Judgment and Decision Making, Vol. 13, No. 6, November 2018, pp. 607–621

Boosting intelligence analysts' judgment accuracy: What works, what fails?

David R. Mandel* Christopher W. Karvetski† Mandeep K. Dhami‡

Abstract

A routine part of intelligence analysis is judging the probability of alternative hypotheses given available evidence. Intelligence organizations advise analysts to use intelligence-tradecraft methods such as Analysis of Competing Hypotheses (ACH) to improve judgment, but such methods have not been rigorously tested. We compared the evidence evaluation and judgment accuracy of a group of intelligence analysts who were recently trained in ACH and then used it on a probability judgment task to another group of analysts from the same cohort that were neither trained in ACH nor asked to use any specific method. Although the ACH group assessed information usefulness better than the control group, the control group was a little more accurate (and coherent) than the ACH group. Both groups, however, exhibited suboptimal judgment and were susceptible to unpacking effects. Although ACH failed to improve accuracy, we found that recalibration and aggregation methods substantially improved accuracy. Specifically, mean absolute error (MAE) in analysts' probability judgments decreased by 61% after first coherentizing their judgments (a process that ensures judgments respect the unitarity axiom) and then aggregating their judgments. The findings cast doubt on the efficacy of ACH, and show the promise of statistical methods for boosting judgment quality in intelligence and other organizations that routinely produce expert judgments.

Keywords: probability judgment, accuracy, coherence, intelligence analysis, recalibration, aggregation, coherentization

1 Introduction

Intelligence organizations routinely call upon their analysts to make probability judgments and test hypotheses under conditions of uncertainty. These expert judgments can inform important policy decisions concerning national and international security. Traditionally, analysts have been expected to accumulate domain expertise and apply this along with critical thinking skills to arrive at timely and accurate assessments for decision-makers. In the US, developers of analytic tradecraft (i.e., the methods developed within the intelligence community to support its analytic functions) such as Richards Heuer Jr. and Jack Davis introduced so-called "structured analytic techniques" (SATs) to support the analyst in the assessment process, but these methods were largely optional tricks-of-the-trade. That state of affairs changed following two notable geopolitical events (i.e., the September 11, 2001, terrorist attacks by Al Qaeda and the 2003 invasion of Iraq) that were attributed in part to striking intelligence failures. These events prompted reviews of the intelligence community with ensuing organizational reforms that, among other things, aimed at debiasing intelligence analysts' judgments (Belton & Dhami, in press). In the US, the Intelligence Reform and Terrorism Prevention Act of 2004 mandated the use of SATs in intelligence production and SATs became a staple topic in most analytic training programs (Chang, Berdini, Mandel & Tetlock, 2018; Coulthart, 2017; Marchio, 2014). Much the same set of organizational reforms was enacted in other Western countries such as the UK (e.g., Butler, 2004).

Although the number of SATs has skyrocketed over the last decade (Dhami, Belton & Careless, 2016; Heuer & Pherson, 2014), as others have lamented in recent years (Chang et al., 2018; Dhami, Mandel, Mellers & Tetlock, 2015; National Research Council, 2011; Pool, 2010), there has been little effort to test their effectiveness. Instead, most SATs have been adopted on the basis of their perceived face validity with the belief that, although imperfect, they must be better than nothing. At the same time, the intelligence community has rarely considered using post-analytic techniques to improve judgment (Mandel & Tetlock, 2018). For instance, Mandel and Barnes (2014) showed that intelligence analysts' strategic forecasts were underconfident, but that much of this bias could be eliminated by recalibrating their judgments to make them more extreme (also see Baron, Mellers, Tetlock, Stone & Ungar, 2014; Turner, Steyvers, Merkle, Budescu

This research was supported by Department of National Defence project #05da, Canadian Safety and Security Program projects #2016–TI–2224 and #2018–TI–2394, and HM Government. This work contributes to the NATO System Analysis and Studies Panel Research Task Group on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making (SAS-114). We thank Jon Baron, Jonathan Nelson, and two anonymous reviewers for helpful feedback on this research.

Copyright: © 2018. The authors license this article under the terms of the Creative Commons Attribution 3.0 License.

*Intelligence, Influence and Collaboration Section, Toronto Research Centre, Defence Research and Development Canada. Email: drmandel66@gmail.com.

†BlackSwan Technologies Ltd.

‡Department of Psychology, Middlesex University.

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

& Wallsten, 2014). Similarly, the accuracy of probability judgments can be improved post-judgment by recalibration so judgments respect one or more coherence principles, such as the axioms of probability calculus — a statistical process called *coherentization* (Karvetski, Olson, Mandel & Twardy, 2013). Karvetski et al. further observed that weighting individuals' contributions to aggregated judgments also improved accuracy above the gains achieved using an unweighted arithmetic average. In the present research, we examine the accuracy of intelligence analysts' probability judgments in an experimental task. We examine the effectiveness of ACH as well as recalibration and aggregation methods with the aim of addressing the prescriptive question: what works — and what fails — to improve judgment accuracy?

1.1 The Analysis of Competing Hypotheses Technique

The Analysis of Competing Hypotheses (ACH; Heuer, 1999; Heuer & Pherson, 2014) is one of the most widely known SATs, and one of only nine that is listed in the US Government's (2009) *Tradecraft Primer* (also see UK Ministry of Defence, 2013). The US Government describes ACH as a diagnostic technique whose main function is to externalize analytic hypotheses and evidence. It further claims that ACH helps analysts overcome common cognitive biases, such as primacy effects, confirmation bias, and other forms of premature cognitive closure that can undermine the accuracy of forecasts or other probabilistic assessments. The US Government also asserts that ACH "has proved to be a highly effective technique when there is a large amount of data to absorb and evaluate" (2009, p. 14), yet it does not cite any evidence to support that claim. The UK handbook conveys comparable exuberance for ACH, noting, "The approach is designed to help analysts consider all the evidence in the light of all the hypotheses *as objectively as possible*" (UK Ministry of Defence, 2013, p. 14, italics added).

ACH includes several steps, but the core of the tradecraft method involves generating a matrix in which mutually exclusive and (preferably) collectively exhaustive (MECE) hypotheses are listed in columns and pieces of relevant evidence are listed in rows. The analyst then assesses the consistency of each piece of evidence with each hypothesis starting on the first row and moving across the columns. For each cell, the analyst rates evidence-hypothesis consistency on a 5-point ordinal scale (i.e., −2 = highly inconsistent, −1 = inconsistent, 0 =neutral or not applicable, 1 = consistent, 2 = highly consistent. However, only the negative scores (−1 and −2) are tallied for each hypothesis. For instance, if there were five pieces of information (i.e., five rows) and Hypothesis A had ratings {2, 2, 2, 2, −2} and Hypothesis B had ratings {0, 0, 1, 0, −1}, Hypothesis B with an inconsistency score of −1 would be rated as more likely to be true than Hypothesis A with a score of −2. In other words, ACH requires that analysts disregard evidential *support* for hypotheses in the information integration process. This feature of the method may have been motivated by a misapplication of Popper's (1959) ideas about the merits of falsification as a strategy for scientific discovery. Popper's claim that hypotheses could only be falsified but never proven pertained to universal hypotheses such as "all swans are white" because a single non-white swan is sufficient to disprove the claim. Most hypotheses of interest in intelligence, however, are not universal but rather deal with events in a particular context (e.g., Iran is developing a nuclear weapon), and few could be falsified outright by a single disconfirming piece of evidence (Mandel, in press).

ACH also includes a subsequent evidential editing phase: once the matrix is populated with consistency ratings, the analyst is encouraged to remove evidence that does not appear to differentiate between the alternative hypotheses. However, there is virtually no guidance on how such assessments of information usefulness should be conducted. For instance, the US Government merely instructs, "The 'diagnostic value' of the evidence will emerge as analysts determine whether a piece of evidence is found to be consistent with only one hypothesis, or could support more than one or indeed all hypotheses. In the latter case, the evidence can be judged as unimportant to determining which hypothesis is more likely correct" (2009, p. 15). The UK handbook is more precise, stating "For each hypothesis ask the following question: 'If this hypothesis were true, how likely would the evidence be?'" (UK Ministry of Defence, 2013, p. 15; see also Heuer, 1999). Yet, it vaguely advises analysts to "pay most attention to the most diagnostic evidence — i.e., that which is highly consistent with some hypotheses and inconsistent with others" (p. 17). If evidence is subsequently disregarded, then analysts are expected to recalculate the sum of the negative (inconsistency) ratings. These scores are then meant to reflect the rank ordering of hypotheses by subjective probability, with the hypothesis receiving the smallest inconsistency score being judged as most likely to be true in the set of hypotheses being tested.

ACH is not a normative method for probabilistic belief revision or hypothesis testing, but it has become an institutionalized heuristic that intelligence organizations have deemed to be effective without compelling reasons or evidence (for additional critiques, see Chang et al., 2018; Jones, 2018; Karvetski, Olson, Gantz & Cross, 2013; Pope & Jøsang, 2005; Mandel, in press; Mandel & Tetlock, 2018). As already noted, ACH disregards useful information about evidential support for hypotheses and it requires analysts to self-assess information utility without providing a clear definition of utility, let alone a computational method for estimating such utility. Perhaps even more fundamental is the omission of a clear definition of *consistency*, which could signify a range of meanings, such as the probability of the

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

evidence given the hypothesis, the probability of the hypothesis given the evidence, the plausibility or the necessity of one given the other, or simply a subjective sense of the representativeness of one to the other — namely, the representativeness heuristic (Kahneman & Tversky, 1972). In addition, ACH does nothing to ensure that analysts consider prior probabilities or objective base rates when revising their beliefs about hypotheses in light of new evidence. In sum, there are many reasons to be skeptical about the effectiveness of ACH.

Unfortunately, there is little scientific research on ACH, and what exists must be interpreted cautiously for several reasons, such as small sample sizes (e.g., Convertino, Billman, Pirolli, Massar & Shrager, 2008; Lehner, Adelman, Cheikes & Brown, 2008; Kretz, Simpson & Graham, 2012), lack of control groups (Convertino et al., 2008) or appropriate control groups (Kretz et al., 2012). Moreover, virtually all published studies have omitted critical, quantitative measures of judgment accuracy, focusing instead on distal considerations such as whether ACH reduces (the highly equivocal notion of) "confirmation bias" (Nickerson, 1998). Yet, despite the many serious limitations of research on ACH (and SATs, more generally), the intelligence studies literature has shown little concern regarding the lack of adequate research to support the widespread use of SATs, including ACH. Rather, a recent review article concluded that ACH was "found to be effective and had a highly credible evidence base. . . " (Coulthart, 2017, p. 377). This conclusion is unwarranted not only because of the methodological weaknesses noted earlier, but also because the extant findings are at best equivocal. For instance, whereas Lehner et al. (2008) find that ACH reduced confirmation bias in non-analysts, it had no effect on analysts.

1.2 The present research

A central aim of our research was to examine how the accuracy and logical coherence of intelligence analysts' judgments about the probability of alternative (MECE) hypotheses depended on whether or not analysts were trained in and used ACH on the experimental task. In addition to this SAT, we also explored the value of statistical post-judgment methods for improving expert judgment, such as recalibrating experts' probabilities in ways that remedy certain coherence violations (i.e., non-unitarity and/or non-additivity), and by aggregating experts' judgments using varying group sizes and weighting methods.

We tested the effectiveness of ACH by randomly assigning intelligence analysts from the same population to experimental conditions that either used ACH or did not. One group of analysts was recently trained to use ACH as part of their organization's training and they were required to use ACH on the experimental task. The other group of analysts was drawn from the same analytic cohort (i.e., same organization and taking the same training course) but they were not instructed to use ACH (or any SAT for that matter) and were not exposed to ACH training until after the experiment was completed. The task, which involved a hypothetical scenario, required analysts to assess the probabilities of four MECE hypotheses that corresponded to four tribes in a region of interest. Participants were asked to assess the probabilities that a detained individual (i.e., the target) from the local population belongs to each of the four tribes. Participants were given the tribe base-rates and diagnostic conditional probabilities for 12 evidential cues (e.g., "speaks Zimban"), along with the cue values (6 present and 6 absent) for the target. Furthermore, two tribes (Bango and Dengo, hereafter B and D) were grouped as friendly (F), whereas the other two (Acanda and Conda, hereafter A and C) were grouped as hostile (H).

If ACH proponents' claims about the technique's effectiveness are warranted, we should find greater probabilistic judgment accuracy in the ACH condition than in the control condition. As noted earlier, to the best of our knowledge, there is no clear evidence to support the claim that ACH improves probabilistic judgment accuracy. Indeed, one non-peer-reviewed study that compared various degrees of ACH support (e.g., ACH on its own or with additional training) across experimental groups found that accuracy was best among those participants in the no-ACH control group (Wheaton, 2014). However, insufficient information was provided to interpret these results with any confidence.

In addition, if proponents' claims about the effectiveness of ACH in promoting soundness of judgment are true, we might expect to find that analysts recently trained in and aided by ACH produce probability judgments that are more coherent than those unaided by ACH. We tested this proposition by examining the degree to which probability judgments in both groups respect the axioms of unitarity and additivity. To do so, we drew on predictions of support theory, a non-extensional descriptive account of subjective probability which posits that one's probability judgments are a function of his or her assessments of evidential support for a focal hypothesis and its alternative (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994). Support theory predicts an *unpacking effect*, in which the sum of the probabilities assigned to a MECE partition with more than two subsets of an event, x, exceeds $P(x)$. Unpacking effects have been shown in several studies (Ayton, 1997; Fox, Rogers & Tversky, 1996; Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994). For instance, in two experiments with undergraduate participants, Mandel (2005) found that the mean *unpacking factor* — namely, the ratio of the sum of unpacked probability estimates to the packed estimate — was 2.4 comparing forecasts of terrorism (i.e., the packed forecast) to forecasts of terrorism unpacked into acts committed by Al Qaeda or by operatives unaffiliated with Al Qaeda. No research has yet examined whether intelligence analysts'

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

probability judgments are susceptible to the unpacking effect. In the present research, the unpacking effect would be observed if $P(A)+P(C) > P(H)$ and/or $P(B)+P(D) > P(F)$. According to the additivity axiom, these inequalities should be equalities, given that $A \cap C = \varnothing$ and $A \cup C \equiv H$; likewise $B \cap D = \varnothing$ and $B \cup D \equiv F$.

Extending our investigation into the coherence of analysts' probability judgments, we further tested whether analysts' judgments respect the unitarity axiom, which states that the probabilities assigned to a MECE set of hypotheses should sum to unity. Support theory predicts that partitions of a sample space into more than two subsets will yield an unpacking effect. Thus, in the present research, support theory predicts $P(A)+P(B)+P(C)+P(D) > 1.0$, in violation of the unitarity axiom, which requires that these probabilities sum to unity. The unitarity axiom also requires that the binary complements $P(H)$ and $P(F)$ sum to 1.0, although support theory predicts agreement with the axiom in the case of binary complements. Some studies find agreement with support theory's prediction for binary complements (e.g., Dhami & Mandel, 2013; Rottenstreich & Tversky, 1997; Wallsten, Budescu & Zwick, 1993), whereas others find that the sum of the probabilities people assign to binary complements is less than unity (e.g., Baratgin & Noveck, 2000; Macchi, Osherson & Krantz, 1999; Mandel, 2008; Sloman, Rottenstreich, Wisniewski, Hadjichristidis & Fox, 2004). Consistent with the latter studies, Mandel (2015b) found that intelligence analysts who were given a series of binary classification tasks to complete provided total probabilities for binary complements that fell significantly short of unity, although analysts' performance was improved through training in Bayesian reasoning using natural sampling trees. In the present research, we tested whether ACH would have a beneficial effect on the degree to which analysts' posterior probability judgments respected the unitarity axiom.

Our investigation into the coherence of analysts' probability judgments was also motivated by the aim of testing the value of statistical, post-judgment methods for improving judgment accuracy. As noted earlier, recent research shows that coherentizing probability judgments so that they respect axioms of probability calculus such as additivity and unitarity can significantly improve judgment accuracy (Karvetski, Olson, Mandel et al., 2013). Moreover, individual differences in the coherence of individuals' judgments can be exploited as a basis for performance weighting contributions to aggregated estimates, making the "crowds wiser" than they would tend to be if each member's contribution had equal weight (Osherson & Vardi, 2006; Predd, Osherson, Kulkarni & Poor, 2008; Tsai & Kirlik, 2012; Wang, Kulkarni, Poor & Osherson, 2011). Karvetski, Olson, Mandel et al. (2013) found that the accuracy of probability judgments about the truth of answers to general knowledge questions was improved through coherentizing the judgments, and a further substantial improvement was achieved by coherence weighting the coherentized judgments. In the present research, we examined how effective coherentization and coherence weighting are for improving the accuracy of intelligence analysts' probability judgments. We compared coherentized judgments to raw probability judgments generated with or without the use of ACH. We also compared coherence-weighted aggregate estimates to an equal-weight linear opinion pool (LINOP), which is the arithmetic average across judges (Clemen & Winkler, 1999). Our interest in this issue was two-fold: First, we aimed to assess the external validity of earlier findings in this nascent area of research on coherentization and coherence-weighted aggregation. Second, we aimed to test whether these post-judgment methods hold promise for organizations, such as intelligence agencies, that generate expert judgment as a product or service.

A further aim of this research anticipated both a possible benefit and a possible drawback of ACH. We hypothesized that ACH will not foster greater accuracy in probability judgment because, as we noted earlier, there are processes in the technique, such as disregarding evidential support in information integration, that are normatively indefensible. However, ACH does require analysts to evaluate each piece of information in relation to each hypothesis on the same criterion (consistency). We hypothesized that this might improve analysts' abilities to extract the usefulness of the evidence. Accordingly, we asked analysts to rate the information usefulness of each of the evidential cues presented and we examined how well these ratings correlated, on average, with the probability gain of the cue, a measure of the extent to which knowledge of the cue value is likely to improve classification accuracy (Baron, 1981, cited in Baron, 1985; Nelson, 2005).

A related aim of ours was to examine whether analysts who display stronger correlations with sampling norms also show better probability judgment accuracy, and whether this "meta-relationship" might differ between ACH and control groups. For instance, ACH proponents might be willing to wager that analysts who use ACH are more likely to reliably encode the information value *and* to use that information to their advantage by making more accurate judgments.

2 Method

2.1 Participants

Fifty UK intelligence analysts participated in the experiment during regular working hours and did not receive additional compensation for their participation. All participants were pre-registered for intelligence training and were asked by the trainers to participate in the experiment. Mean age was 27.79 years (SD = 5.03) and mean length of experience working as an analyst was 14.08 months (SD = 29.50). Out of 44 participants who indicated their sex, 25 (57%) were male.

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

TABLE 1: Informational features of experimental task. Values represent cue likelihoods.

	Tribe (base rate)				
Evidential cues	Acanda (.05)	Bango (.20)	Conda (.30)	Dengo (.45)	Feature Present in Target
Under 40 years	.10	.10	.90	.90	Yes
Use social media	.75	.50	.25	.50	Yes
Speak Zebin	.50	.75	.50	.25	Yes
Employed	.25	.25	.10	.10	Yes
Practice religion	.90	.90	.10	.10	No
From large family	.25	.50	.75	.50	No
Educated to age 16	.50	.25	.50	.75	No
Have high-SES	.75	.75	.90	.90	No
Speak Zimban	.75	.25	.75	.25	Yes
Have political affiliation	.75	.25	.75	.25	No
Wear traditional clothing	.75	.50	.60	.40	Yes
Fair coloured skin	.25	.50	.40	.60	No

2.2 Design and procedure

Participants were randomly assigned in balanced numbers to one of two conditions of the tradecraft factor: the ACH (i.e., tradecraft) condition or the no-ACH (i.e., no tradecraft) control condition. In the ACH condition, participants completed their scheduled ACH training, which was based on Heuer and Pherson (2014) and related material from Pherson Associates, LLC. Participants in the control condition received ACH training after the experiment. Participants completed a paper and pencil questionnaire and were subsequently debriefed in small group sessions within the organization in which they worked. However, participants worked individually on the task. Participants in the ACH condition were instructed to approach the judgment task using the eight steps of the ACH method, whereas participants in the control condition were free to use whatever approach they favored. The experiment received ethical approved from the institutional review board of Middlesex University.

2.3 Materials

Participants read about a fictitious case in which they were required to assess the tribe membership of a randomly selected person from a region called Zuma.[1] They read that there were four tribes (*A*-*D*) that constituted 5%, 20%, 30%, and 45% of Zuma, respectively. Each tribe was then described in terms of 12 probabilistic cue attributes. For instance, for the Acanda tribe (i.e., Tribe *A*) the description read:

> Acanda: 10% of the tribe is under 40 years of age, 75% use social media, 50% speak Zebin (one of two languages spoken in Zuma), 25% are employed, 90% practice a religion, 25% come from a large family (i.e., more than 4 children), 50% have been educated up to the age of 16, 75% have a reasonably high socio-economic status relative to the general population, 75% speak Zimban (one of two languages spoken in Zuma), 75% have a political affiliation, 75% wear traditional clothing, and 25% have fair coloured skin.

Next, the target's cue attributes were described as follows:

> The target is under 40 years of age, uses social media, speaks Zebin, is employed, does not practice a religion, does not come from a large family, does not have education up to age 16, does not have a reasonably high socio-economic status, speaks Zimban, is not politically affiliated, wears traditional clothing, and does not have fair coloured skin.

Thus, the target had positive values for half of the cues and negative values for the other half. Furthermore, analysts were told to assume that the target's answers were truthful (due to the administration of a truth serum) in order to ameliorate any possible effects of participants perceiving the information as unreliable or deceptive. Table 1 summarizes the informational features of the task.

In the ACH condition, participants were asked to complete the eight steps of the ACH method (see supplementary materials for full instructions), which included: (a) identifying all possible hypotheses, (b) listing significant evidence

[1]Full instructions for ACH and control conditions are available as supplements.

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

that is relevant for evaluating the hypotheses, (c) creating a matrix with all the hypotheses as columns and all items of relevant information as rows and then rating the consistency of each piece of evidence with each hypothesis, (d) revising the matrix after omitting non-diagnostic evidence, (e) calculating the inconsistency scores by taking the sum of the inconsistent values and using that to draw tentative conclusions about the relative likelihood of the hypotheses, (f) analyzing the sensitivity of conclusions to a change in the interpretation of a few critical items of relevant information, (g) reporting conclusions, and (h) identifying indicators for future observation.

By comparison, in the control condition, participants were asked to "consider the relative likelihood of all of the hypotheses, state which items of information were the most diagnostic, and how compelling a case they make in identifying the most likely hypothesis, and also say why alternative hypotheses were rejected." They were provided with two pages of blank paper on which to respond (none asked for more paper).

All participants completed the same final page of the questionnaire. The first four questions prompted analysts for the probability that the target belonged to each of the four tribes (*A*-*D*). Next, they were asked for the probability that the target was friendly and also for the probability that the target was hostile. Probability judgments were made on a 101-point scale that shows numeric probabilities starting at 0 and continuing at every 5% increment up to 100. The instructions noted that 0% meant "impossible" and 100% meant "absolutely certain." Next participants rated on an unnumbered 11-point scale, ranging from *not at all* to *completely*, how useful each of the 12 cues was in assessing which the target's tribe membership. For the purpose of statistical analysis, these ratings were entered as values ranging from 1 to 11. We examined analysts' responses to the scale measures of probability and information usefulness.

2.4 Coherentization and coherence weighting

As described previously, more often than not, individuals produce probability estimates that are incoherent and violate probability axioms, and there is evidence that more coherent estimates are associated with more accurate estimates (Mellers, Baker, Chen, Mandel & Tetlock, 2017). Given a set or vector of elicited probabilities that is incoherent, the *coherent approximation principle* (CAP; Osherson & Vardi, 2006; Predd et al., 2008) was proposed to obtain a coherent set of probabilities that is minimally different in terms of Euclidean distance from the elicited probabilities with the goal of improving accuracy. This "closest" set of coherent probabilities is found by projecting the incoherent probabilities onto the coherent space of probabilities. An incoherence metric can then be defined as the Euclidean distance from an incoherent set of probabilities to the closest coherent set of probabilities. For example, if an analyst in the present research provided probability judgments of .2, .3, .4, and .3 for the four MECE hypotheses *A*-*D*, respectively, these estimates are incoherent because they sum to a value greater than 1 and thus violate the unitarity constraint. Using the CAP and (if needed) quadratic programming (see Karvetski, Olson, Mandel et al., 2013) a coherent set of recalibrated probabilities can be obtained, which minimizes the Euclidean distance between the point {.2, .3, .4, .3} and all quartet vectors with values between 0 and 1, such that the sum of the four values is 1. For this example, the probabilities of .15, .25, .35, and .25 represent the closest coherent set, with minimum distance as follows:

$$\sqrt{(.2-.15)^2+(.3-.25)^2+(.4-.35)^2+(.3-.25)^2}=.10.$$

The resulting value, moreover, represents an incoherence metric, expressed, more generally, as

$$IM=\sqrt{\sum_{i=1}^{K}(y_i-y_i^c)^2}. \quad (1)$$

In Equation 1, IM is calculated over the sum of k judgments that form a related set, and notably IM is zero when elicited judgments are perfectly coherent. The CAP is not limited to using only the unitarity constraint but can be applied with any set of coherence constraints that can be defined mathematically as an optimization program.

As noted earlier, variations in IM across individuals can also be used as a basis for performance-weighted aggregation. With IM_j as the incoherence metric for the j^{th} individual in an aggregate, a weighting function should satisfy general properties. First, it should be strictly decreasing as IM_j increases, thus assigning harsher penalties to more incoherent individuals. Because weights are normalized during the aggregation, only the ratio values of weights are relevant. Thus, the function can be arbitrarily scaled in the [0, 1] interval, with 1 representing a perfectly coherent judge. In the present research, we use a weighting function similar to that of Wang et al. (2011)

$$\omega_j=e^{-IM_j\cdot\beta}. \quad (2)$$

The weighting function assigns full weight to the j^{th} individual if $IM_j=0$ or if $\beta=0$. In the former case, this is due to the perfect coherence of j's raw estimates, while in the latter case the weighting function is nondiscriminatory and equivalent to taking the arithmetic average across individuals.

Next, we define the coherence-weighted average of n (where $2\leq n\leq N$) individuals' coherentized probability judgments of the i^{th} hypothesis as

$$\bar{y}_i^{cc}=\frac{\sum_{j=1}^{n}\omega_j y_{ij}^c}{\sum_{j=1}^{n}\omega_j}. \quad (3)$$

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

Again, if $\beta = 0$, we have an equal-weighted (arithmetic) average of the coherentized judgments

$$\bar{y}_i^c = \frac{1}{n}\sum_{j=1}^{n} y_{ij}^c. \quad (4)$$

Note that the coherence constraints on y_{ij}^c imply that set of all coherent probabilities is a convex set, and any linear combination of elements from a convex set is again an element of the same set. Therefore, the aggregated estimates must also be coherent and do not have to be coherentized again.

In the present research, we let $\beta = 5$, and we later show that the results are not sensitive to the exact value chosen. Choosing a sufficiently large value alleviates the issue with the "fifty-fifty blip", which results when an individual expresses epistemic uncertainty by responding .5 over multiple judgments (Bruine de Bruin Fischbeck, Stiber & Fischhoff, 2002). In the present research, if an analyst entered .5 for each hypothesis, *A*-*D*, the values would sum to 2, and the participant's *IM* score would be .50. In the weighting function, we have $\omega(.50) = .082$. This participant would be assigned only 8.2% of the weight that would be assigned to a perfectly coherent participant.

2.5 Metrics

The primary measure of accuracy we use is mean absolute error (*MAE*), which in this research computes the mean absolute difference between a human-originated judgment (i.e., raw, transformed, or aggregated), y_i, and the corresponding posterior probabilities derived from Bayes theorem assuming class conditional independence (i.e., a "naïve Bayes" model), x_i. We acknowledge that this simplifying assumption is not necessitated by the task. However, we believe it is reasonable to assume that participants did not perceive conditional dependence and subsequently take it into account — at least we found no evidence to support such a conclusion in participants' responses. Using the naïve Bayes model, $x_A = .08$, $x_B = .15$, $x_C4 = .46$, and $x_D = .31$. Accordingly,

$$MAE = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n}|y_{ij} - x_i|. \quad (5)$$

The summation over i refers to the set of hypotheses (i.e., in this research, $k = 4$).

An advantage of *MAE* over mean squared error or root mean squared error is that it is less susceptible to outliers (Armstrong, 2001; Willmott & Matsuura, 2005). In addition, *MAE* is decomposable into quantity disagreement (*QD*) and allocation disagreement (*AD*):

$$QD = |ME|, \qquad \text{where} \qquad MD = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - x_i). \quad (6)$$

$$AD = MAE = QD. \quad (7)$$

QD is the absolute value of mean error (*ME*) or bias. *AD* represents remaining inaccuracy after removal of *QD* (i.e., absolute bias), which necessarily involves a fair balance between under- and over-estimations of correct values (i.e., any imbalance is part of *QD*). Coherentization reduces *MAE* by eliminating *QD*.

As noted earlier, we used a measure of classification accuracy improvement called probability gain (Nelson, 2005) to assess analysts' accuracy in rating cue usefulness:

$$\text{probability gain}\,(Q) = \left[\sum_{q_j} P(q_j)\,\max P(h_i|q_j)\right]. \quad (8)$$

3 Results

3.1 Coherence of probability judgments

We tested the coherence of analysts' probability judgments as a function of tradecraft using the following logical constraints:

$$y_A + y_B + y_C + y_D = 1 \qquad \text{unitary, quarternary partition.} \quad (9)$$

$$y_H + y_F = 1 \qquad \text{unitary, binary partition.} \quad (10)$$

$$y_H = y_A + y_C,\ y_F = y_B + y_D \qquad \text{additivity, two binary partitions.} \quad (11)$$

Equations 9 and 10 reflect the unitarity axiom and Equation 11 reflects the additivity axiom.[2] In violation of Equation 9 and showing a strong unpacking effect, the sum of the probabilities assigned to the four MECE hypotheses significantly exceeded unity in the control condition ($M = 1.54$ [1.33, 1.76], $t[24] = 13.63$, $d = 0.96$, $p < 001$) and in the ACH condition ($M = 1.77$ [1.56, 1.97], $t[24] = 19.53$, $d = 1.69$, $p < .001$). The unpacking effect did not significantly differ between conditions, but the difference in the size of these effects was nevertheless of medium effect size by Cohen's (1992) standards and favored the control group, $\Delta = 0.22$ [−0.08, 0.55], $t[45.8)] = 1.55$, $d = 0.45$, $p = .13$.

In contrast, but consistent with several studies also finding unitarity for binary complements (e.g., Brenner & Rottenstreich, 1999; Dhami & Mandel, 2013; Mandel, 2005; Tversky & Koehler, 1994), the total probability assigned to the binary complements, H and F, did not significantly differ from unity in either the control condition ($M = 0.98$ [0.90, 1.04], $t[24] = 27.38$, $d = 0.12$, $p < .001$) or the ACH condition ($M = 0.95$ [0.83, 1.00], $t[24] = 23.45$, $d = 0.25$,

[2]Square brackets show bootstrapped 95% confidence intervals from 1,000 bias-corrected and accelerated samples, Δ denotes the mean difference between conditions, and d refers to the effect size estimator, Cohen's d.

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

$p < .001$). Thus, on average, analysts respected the unitarity constraint imposed by Equation 10.

Turning to tests of additivity, we computed the sum of the (signed) non-additivity (SSN):

$$SSN = (y_A + y_C - y_H) + (y_B + y_D - y_F). \qquad (12)$$

If Equation 12 is respected, $SSN = 0$. However, it is evident that implicit disjunctions were assigned significantly less probability than what was assigned, in sum, to their constituents in both the ACH condition ($M = 0.82$ [0.64, 0.99], $t[24] = 8.98$, $d = 1.80$, $p < .001$) and the control condition ($M = 0.56$ [0.37, 0.74], $t[24] = 5.34$, $d = 1.07$, $p < .001$). In addition, mean additivity violation, consistent with the unpacking effect, was marginally greater in the ACH condition than in the control condition, $\Delta = 0.25$ [−0.05, 0.57], $t(48) = 1.81$, $d = 0.52$, $p = .08$. Once again, this difference was of medium effect size.

3.2 Accuracy of probability judgments

As noted earlier, we compared the accuracy of analysts' untransformed (i.e., not coherentized) probability judgments for the four-way MECE partition (i.e., Tribes *A*-*D*) using analysts' *MAE* calculated over the four estimates. Although there was a significant degree of inaccuracy in both the control condition (*MAE* = 0.21 [0.17, 0.26], $t[24] = 9.69$, $d = 1.94$, $p < .001$) and the ACH condition (*MAE* = 0.26 [0.22, 0.29], $t[24] = 14.39$, $d = 2.88$, $p < .001$), the effect of tradecraft was not significant, $\Delta = 0.04$ [−0.02, 0.11], $t(45.9) = 1.49$, $d = 0.43$, $p = .14$. Nevertheless, as the effect-size estimate reveals, there was a medium-sized effect of tradecraft that, once again, favored the control group.

The observed *MAE* in the sample was also compared to that obtained from 10,000 random draws of probability values for each of the four hypotheses, *A*-*D* (i.e., where each probability was drawn from a uniform distribution over the [0, 1] interval — a simulated dart-throwing chimp, to use Tetlock's [2005] metaphor). *MAE* for the random judgments was 0.33. Thus, analysts performed significantly better than chance, analysts' *MAE* = 0.23 [0.21, 0.26], $t(49) = 6.69$, $d = 0.95$, $p < .001$.

Given that the *QD* decomposition of *MAE* calculated over the four MECE hypotheses is directly related to unitarity violation and, further, given that we have established that this type of coherence violation is greater in the ACH condition than in the control condition, we can verify that the proportion of total inaccuracy (*MAE*) accounted for by *QD* is greater in the ACH condition than in the control condition. In fact, this was confirmed: The *QD*/*MAE* proportion was .73 [.60, .86] in the ACH condition and .50 [.32, .67] in the control condition, a significant effect of medium size, $\Delta = .23$ [.04, .45], $t(45.7) = 2.08$, $d = 0.60$, $p = .04$.

Although the preceding analyses do not indicate that ACH helps to improve analysts' probability judgments, critics might argue that the method is not aimed at minimizing absolute error but rather at improving the rank ordering of alternative hypotheses in terms of their probability of being correct. To address this point, we calculated the rank-order (Spearman) correlation between each analyst's four raw probability judgments of *A*-*D* and the probability vector of the naïve Bayes model. The mean correlations in the ACH condition ($M = .29$ [.02, .55]) and the control condition ($M = .24$ [−.08, .55] did not significantly differ, $t[46.9] = 0.28$, $d = 0.08$, $p = .78$. Therefore, we find no support for the hypothesis that ACH helped analysts to better assess the relative probability of the four hypotheses.

3.3 Information usefulness

As noted earlier, we hypothesized that the consistency rating process in ACH, which requires analysts to assess each piece of evidence for consistency with each hypothesis, and the subsequent diagnosticity assessment process, which requires analysts to consider information usefulness, might help analysts capture variation in information utility. Accordingly, we computed the Pearson correlation between each analyst's ratings of the information usefulness of the 12 cues and the probability gain values for those cues. Providing support for the preceding hypothesis, the mean correlation in the ACH condition ($M = .68$ [.61, .75]) was significantly greater than the mean value in the control condition ($M = .17$ [−.02, .35]), and the effect size was very large, $t[29.5] = 5.35$, $d = 1.59$, $p < .001$.

Next, we examined whether these correlations were themselves related to analysts' *MAE* scores. Overall, this correlation was non-significant, $r(49) = -.14$ [−.39, .15], $p = .53$. However, the observed relationship was strikingly different between the two conditions. The correlation was negligible in the ACH condition, $r(24) = -.10$ [−.44, .28], $p = .63$, but it was significant and of medium-to-large effect size in the control condition, $r(24) = -.42$ [−.69, −.07], $p = .045$. Although analysts using ACH were more likely than analysts in the control condition to track the variation in probability gain with their usefulness ratings, the degree to which the ACH group tracked probability gain had almost no correspondence to their accuracy, whereas it did for the control group.

3.4 Recalibrating probability judgments

The substantial degree of nonadditivity observed in analysts' probability judgments implies that recalibration procedures that coherentize the judgments will not only ensure coherence, they will also benefit accuracy by eliminating the *QD* component of *MAE*. Thus, we coherentized analysts' probability judgments of *A*-*D* so that they respected the unitar-

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

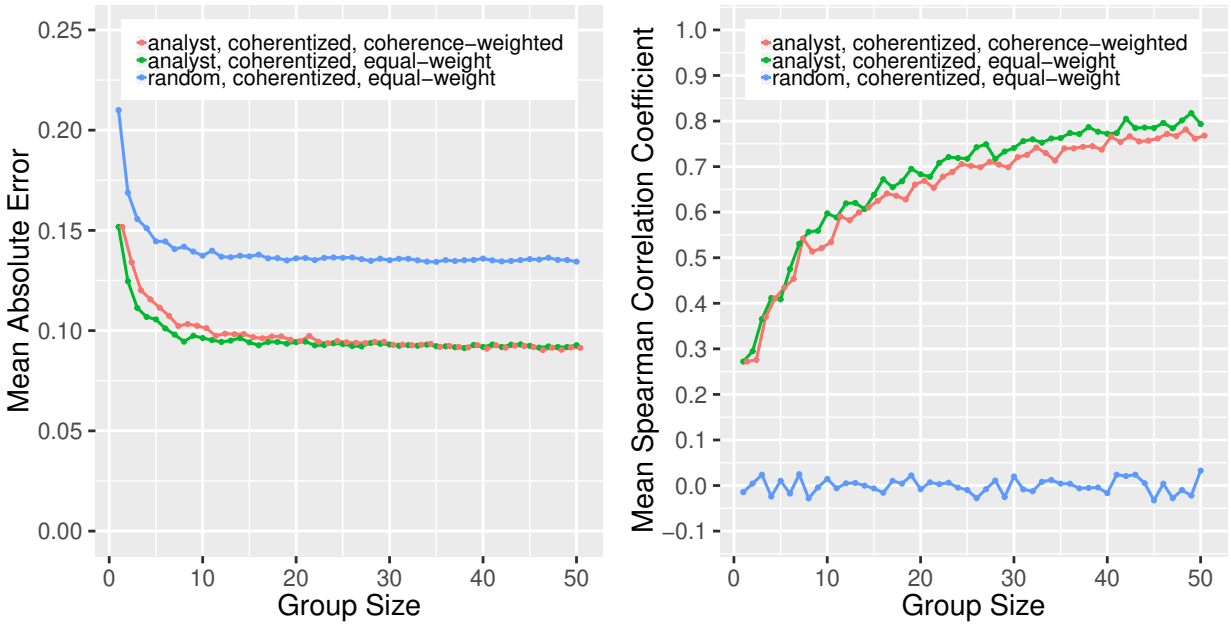

FIGURE 1: Accuracy of probability judgments by group size and aggregation method.

ity constraint in Equation 9.[3] The coherentized probability judgments (*MAE* = 0.15 [0.13, 0.18]) were significantly more accurate than the raw judgments (*MAE* = 0.23 [0.21, 0.26]), Δ = −0.08 [−0.11, −0.06], t[49] = 6.77, d = 0.96, $p < .001$. This represents a 35% reduction in *MAE* and approximately a 1 *SD* improvement. Recall that the proportion of *MAE* attributable to *QD* was significantly greater in the ACH condition than in the control condition. This suggests that the effect of coherentizing will be stronger in the ACH condition. In fact, d = 0.69 in the control condition and d = 1.37 in the ACH condition. Therefore, the *SD* improvement is roughly twice as large in the ACH condition as it is in the control condition. Moreover, after coherentizing, the effect of tradecraft on accuracy is negligible, Δ = 0.01 [−0.04, 0.07].

We once again compared analysts' judgment accuracy to the performance of the average dart-throwing chimp. However, this time we coherentized the randomly generated probabilities, which yielded *MAE* = 0.21, a value that was significantly inferior to the observed coherentized *MAE* of 0.15, t[49] = 4.56, d = 0.64, $p < .001$. An alternative method of assessing chance is to define it in terms of all possible permutations of the probabilities actually provided by each participant, rather than as a uniform distribution. Using this definition, the superiority of the analysts over chance was still apparent but not as large: 0.16 for chance, 0.13 for the participants (t(49) = 2.75, p = .008), a difference of about 0.03 rather than 0.06.[4] This analysis suggests that probability judgments were in the right range, but they were conveying very little information about the relative probabilities of the four hypotheses, this reducing the power of the experiment to detect group differences.

3.5 Aggregating probability judgments

Coherentization yielded a large improvement in the accuracy of analysts' probability judgments. We examined how much further improvement in accuracy might be achieved by aggregating analysts' probability judgments. To do so, we generated 1,000 bootstrap samples of statistical group sizes ranging from 1 (i.e., no aggregation) to 49 in increments of two. We aggregated probability judgments in two ways: using an unweighted arithmetic average of coherentized probability judgments and using a coherence-weighted average of such estimates.[5] We examined the effect of aggregation on *MAE* as well as on the average Spearman correlation between the aggregated estimates and the vector of values from the naïve Bayes model. As a benchmark, we also examined the effect of these aggregation methods on random responses, where each data point is based on 1,000 simulations of probability judgments from a uniform distribution over the [0, 1] interval.

[3] An alternative form of coherentization that used Equations 9 and 11 was also tested but found to be virtually indistinguishable. Thus, we used the simpler form.

[4] Jon Baron conducted this analysis and used normalization (i.e., dividing each stated probability by the sum of the four) rather than coherentization as the recalibration method.

[5] For coherence-weighted aggregation, $\beta = 5$. However, as shown in the supplementary figure, the effect of coherence weighting was robust across a wide parametric range.

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

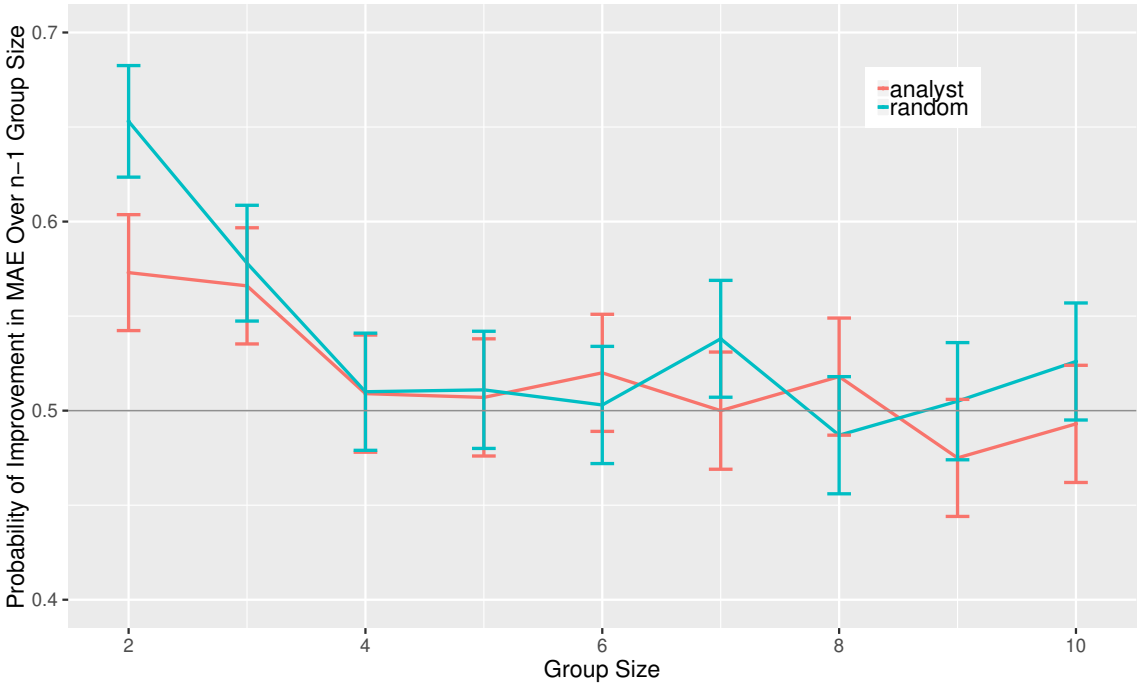

FIGURE 2: Probability of improvement achieved by increasing group size by one member. Bars show 95% confidence intervals from 1,000 bootstrap samples. The reference line shows the probability of improvement by chance.

As Figure 1 shows, these analyses yield several important findings. First, they confirm that, when aggregated, analysts' judgments are substantially more accurate than aggregated random judgments. Second, it is evident from the left panel in Figure 1 that aggregation greatly improves accuracy in analysts' judgments, but to a degree comparable to that observed in the randomly generated response data. This suggests that most of the error reduction observed is due to variance reduction from averaging and should not be attributed to an eking out of any crowd wisdom, as clearly there is no wisdom in the random response data.[6] Third, it is equally evident from the right panel in Figure 1 that aggregation over increasingly larger group sizes steadily increases the correct rank ordering of probabilities. This effect is clearly not manifested in random response data, where aggregation has no benefit. Fourth, aggregation with coherence weighting did not outperform aggregation with equal weighting; in fact, it slightly underperformed. Finally, the left panel in Figure 1 shows that most error reduction due to aggregation was achieved with small group sizes. Figure 2 clarifies that there was a significantly greater the proportion of cases where *MAE* was lower for a group size of two than for single individuals (i.e., the probability of improvement), and likewise the stepwise increase in group size from two to three significantly increased the proportion with lower *MAE* scores. However, no additional stepwise increase in group size yielded significant improvements.

Finally, we assessed the proportional gain in accuracy achieved by recalibration and equal-weight aggregation when $n = 50$. As noted earlier, coherentizing the disaggregated judgments yielded a 35% reduction in *MAE*. If we combine coherentization with equal-weight aggregation of the full sample of 50 analysts, we obtain $MAE = 0.09$, a 61% reduction in *MAE* over the value for analysts' original probability judgments (i.e., $MAE = .23$). That is, 61% of the inaccuracy of analysts' probabilistic assessments of the target's category membership was eliminated by first coherentizing those assessments and then taking an unweighted average of them prior to scoring.

4 Discussion

Although intelligence organizations routinely train and advise analysts to use tradecraft methods, such as ACH, to mitigate cognitive biases and thereby improve the coherence and accuracy of their assessments, there has been a dire lack of research on their effectiveness. The present research conducted such a test and found that ACH failed to improve intelligence analysts' probabilistic judgments about alternative hypotheses. It even had a small detrimental effect on some measures of coherence and accuracy. In such cases, the comparison between conditions yielded a medium effect

[6]Note that aggregation of random responses will bring all responses closer to .25. In the limit, the *MAE* of this constant response may be lower than that for a set of responses with excessive variability.

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

size in favor of the control group. To better understand the advantage of *not* using ACH in the present research task, it is helpful to convert the effect size into a stochastic superiority or probability of superiority estimate equal to the area under the receiver-operator characteristic curve in signal detection theory (Grissom & Kim, 2005; Ruscio & Mullen, 2012; Vargha & Delaney, 2000). The probability of superiority is the probability that a randomly selected member of one condition will outperform a randomly selected member of another condition. For accuracy, for instance, the effect size, $d = 0.45$, yields a probability of superiority estimate equal to .62 favoring the control condition. That is, if one analyst were randomly drawn from the ACH condition and another randomly drawn from the control condition, there would be a 62% chance of the former having *worse* accuracy than the latter.

Comparable probabilities of superiority favoring the control condition likewise are obtained in tests of unitarity and additivity. In each case, coherence violations conformed to the unpacking effect predicted by support theory (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994). As noted earlier, the unpacking effect refers to the tendency for people to assign greater total probability to the sum of a MECE partition of a disjunctive event (in the case of additivity) or an event space (in the case of unitarity). The unpacking effect has been shown to undermine the logical coherence of geopolitical assessments (Tetlock & Lebow, 2001), which suggests that such forms of incoherence can undermine strategic intelligence assessments. Indeed, compared to regular forecasters, elite super-forecasters of geopolitical topics tend to display greater coherence on other, unrelated probabilistic tasks (Mellers et al., 2017). It should concern the intelligence community that a commonplace analytic tradecraft technique served to increase (rather than reduce) this form of judgmental error.

Of course, critics might argue that perhaps analysts interpreted the request for probabilities as requiring only a relative probability assessment of the hypotheses. After all, ACH is primarily aimed at ranking hypotheses by likelihood, and for that reason our control analysts were also instructed to assess the relative likelihoods of the hypotheses. However, we elicited probabilities for each hypothesis separately on a scale covering the [0, 1] interval. Moreover, the four probabilities (*A*-*D*) that are bound by the unitarity axiom were elicited in immediate succession, an elicitation feature shown to mitigate incoherence (Mandel, 2005). Therefore, we expect to find even greater incoherence in analytic practice where the logical relations between assessments are likely to be obscured. Finally, we found that the rank-order correlations between analysts' judgments and the correct values were small, on average, having only about 7% shared variance.

Another striking result of the present research concerns the relationship between the quality of analysts' information usefulness evaluations and the quality of their probability judgments regarding the alternative hypotheses. Although analysts who used ACH provided ratings of probabilistic cue usefulness that were more strongly correlated with the cues' probability gain values than analysts who did not use ACH, the former group's assessments of information usefulness did virtually nothing to guide them to exploit the knowledge effectively to boost accuracy in probability judgments. In contrast, among analysts in the control group, there was substantially better correspondence between accuracy and the degree to which their usefulness ratings tracked probability gain. Analysts in the control group whose usefulness ratings tracked probability gain were better poised than analysts in the ACH group to use that knowledge to improve the accuracy of their probability assessments. This finding was unanticipated and should ideally be tested for reproducibility in future research.

While speculative, one explanation for the disconnect between accurate evaluation of information usefulness and accuracy of probability judgments is that the consistency-encoding phase in ACH prompts analysts to adopt a perspective that is evidence-contingent rather than hypothesis-contingent. That is, analysts are taught to evaluate evidence-hypothesis consistency *within* pieces of evidence and *across* hypotheses rather than the other way around. This approach is deliberate, reflecting Heuer's (1999) belief that analysts are susceptible to confirmation bias and thus need to be made to focus on evidence rather than their preferred hypothesis. The evidence-contingent approach should prompt consideration of information usefulness given that the consistency between a piece of evidence and each hypothesis being evaluated is assessed before proceeding to another piece of evidence. However, we see that information integration within hypotheses is left to the questionable "sum of the inconsistency scores" rule in ACH. Unlike a normative (e.g., Bayesian) approach, this rule merely serves as a summator and, moreover, selectively so by choosing to ignore scores that indicate degree of positive consistency. The integration rule is also exceptionally coarse in its treatment of evidence, assigning one of only three levels (−2, −1, 0) to each piece of evidence, and such coarseness is likely to impede judgment accuracy (Friedman, Baker, Mellers, Tetlock & Zeckhauser, 2018).

Moreover, ACH does virtually nothing as an analytic support tool to ensure that analysts consistently map evidential strength onto −1 and −2 ratings. Consider two hypotheses, A and B. Assume that given five pieces of evidence, three analysts, X, Y, and Z agreed on the following. All five pieces of evidence are inconsistent with A and three pieces are inconsistent with B. Assume further that compared to Y, X has a low threshold for assigning −2 ratings, and Z has a high threshold. All three analysts might agree that the five pieces of evidence are inconsistent with A, but not strongly so, and they would assign -1 for each piece. They might further agree that the three pieces of evidence that are inconsistent

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

with B are stronger in their inconsistency than in the case of A, but given their differing thresholds for assigning −2 ratings, they may vary in their ratings. For instance, X might assign −2 to the three pieces that are inconsistent with B, Y might assign, two −2 ratings and one −1 rating, and Z might assign −1 ratings to each of the three pieces of evidence inconsistent with B. If so, in spite of the substantial agreement among analysts, using ACH, X would judge A less probable than B, Y would judge A and B as equally probable, and Z would judge A as more probable than B!

The present findings indicate that ACH is ineffective as a means of supporting analysts in assessment tasks requiring the integration of uncertain evidence in order to evaluate a set of hypotheses. The findings challenge a widespread assumption among tradecraft professionals in intelligence organizations that, although ACH (and SATs, in general) might not always help the analyst, at least they don't hurt the analyst (Mandel & Tetlock, 2018). Two of the authors (DRM and MKD) who have worked for several years with analytic tradecraft professionals have repeatedly encountered a "nothing to lose" attitude when it comes to SAT training and on-the-job use. Yet, our findings suggest that, in fact, ACH can impede the quality of intelligence assessments. It can do so in two ways: first, by undermining the coherence and accuracy of estimates and, second, by fostering a disconnection between evidence evaluation and hypothesis evaluation. We therefore urge intelligence organizations to be more circumspect about the benefits of training analysts to use ACH and other SATs that have not received adequate testing.

Indeed, a commonplace rebuttal from intelligence professionals to any criticism of tradecraft methods is that although they aren't perfect, intelligence organizations can't just "do nothing." The idea of leaving analysts to their own "intuitive" reasoning is thought to — and often does — result in bias and error. Our findings challenge this assumption since analysts who were left to their own devices performed better than analysts who used ACH.

SAT proponents are likely to object and claim that our findings lack external validity. After all, intelligence analysts seldom are presented with such neat problems where all evidence is precisely quantified and expresses relative frequencies and where the full set of pertinent hypotheses is explicit and, further, it is evident that these hypotheses are also neatly partitioned (i.e., MECE). We agree that in these and other respects the experimental task we used lacks mundane realism. However, we disagree with the implications that proponents would likely draw from such observations. Intelligence problems are murkier in many respects — the quality of evidence will be variable, the hypotheses might be unclearly defined and will often fail to yield a MECE set, and analysts are likely to give no more than vague probability estimates on coarse verbal probability scales (Dhami et al., 2015; Friedman et al., 2018; Mandel, in press; Mandel & Barnes, 2018). We see no compelling reason why ACH should help under those conditions when it does not help hypothesis evaluation under the much more modest requirements of the present experimental task. Indeed, it is possible that ACH can do even more harm to judgment when analysts use it on the job.

Clearly, it would be beneficial to conduct research in the future that uses tasks that are more challenging in the respects noted while permitting unambiguous evaluation of the merits of ACH or other SATs. However, the present research already shows that ACH is not an *all-purpose* judgment corrective for problems involving the evaluation of multiple hypotheses on the basis of uncertain evidence. In fact, the poor performance of both groups of analysts in this research raise a more basic question: why were they so inaccurate on a task (even in terms of their relative probability judgments) that is arguably much easier than the types of so-called puzzles and mysteries they encounter on the job? This may ultimately prove to be a more important finding than the relative performance between conditions. In the present task, analysts had unambiguous sources of accurate information that they could exploit, yet most were at a loss do so regardless of whether they used ACH or not. Our findings therefore raise a fundamental question about the competence of analysts to judge probabilities. Given the small and homogeneous sample of analysts we tested, it would be wrong to draw sweeping generalizations. Yet, if our findings do generalize across a wide range of analyst samples, it should prompt the intelligence community and the bodies that provide intelligence oversight to take stock of the practical significance of the findings and study the putative causes of poor performance.

We also respond to SAT proponents by noting that "doing nothing" is not the only alternative to using conventional analytic tradecraft techniques such as ACH. In this research, we examined two promising statistical methods that intelligence organizations could use to improve probability judgments after analysts had provided judgments — methods we accordingly describe as *post-analytic*. One method, coherentization, exploits the logical structure of related queries by recalibrating probability assessments so that they conform to one or more axioms of probability calculus. As noted earlier, Karvetski, Olson, Mandel et al. (2013) showed that such methods substantially improve the accuracy of probability judgments. Likewise, in the present experiment, a large improvement in analysts' accuracy was achieved by coherentizing analysts' probability judgments such that they respected the unitarity axiom. This method fully counteracted the unpacking effect exhibited by analysts in this research, especially those who were instructed to use ACH. We view CAP-based coherentization as illustrative rather than definitive. Other recalibration methods might be even more effective or easier to apply. For instance, in the present research, we could have coherentized probabilities by simple normalization (i.e., dividing each by their sum), as researchers sometimes do as a step in the statistical analysis of

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

probability judgment data (e.g., Prims & Moore, 2017, Study 2). This method (as noted in Section 3.4) would have yielded even slightly better accuracy than CAP-based coherentization (*MAE* = 0.13 for normalization, vs. 0.15 for CAP). Our study is clearly not designed to examine such competitions given it relies on a single vector of values defining probabilistic accuracy. However, our findings suggest that research comparing optimization methods using such techniques under a broad range of task conditions are needed.

Another post-analytic method intelligence organizations could use to boost the accuracy of probabilistic assessments is to aggregate them across small numbers of analysts. We found that substantial benefits to accuracy were achieved by taking the arithmetic average of as few as three analysts. These findings are consistent with earlier studies showing that most of the advantage from aggregating can be achieved with between two to five judges (e.g., Ashton & Ashton, 1985; Libby & Blashfield, 1978; Winkler & Clemen, 2004). Moreover, we found that a simple equal-weighted aggregate of analysts' judgments yielded comparable benefit to the more complex coherence-weighted aggregation method. This result was unexpected given the superior performance coherence weighting afforded over equal weighting in recent studies (Karvetski, Olson, Mandel et al., 2013; Wang et al., 2011). A key difference between the tasks in the present research and Karvetski, Olson, Mandel et al. (2013) is that the former included all information relevant to solving the task, whereas the latter relied on participants' knowledge of world facts, such as who was the first person to walk on the moon. Thus, whereas in the present research, coherentization may have already reaped most of the benefit achievable through coherence weighting, in the earlier studies coherence weighting might also have benefited accuracy by predicting how knowledgeable participants were.

More generally, the present results indicate that intelligence organizations should be exploring how to effectively incorporate processes for eliciting judgments from multiple analysts and then aggregating them in order to reduce judgment error. At present, intelligence organizations rarely capitalize on statistical methods such as the recalibration and aggregation approaches shown to be effective in the present research. Instead, the management of intelligence production tends to rely on traditional methods such as having sole-source analysts provide input to an all-source analyst (an approach that is common at the operational level), or by having a draft intelligence report reviewed by peers with relevant domain expertise and by the analyst's director (an approach often employed at the strategic level). Still, we caution not to infer too much from the aggregation results. It is tempting to suggest that the aggregate divines the wisdom of crowds, as Surowieki (2004) put it, yet our finding that aggregation of random response data yielded comparable error reduction as in analysts' judgments clearly challenges that interpretation as there was no wisdom in the random data to divine. Our analysis of how aggregates can improve *relative* probability assessment, however, showed a large improvement in accurately capturing the rank ordering of probabilities, and this benefit was entirely absent in the random response data, which suggests that aggregation did in fact boost the signal-to-noise ratio in analysts' ordered probability judgments.

To conclude, we argue that the intelligence community should look to recent examples of research that illustrate how organizations could better integrate recalibration and aggregation methods pioneered in decision science into day-to-day analytic practices. One example involves the systematic monitoring of probabilistic forecast accuracy within intelligence organizations (e.g., Mandel, 2015a; Mandel & Barnes, 2018). The results of such monitoring have shown that analysts' forecasts tend to be underconfident, and that the calibration of intelligence units can be improved post-judgment through an organizational recalibration process that "extremizes" overly-cautious forecasts (Mandel & Barnes, 2014; Baron et al., 2014; Turner et al., 2014). Another example is the introduction in the US intelligence community of a classified prediction market that poses forecasting questions not unlike those worked on by strategic analysts as part of their routine assessment responsibilities. Stastny and Lehner (2018) showed that analysts' forecasts within the prediction market, which aggregated the forecasters' estimates but also shared the aggregated estimates with the forecasters, were substantially more accurate than the same forecasts arrived at through conventional analytic means. These examples illustrate the benefits to analytic accuracy and accountability that intelligence organizations could accrue if they leveraged post-analytic mathematical methods for boosting the quality of expert judgment.

References

Armstrong, J. S. (2001). Evaluating forecasting methods. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners*. Norwell, MA: Kluwer.

Ashton, A. H., & Ashton, R. H. (1985). Aggregating subjective forecasts: some empirical results. *Management Science*, *31*, 1499–1508.

Ayton, P. (1997). How to be incoherent and seductive: Bookmakers' odds and support theory. *Organizational Behavior and Human Decision Processes, 72*, 99–115.

Baratgin, J., & Noveck, I. (2000). Not only base-rates are neglected in the Lawyer-Engineer problem: an investigation of reasoners' underutilization of complementarity. *Memory & Cognition*, *28*, 79–91.

Baron, J. (1985). *Rationality and intelligence*. Cambridge, UK: Cambridge University Press.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*, 133–145.

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

Belton, I., & Dhami, M. K. (in press). Cognitive biases and debiasing in intelligence analysis. In R. Viale & K. Katzikopoulos (Eds.), *Handbook on bounded rationality*. London: Routledge.

Brenner, L. A., & Rottenstreich, Y. (1999). Focus, repacking and the judgment of grouped hypotheses. *Journal of Behavioral Decision Making, 12*, 141–148.

Bruine de Bruin, W., Fischbeck, P. S., Stiber, N. A., & Fischhoff, B. (2002). What number is "fifty-fifty"?: Redistributing excessive 50% responses in elicited probabilities. *Risk Analysis*, *22*, 713–723.

Butler, Lord (2004). *Review of intelligence on weapons of mass destruction, report of a committee of privy councillors (the Butler Report)*, HC 898, London, The Stationery Office.

Chang, W., Berdini, E., Mandel, D. R., & Tetlock, P. E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security, 33*, 337–356.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Anaysis, 19*, 197–203.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Convertino, G., Billman, D., Pirolli, P., Massar, J. P., & Shrager, J. (2008). The CACHE study: Group effects in computer-supported collaborative analysis. *Computer Supported Cooperative Work*, *17*, 353–393.

Coulthart, S. J. (2017). An evidence-based evaluation of 12 core structured analytic techniques. *International Journal of Intelligence and CounterIntelligence, 30*, 368–391.

Dhami, M. K., Belton, I. K., & Careless, K. E. (2016). Critical review of analytic techniques. *2016 European Intelligence and Security Informatics Conference*, 152–155. http://dx.doi.org/10.1109/EISIC.2016.33.

Dhami, M. K., & Mandel, D. R. (2013). How do defendants choose their trial court? Evidence for a heuristic processing account. *Judgment and Decision Making, 8,* 552–560.

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science, 106*, 753–757.

Fox, C. R., Rogers, B., & Tversky, A. (1996). Option traders exhibit subadditive decision weights. *Journal of Risk and Uncertainty, 13*, 5–19.

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

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis.* Washington, DC: Center for the Study of Intelligence.

Heuer, R. J., Jr., & Pherson, R. H. (2014). *Structured analytic techniques for intelligence analysis*. Washington, DC: CQ Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

Karvetski, C. W., Olson, K. C., Gantz, D. T., & Cross, G. A. (2013). Structuring and analyzing competing hypotheses with Bayesian networks for intelligence analysis. *EURO Journal on Decision Processes, 1*, 205–231.

Karvetski, C. W., Olson, K. C., Mandel, D. R., & Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis, 10*, 305–326.

Kretz, D. R., Simpson, B. J., & Graham, C. J. (2012). A game-based experimental protocol for identifying and overcoming judgment biases in forensic decision analysis. *2012 IEEE Conference on Technologies for Homeland Security*, 439–444. Waltham, MA: IEEE. http://dx.doi.org/10.1109/THS.2012.6459889.

Jones, N. (2018). Critical epistemology for Analysis of Competing Hypotheses. *Intelligence and National Security, 33*, 273–289.

Lehner, P. E., Adelman, L., Cheikes, B. A., & Brown, M. J. (2008). Confirmation bias in complex analyses. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 38*, 584–592.

Libby, R., & Blashfield, R. K. (1978). Performance of a composite as a function of the number of judges. *Organizational Behavior and Human Performance, 21*, 121–129.

Macchi, L., Osherson, D., and Krantz, D. H. (1999). A note on superadditive probability judgment. *Psychological Review, 106*, 210–214.

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

Mandel, D. R. (2008). Violations of coherence in subjective probability: a representational and assessment processes account. *Cognition, 106*, 130–156.

Mandel, D. R. (2015a). Accuracy of intelligence forecasts from the intelligence consumer's perspective. *Policy Insights from the Behavioral and Brain Sciences, 2,* 111–120.

Mandel, D. R. (2015b). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology 6*, article 387, 1–12. http://dx.doi.org/10.3389/fpsyg.2015.00387

Mandel, D. R. (in press). Can decision science improve intelligence analysis? In S. Coulthart, M. Landon-Murray, & D. Van Puyvelde (Eds), *Researching national security intelligence: A reader*. Washington, DC: Georgetown University Press.

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences, 111*, 10984–10989.

Mandel, D. R., & Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making, 31*, 127–137.

Mandel, D. R., & Tetlock, P. E. (2018). *Correcting judgment correctives in national security intelligence*. Manuscript submitted for publication.

Marchio, J. (2014). Analytic tradecraft and the intelligence community: Enduring value, intermittent emphasis. *Intelligence and National Security*, *29*, 159–183.

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Making, 12*, 369–381.

National Research Council (2011). *Intelligence analysis for tomorrow: Advances from the behavioral and social sciences*. Washington, DC: National Academies Press.

Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact and information gain. *Psychological Review, 112*, 979–999.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220.

Osherson, D., & Vardi, M. Y. (2006). Aggregating disparate estimates of chance. *Games and Economic Behavior*, *56*, 148–173.

Pool, R. (2010). *Field evaluation in the intelligence and counterintelligence context: Workshop summary*. Washington, DC: National Academies Press.

Pope, S., & Jøsang, A. (2005). Analysis of competing hypotheses using subjective logic. In *10th CCRTS: The Future of Command and Control*, pp. 1–30.

Popper, K. (1959). *The logic of scientific discovery*. London, UK: Hutchison & Co.

Predd, J. B., Osherson, D. N., Kulkarni, S. R., & Poor, H. V. (2008). Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Analysis, 5*, 177–189.

Prims, J. P., & Moore, D. A. (2017). Overconfidence over the lifespan. *Judgment and Decision Making, 12*, 29–41.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review*, *104*, 406–415.

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research, 47*, 201–223.

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *30*, 573–582.

Stastny, B. J., & Lehner, P. E. (2018). Comparative evaluation of the forecast accuracy of analysis reports and a prediction market. *Judgment and Decision Making, 13*, 202–211.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Double Day.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Tetlock, P. E., & Lebow, R. N. (2001). Poking holes in counterfactual covering laws: Cognitive styles and historical reasoning. *American Political Science Review, 95,* 829–843.

Tsai, J., & Kirlik, A. (2012). Coherence and correspondence competence: Implications for elicitation and aggregation of probabilistic forecasts of world events. *Proceedings of Human Factors and Ergonomics Society 56th Annual Meeting*, pp. 313–317, Thousand Oaks, CA: Sage.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*, 261–289.

Tversky, A., & Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychological Review, 101*, 547–567.

UK Ministry of Defence (2013). *Quick wins for busy analysts*. London: author.

US Government (2009). *A tradecraft primer: Structured analytic techniques for improving intelligence analysis*. Washington, DC: Center for the Study of Intelligence Analysis.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25,* 101–132.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39*, 176–190.

Wang, G., Kulkarni, S. R., Poor, H. V., & Osherson, D. N. (2011). Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis*, *8*, 128–144.

Wheaton, K. (2014). *Reduce bias in analysis: Why should we care?* Retrieved from http://sourcesandmethods.blogspot.com/2014/03/reduce-bias-in-analysis-why-should-we.html.

Winkler, R. L., & Clemen, R.T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis*, *1*, 167–176.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*, 79–82.

https://doi.org/10.1017/S1930297500006628 Published online by Cambridge University Press