2023) also permitted additional tests in which the expectation of mean judgment extremity is held constant, while attribute consistency is varied. These include the comparison of high-consistency C3 ($M = .16$, $SD = .15$) with low-consistency A5 ($M = .21$, $SD = .20$) and E1 ($M = .24$, $SD = .23$). The C3-A5 comparison was not significant, $t(51) = 1.35$, $p = .183$. The C3-E1 comparison was significant, $t(51) = 2.13$, $p = .038$. These tests suggest when mean judgment extremity is controlled, there is some evidence for the attribute consistency hypothesis. However, the evidence is inconclusive as only one of the 2 tests was statistically significant and the sample size of the study is small.

7. General discussion

Expert assessments require disparate sources of information, and information naturally varies in quality for many reasons. Accordingly, information is often encoded with meta-informational attributes cueing information quality (Irwin and Mandel, 2019). Those attributes, however, may conflict with each other and if they do, the inferences that people draw from them about information quality may become less reliable, as earlier work suggests (Lichtenstein et al., 1975; Slovic, 1966).

Across 2 experiments, using a new scoring rule (MANE) that provides a more conservative test of the attribute consistency hypothesis than that conducted by Mandel et al. (2023), we observed that the intraindividual reliability of information quality judgments varied as a function of attribute consistency, although not as predicted by the attribute consistency hypothesis. In our studies and in a reanalysis of data from Mandel et al. (2023), the lowest level of error was invariably associated with the medium level of attribute consistency. The findings, therefore, suggest that the earlier report of evidence in support of the attribute consistency hypothesis was due to the selection of a scoring rule that had a statistical expectation of lower error for high-consistency attribute pairs.

Although there was some evidence to support the attribute consistency hypothesis from pairwise tests that held expected mean judgment extremity constant, on balance, the evidence for this hypothesis is inconclusive. A more definitive test of the hypothesis would require a large sample evaluating stimuli that control for expected average judgment extremity, preferably across a range of dependent measures of information quality. As well, future studies could compare the reliability and perceived usefulness of the Admiralty Code to alternative methods that encode qualitative meaning at the 'cell' level. For example, Icard (2023, 2024) proposes a 3 (Honesty of Source: Honest vs. Imprecise vs. Dishonest) × 3 (Truth of Content: True vs. Indeterminate vs. False) matrix wherein the 9 categorizations (i.e., 'cells') are qualitatively well described. The effectiveness of this and alternative proposals could inform next-generation models for information evaluation in intelligence.

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

The present findings extend previous work in other respects. We observed similar consistency-reliability results among professional analysts (Experiment 1) and nonexperts (Experiment 2), suggesting that both groups judge similarly despite differing in formal intelligence training, including with respect to the Admiralty Code. As well, reliability estimates tended to correlate across levels of attribute consistency in both experiments. These findings suggest that there are stable individual differences in the ability to fuse information from multiple attributes. The CRT task did not predict reliability in Experiment 2. However, future work could examine the impact of other potentially relevant individual differences (e.g., need for cognition, conscientiousness) as well as the promise of more structured judgment elicitation methods on the assessment of information quality. For instance, Irwin and Mandel (2019) outlined a list of questions that analysts might routinely ask themselves when conducting information evaluation and Mandel and Irwin (2024) suggest that various judgment tasks in intelligence such as information evaluation could be crowdsourced and subsequently aggregated to improve accuracy and reliability.

Our findings also extend the earlier work by testing how the meta-informational attributes of source reliability and information credibility might be differentially weighted when considering each information-quality measure. Only trustworthiness ratings differentially weighted these attributes, giving more weight to source reliability. The lack of evidence for unequal weighting of source reliability and information credibility for other information-quality measures raises prescriptive questions. Given that information credibility directly pertains to the information quality, whereas source reliability is an indirect indicator, one might have expected the former to be weighted more heavily, and some might argue that it should. Perhaps in a more naturalistic setting, reasoning for differential weighting between source reliability and information accuracy might be salient in response to key contextual information at that time. For example, Samet (1975) presented participants with plausible raw intelligence and found that information credibility (or 'accuracy' in their version) was weighted more heavily than source reliability when judging event likelihood. Samet's (1975) inclusion of plausible raw intelligence could have introduced additional contextual interactions not present in the current work. The current findings suggest that 'all else being equal', source reliability and information credibility are also treated 'equal' when judging certain aspects of information quality and that this is true of experts and nonexperts alike.

A final question we addressed was whether accuracy, informativeness, and/or trustworthiness predicted intent to use that information (i.e., the rated probability of using that information) given that, in practice, intent to use the information is arguably most consequential in terms of information quality indices. That is, if an analyst decides not to use the information, it cannot influence the analyst's substantive assessment. Across both experiments, accuracy ratings significantly predicted likelihood of use regardless of the levels of source reliability or information credibility (low vs. high). The other indices of information quality did not consistently predict likelihood of use ratings across experiments—although in Experiment 2, there was some evidence that informativeness also predicted likelihood of use. While information quality indices might be similarly affected by source reliability and information credibility, these results suggest that accuracy, in particular, is predictive of one's intention to use the information. Thus, accuracy appears to be an essential factor in opting whether to use information and earlier works focusing on eliciting accuracy judgments likely indirectly capture the likelihood of using information. These findings suggest that it may therefore be beneficial to focus on methods for reliably assessing information accuracy, and that intelligence doctrine might do well to be explicit about the connection between judged information accuracy and decision-making regarding information use or disuse.

7.1. Limitations

In the current work, we examined a single information-evaluation system and, for experimental control, did not incorporate contextual information along with the meta-informational attributes that would typically accompany the alphanumeric coding in day-to-day administration. This could potentially

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

affect the generality of the present findings. We are also aware that being an experienced intelligence analyst does not necessarily imply one is also an expert in applying the Admiralty Code to effectively manage complex situations. Indeed, this could explain why performance was similar across the 2 samples (of experts and novices). Future work could focus on isolating a particularly expert sample (e.g., human intelligence analysts who interact with such meta-information on a regular basis), though collecting an adequately powered sample of such individuals is nontrivial.

Despite the points we raise above, the generality of the current findings is expected to be higher than in previous relevant work for 3 main reasons. First, the task that participants completed was based on a real-world information-evaluation system (the Admiralty Code) which has been in use for the better half of a century and across many domains including intelligence and national security. Second, Experiment 2 was specifically designed as a generalizability test (i.e., from an expert sample to a nonexpert sample) comprising a large sample collected from a general online pool and the results across experiments were similar in their qualitative results. Third, where possible, samples were diverse in their demographics including sex, age, educational attainment, and expertise (i.e., professional analysts in Experiment 1, general online sample in Experiment 2).

Data availability statement. The data, analyses code, and supplementary materials for Experiments 1 and 2 are all available at the following external repository: osf.io/nzvux/files/osfstorage.

Acknowledgments. We thank Daniel Irwin and Irina Levit for their research assistance. We also thank Jon Baron and 2 anonymous reviewers for their feedback on earlier drafts of this article. This research contributes to the North Atlantic Treaty Organization System Analysis and Studies Panel Research Task Group on Anticipatory Intelligence for Superior Decision-Making (SAS-189), co-chaired by the last author and including national representation from the second and third authors.

Funding statement. This research was funded by Canadian Safety and Security Program project CSSP-2018-TI-2394 and the Department of National Defence Accelerate Command, Control and Intelligence Project Activity AC2I-031 (Enhanced Indications and Warning) under the direction of the last author.

Competing interest. The authors declare none.

References

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379–384. https://doi.org/10.3758/BF03192707

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. https://doi.org/10.1037/e463552004-001

Batini, C., & Scannapieco, M. (2016). *Data and information quality*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24106-7

Bertino, E., Lim, H. S. (2010). Assuring data trustworthiness: Concepts and research challenges. In: Jonker, W., Petković, M. (Eds.) *Secure data management 2010* (pp. 1–12), Springer. https://doi.org/10.1007/978-3-642-15546-8_1

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior*, *26*, 353–364. https://doi.org/10.1023/A:1015380522722

Carter, D. L. (2009). *Law enforcement intelligence: A guide for state, local, and tribal law enforcement agencies, second edition*. Office of Community Oriented Policing Services, U.S. Department of Justice.

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

Dai, C., Lin, D., Bertino, E., Kantarcioglu, M. (2008). An approach to evaluate data trustworthiness based on data provenance. In: Jonker, W., Petković, M. (Eds.) *Secure data management 2008* (pp. 82–98), Springer. https://doi.org/10.1007/978-3-540-85259-9_6

Eppler, M. J., & Wittig, D. (2000). Conceptualizing information quality: A review of information quality frameworks from the last ten years. *Proceedings of the 2000 Conference on Information Quality*, *20*, 83–91. http://mitiq.mit.edu/iciq/Documents/IQ%20Conference%202000/Papers/ConceptIQaReviewofIQFramework.pdf

Erceg, N., & Bubić, A. (2017). One test, five scoring procedures: Different ways of approaching the cognitive reflection test. *Journal of Cognitive Psychology*, *29*(3), 381–392. https://doi.org/10.1080/20445911.2016.1278004

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed). Sage Publications. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 25–42. https://doi.org/10.1257/089533005775196732

Gil, Y. & Artz, D. (2007). Towards content trust of web resources. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, *5*(4), 227–239. https://doi.org/10.1016/j.websem.2007.09.005

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. https://mail.ijlter.org/index.php/ijlter/article/view/494/234

Hoffman, P. J., Slovic, P., & Rorer, L. G. (1968). An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment. *Psychological bulletin*, *69*(5), 338. https://doi.org/10.1037/h0025665

Huang, K.-T., Lee, Y.W., & Wang, R.Y. (1999). *Quality information and knowledge*. Prentice Hall.

Icard, B. (2023). Facts versus interpretations in intelligence: A descriptive taxonomy for information evaluation. *Intellectica, Cognition and Intelligence*, *78*(1), 89–105. https://doi.org/10.31234/osf.io/q8avs

Icard, B. (2024). A dynamic logic for information evaluation in intelligence. SSRN Preprint. https://doi.org/10.2139/ssrn.4627928

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. (2023). Communicating uncertainty in national security intelligence: Expert and non-expert interpretations of and preferences for verbal and numeric formats. *Risk Analysis*, *43*, 943–957.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(6), 1797. https://doi.org/10.1037/0096-1523.26.6.1797

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. https://doi.org/10.1017/S1930297500008159

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, *43*(24), 2539–2558. https://doi.org/10.1016/S0042-6989(03)00458-9

Lambert, T. A., & Peytcheva, M. (2020). When is the averaging effect present in auditor judgments? *Contemporary Accounting Research*, *37*(1), 277–296. https://doi.org/10.1111/1911-3846.12512

Landy, M. S., Banks, M. S., & Knill, D. C. (2011). Ideal-observer models of cue integration. In J. Trommershauser, K. , Kording, & M. Landy (Eds.), *Sensory cue integration* (pp. 5–29). Oxford University Press.

Lawrence MA (2016). ez: Easy analysis and visualization of factorial experiments. R package version 4.4-0. https://CRAN.R-project.org/package=ez.

Leitzke, B. T., & Pollak, S. D. (2016). Developmental changes in the primacy of facial cues for emotion recognition. *Developmental Psychology*, *52*(4), 572. https://doi.org/10.1037/a0040067

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

Lichtenstein, S., Earle, T. C., & Slovic, P. (1975). Cue utilization in a numerical prediction task. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(1), 77–85. https://doi.org/10.1037/0096-1523.1.1.77

Lusk, C.M. (1993). Assessing components of judgments in an operational setting. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making*. (pp. 309–322). Springer. https://doi.org/10.1007/978-1-4757-6846-6_20

Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and framework for data and information quality research. *Journal of Data and Information Quality*, *1*(1), 1–22. https://doi.org/10.1145/1515693.1516680

Mandel, D. R. (2020). The occasional maverick of analytic tradecraft. *Intelligence and National Security*, *35*(3), 438–443. https://doi.org/10.1080/02684527.2020.1723830

Mandel, D. R., & Irwin, D. (2024). Beyond bias minimization: Improving intelligence with optimization and human augmentation. *International Journal of Intelligence and Counterintelligence*, *37*(2), 649–665. https://doi.org/10.1080/08850607.2023.2253120

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

Marcoci, A., Burgman, M., Kruger, A., Silver, E., McBride, M., Thorn, F. S., Fraser, H., Wintle, B. C., Fidler, F., & Vercammen, A. (2019a). Better together: Reliable application of the post-9/11 and post-Iraq US intelligence tradecraft standards requires collective analysis. *Frontiers in Psychology*, *9*, 2634. https://doi.org/10.3389/fpsyg.2018.02634

Marcoci, A., Vercammen, A., & Burgman, M. (2019b). ODNI as an analytic ombudsman: Is Intelligence Community Directive 203 up to the task? *Intelligence and National Security*, *34*(2), 205–224. https://doi.org/10.1080/02684527.2018.1546265

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

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. https://doi.org/10.21236/ADA060321

Miyazaki, A. D., Grewal, D., & Goodstein, R. C. (2005). The effect of multiple extrinsic cues on quality perceptions: A matter of consistency. *Journal of Consumer Research*, *32*(1), 146–153. https://doi.org/10.1086/429606

Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Current Biology*, *18*(9), 689–693. https://doi.org/10.1016/j.cub.2008.04.021

https://doi.org/10.1017/jdm.2025.10007 Published online by Cambridge University Press

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.

Nurse, J. R. C., Rahman, S. S., Creese, S., Goldsmith, M., & Lamberts, K. (2011). Information quality and trustworthiness: A topical state-of-the-art review. In: *The international conference on computer applications and network security (ICCANS), KAR id: 67536*.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, *48*, 341–348. https://doi.org/10.3758/s13428-015-0576-1

Petrie A (2020). *Regclass: Tools for an introductory class in regression and modeling.* R Package Version 1.6. https://CRAN.R-project.org/package=regclass

Piazza, J., & Sousa, P. (2014). Religiosity, political orientation, and consequentialist moral thinking. *Social Psychological and Personality Science*, *5*(3), 334–342. https://doi.org/10.1177/1948550613492826

R Core Team (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

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

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423. https://doi.org/10.1037/a0025391

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

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, *58*(12), 1720–1733. https://doi.org/10.1002/asi.20652

United States Department of the Army . (1951). *Field manual FM 30-5, combat intelligence.* Washington, DC.

Wall, M., Liefeld, J., & Heslop, L. A. (1991). Impact of country-of-origin cues on consumer judgments in multi-cue situations: A covariance analysis. *Journal of the Academy of Marketing Science*, *19*, 105–113. https://doi.org/10.1007/BF02726002

Zaki, J. (2013). Cue integration: A common framework for social cognition and physical perception. *Perspectives on Psychological Science*, *8*(3), 296–312. https://doi.org/10.1177/1745691613475454

Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News*, *2*(3), 7–10. https://CRAN.R-project.org/doc/Rnews/

Cite this article: Kelly, M. O., Budescu, D. V., Dhami, M., and Mandel, D. R. (2025). The effect of source reliability and information credibility on judgments of information quality in intelligence analysis. *Judgment and Decision Making*, e36. https://doi.org/10.1017/jdm.2025.10007