Received: 19 September 2018 | Revised: 7 March 2019 | Accepted: 18 March 2019

DOI: 10.1002/acp.3550

RESEARCH ARTICLE

WILEY

The "analysis of competing hypotheses" in intelligence analysis

Mandeep K. Dhami[1] | Ian K. Belton[2] | David R. Mandel[3]

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

[2] Department of Management Science, University of Strathclyde, Glasgow, UK

[3] Intelligence, Influence and Collaboration Section, DRDC-Toronto Research Centre, Toronto, Ontario, Canada

Correspondence
Professor Mandeep K. Dhami, Department of Psychology, Middlesex University, Hendon, The Burroughs, London NW4 4BT, UK.
Email: m.dhami@mdx.ac.uk

Funding information
Canadian Safety and Security Program, Grant/Award Number: CSSP-2016-TI-2224; Canadian Department of National Defence, Grant/Award Number: 05da; HM Government, Grant/Award Number: NA

Summary

The intelligence community uses "structured analytic techniques" to help analysts think critically and avoid cognitive bias. However, little evidence exists of how techniques are applied and whether they are effective. We examined the use of the analysis of competing hypotheses (ACH)—a technique designed to reduce "confirmation bias." Fifty intelligence analysts were randomly assigned to use ACH or not when completing a hypothesis testing task that had probabilistic ground truth. Data on analysts' judgement processes and conclusions were collected using written protocols that were then coded for statistical analyses. We found that ACH-trained analysts did not follow all of the steps of ACH. There was mixed evidence for ACH's ability to reduce confirmation bias, and we observed that ACH may increase judgement inconsistency and error. It may be prudent for the intelligence community to consider the conditions under which ACH would prove useful and to explore alternatives.

KEYWORDS

confirmation bias, hypothesis testing, intelligence analysis, judgement and decision making

1 | INTRODUCTION

Intelligence analysts are required to assess evidence to test alternative accounts of a current or future situation. In performing such a cognitively complex task, analysts may resort to using simple strategies that can bias their thinking and result in judgement errors (Belton & Dhami, in press). In particular, it is argued that analysts may suffer from "confirmation bias" (Heuer, 1999). This can manifest itself in a number of ways (see Klayman, 1995; Nickerson, 1998). Analysts are often portrayed as not considering alternative hypotheses; searching for evidence supporting rather than disconfirming their prior beliefs; reaching conclusions about a hypothesis based on the presence of supporting rather than conflicting evidence; and insufficiently adjusting their belief in a hypothesis when existing (supporting) evidence is discredited (e.g., Cook & Smallman, 2008; Lehner, Adelman, Cheikes, & Brown, 2008; Lehner et al., 2009). Indeed, confirmation bias is a popular explanation for intelligence failures such as the Iraq weapons of mass destruction mis-estimate (Jervis, 2006).

In an effort to assist analysts to think critically and avoid bias, the intelligence community has adopted the use of "structured analytic techniques." The analysis of competing hypotheses (ACH; Heuer, 1999, 2005) is one such technique. It is designed to help analysts avoid "confirmation bias" in several respects, namely, by explicitly requiring them to (a) consider alternative hypotheses; (b) rate evidence as inconsistent (or consistent) with each hypothesis under consideration; (c) adjust their belief in a hypothesis in accordance with evidence diagnosticity (or credibility); (d) select the most likely hypothesis based solely on (it being the one with the least) inconsistent evidence; and (e) identify indicators that will disconfirm (or confirm) a hypothesis in the future.

Critics of ACH have noted several shortcomings (e.g., Chang, Berdini, Mandel, & Tetlock, 2018; Jones, 2017; Mandel, in press; Murukannaiah, Kalia, Telang, & Singh, 2015). It is vague in multiple respects. For instance, it is unclear how hypotheses should be selected; what criteria should be

used to rate evidence as being consistent or inconsistent with a hypothesis; or what criteria should be used to judge evidence diagnosticity. This vagueness permits the analyst's judgement process to become unreliable. Finally, ACH does not represent some features of relevant normative methods, such as Bayesianism, for revising beliefs in the face of uncertain evidence. For instance, it provides no guidance on how prior beliefs should be revised in light of new evidence, and so it may be prone to base rate neglect (Bar-Hillel, 1980). ACH's information integration rule involves merely counting the number of weighted inconsistent evidence items for any given hypothesis, while discounting the amount of supporting evidence. Consequently, ACH will diverge from the predictions of Bayes theorem under some conditions, such as when the prior probability distribution over the set of hypotheses is far from uniform (as in the experiment reported here).

Despite these limitations, the intelligence community continues to hold the belief that ACH encourages critical thinking and cognitively debiases analysts (e.g., Marrin, 2008). Indeed, ACH is one of the few techniques listed in the U.S. Government's (2009) *Tradecraft Primer*. ACH also features prominently in the UK Ministry of Defence's (2013) *Quick Wins for Busy Analysts* handbook. The popularity of ACH is surprising given the dearth of empirical research testing its utility (Chang et al., 2018; Dhami, Mandel, Mellers, & Tetlock, 2015; Pool, 2010).

2 | PAST RESEARCH ON THE ANALYSES OF COMPETING HYPOTHESES

The small body of past research is conceptually vague in terms of the features of ACH being tested, although there is a general focus on measuring some aspects of confirmation bias (Convertino, Billman, Pirolli, Massar, & Shrager, 2008; Kretz & Granderson, 2013; Kretz, Simpson, & Graham, 2012; Lehner et al., 2008). Specifically, the studies induce confirmation bias in participants before testing ACH by presenting evidence in stages such that it initially favours one hypothesis, and then in later stages, it either balances out across the hypotheses (Convertino et al., 2008), supports a hypothesis it initially conflicted with (Lehner et al., 2008), or conflicts with the hypothesis it initially supported (Kretz et al., 2012; Kretz & Granderson, 2013). Thus, researchers cannot comment on how ACH may reduce other aspects of confirmation bias such as explicitly requiring them to consider alternative hypotheses, rate evidence as inconsistent (or consistent) with each hypothesis under consideration, adjust belief in a hypothesis in accordance with evidence diagnosticity (or credibility), and identify indicators that will disconfirm (or confirm) a hypothesis in the future.

Lehner et al. (2008) reported that ACH reduced confirmation bias (measured in terms of the size of the significant positive correlations between participants' confidence in an initial hypothesis and their ratings of the extent to which subsequent evidence supported that hypothesis) in participants with no analytic experience but not in those with experience. This was partly because the latter group initially demonstrated less bias. ACH, however, did not appear to reduce participants' resistance to change from one hypothesis to another. Convertino et al. (2008) found that confirmation bias (measured in terms of belief in the initially supported hypothesis in later phases and the importance attached to evidence supporting the favoured hypothesis) was evident across all groups studied, but stronger in the group with similar beliefs rather than dissimilar beliefs. Kretz et al. (2012) and Kretz and Granderson (2013) found that participants using ACH did not consistently outperform those using one of two other techniques (i.e., link analysis and information extraction and weighting) in terms of the number of hypotheses generated and how often the chosen hypothesis was supported by the evidence overall.

The aforementioned studies have several shortcomings. They were based on very small samples, precluding statistical testing of the reliability and size of any effects reported. Lehner et al. (2008) studied 24 individuals. Convertino et al.'s (2008) study involved nine, three-member, geographically distributed groups of students. Kretz et al. (2012) and Kretz and Granderson (2013) studied 27 junior engineers without analytic experience. In addition, some past studies did not include relevant control groups against which ACH could be compared. In Convertino et al.'s (2008) study, all groups used a collaborative version of ACH. Kretz et al. (2012) and Kretz and Granderson (2013) compared ACH with two techniques whose primary function is not hypothesis testing (see Dhami, Belton, & Careless, 2016). Furthermore, some of the studies using a control group were confounded by the fact that some "control" participants were familiar with ACH and may have used it (Kretz et al., 2012; Kretz & Granderson, 2013; Lehner et al., 2008). Finally, none of the studies measured whether analysts in the ACH group applied ACH correctly.

3 | THE PRESENT RESEARCH

We examine all of the features of ACH using a comparatively larger sample of practicing intelligence analysts, half of whom were randomly assigned to be trained to use ACH and half of whom were not. The experiment had three main aims. First, we sought to compare the judgement processes of analysts trained (and instructed) to use ACH against analysts from the same cohort not trained in ACH and not instructed to use any particular technique (i.e., control group). According to proponents of ACH, the untrained group ought to demonstrate greater "confirmation bias" than the ACH group in several respects. In the context of our experiment, this bias is conceptualized as: (a) not considering all alternative hypotheses; (b) only evaluating evidence based on whether it is consistent with each hypothesis under consideration; (c) not adjusting belief in a hypothesis in accordance with evidence diagnosticity; (d) selecting the most likely hypothesis based solely on evidence that is consistent with it; and (e) identifying indicators that will only confirm a hypothesis in the future.

The second aim was to measure the extent of within-individual consistency in the judgement processes of the ACH and untrained groups. It is reasonable to expect that analysts taught to use ACH may demonstrate greater consistency in how they approach a hypothesis testing task compared with those who have not been taught to use ACH.

The final aim was to compare the accuracy of the ACH and untrained groups. Although ACH was designed to reduce judgement

bias and error, as Dhami et al. (2016) point out, techniques such as ACH cannot guarantee accuracy. This is partly because they rely on the judgement skills of the analyst and his/her subjective input of the information and interpretation of the outputs. Past research on ACH does not sufficiently comment on its ability to help analysts arrive at the correct solution; however, the implicit belief among the intelligence community is that ACH can improve the accuracy of those analysts who use it as opposed to those who do not.

4 | METHOD

The present study received ethics approval from the Middlesex University Department of Psychology Research Ethics Committee.

4.1 | Participants

Analysts undergoing their regular training at a UK intelligence organization were asked by the trainers to participate in the experiment. In total, 50 analysts participated, and there was no attrition.[1] Fifty-seven per cent of the sample was male. The mean age of the sample was 27.79 years (*SD* = 5.03). The mean number of months' experience working as an analyst was 14.08 (*SD* = 29.50). Half of the sample was randomly allocated to the experimental group and half to the control group. The two groups did not differ significantly on any of the aforementioned demographic variables.

4.2 | Analytic task and measures

ACH training was based on the latest version of ACH (see Heuer & Pherson, 2014; Pherson Associates, LLC, n.d.; see Table A1). The training was delivered by the organization's trainers during a half-day session, and analysts were "assessed" using in-class exercises. (The control group was given ACH training after the experiment)

Analysts each performed an analytic task (i.e., judging the likelihood that a target individual belongs to one tribe) comprising four hypotheses (i.e., tribes) and 12 evidence items (e.g., language spoken). The probability of occurrence of each evidence item (i.e., the diagnostic probability) was provided, as was the base rate information for each hypothesis (see Table A2 for task properties). The task enabled analysts to apply all of the steps of ACH and arrive at a normatively correct conclusion by relying solely on the available information. A statistical analysis using Bayes theorem under the assumption of cue independence (i.e., a naïve Bayes model of the evidence) shows that the most probable hypothesis is that the target is a member of the Conda tribe (46% chance). The probabilities for the other tribes are Dengo (31%), Bango (15%), and Acanda (8%). (Although we acknowledge that the task does not demand the simplifying assumption of cue independence, we found no evidence—such as discussion of inter-cue relationships—to suggest the invalidity of the assumption.)

Analysts were instructed as follows: "In this task, you will be asked to assess the tribe membership of a randomly selected person from a region. The region and groups are fictitious and bear no intended relationship to any real groups in any region on Earth. Your task is to use the information provided to offer the best assessment you can of the target person's tribe membership. After reading the scenario, you will be asked to detail your analysis. Then, you will be asked to assess the likelihood of specific hypotheses and the usefulness of the various pieces of information that you received."

The scenario was as follows: "In the Zuma region of Zanda, there are four tribes called Acanda, Bango, Conda, and Dengo. They represent 5%, 20%, 30%, and 45% of Zuma's population, respectively. Assume that Acanda and Conda are hostile tribes, whereas Bango and Dengo are friendly. Your government would like to improve its understanding of this region and has captured a randomly chosen inhabitant to be interviewed. The inhabitant was given a truth serum and will have provided accurate information. In this sense, your task is already easier than in real life since you don't have to worry about inaccuracies in the information provided. Moreover, you may assume that this target, when released, will have no memory of the capture and his brief absence will not have been noticed by any Zumans. Finally, the sex of the target (male) is non-diagnostic since all tribes have the same ratio of males to females (1:1)."

Participants were then told "Assume that your government has already determined the following information which is at your disposal." See Table A2 for a summary description of the four tribes in terms of the 12 evidence items, and the information about the target.

Data were collected using a written protocol. The ACH group was told "In order to solve the analytic task presented, we would like you to use the technique called 'Analysis of Competing Hypotheses' (ACH). This consists of the steps described below. Please use the space provided to detail your analysis using ACH" (see Table A3). The control group was told "Report your conclusions in the box below. 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. Also say why alternative hypotheses were rejected. (You can use the page overleaf to make any notes you need.)"[2]

Participation took either a morning or afternoon of a scheduled training day. All data collection occurred at the intelligence organization in small groups in training rooms. All tasks were completed using pen and paper. Participants were debriefed at the end.

4.3 | Data coding

The data in the written protocols were coded using a structured coding scheme (a copy is available from the first author). This scheme was divided into three parts. The first enabled coding of variables pertaining to data that could potentially be available for both groups (e.g., selection of tribe membership of the target individual and whether the analyst took account of base rate information). The second part contained codes for variables pertaining to data we would expect to observe for the ACH group only given the contents of their training (e.g., did they draw a ACH matrix?). The final part enabled

[1]Further data collection was not possible because the organization changed its training regime.

[2]After completing the above task, all analysts were asked to each complete a posttask questionnaire, and the results of which are reported in Mandel et al. (2018).

TABLE 1 Frequencies for ACH and untrained (control) groups on main variables of interest

	ACH		Control	
Measure	Yes	No	Yes	No
Identified all hypotheses	23	2	25	0
Identified all relevant evidence	23	2	17	8
Reformatted data	25	0	20	5
Used scoring rule to assess evidence	25	0	20	5
Used base rates	3	22	13	12
Considered evidence diagnosticity	20	5	8	17
Evidence integration*				
Only inconsistent evidence	5	—	0	—
Only consistent evidence	1	—	4	—
Both consistent and inconsistent evidence	19	—	16	—
Conducted sensitivity analysis	15	10	1	24
Consistent in applying evidence scoring rule	1	24	4	5
Consistent in evidence integration	19	6	12	3
Consistent in final conclusion*	16	9	16	0
Correctly ranked hypotheses*	1	24	2	14
Correctly chose most likely hypothesis*	9	16	8	16

Note. *See footnotes in Section 5 for why $n < 25$ in either group.

coding of data that could be available for the untrained group only (e.g., did they reformat the data? If so, how?).

5 | RESULTS

The coded data are presented in Table 1. In order to examine the association between group (ACH or untrained/control) and performance on specific aspects of the analytic task, we analysed the data using chi-square tests of independence supplemented with effect size measures. The results are presented below in order of the three main aims of the study.

5.1 | Analysts' judgement processes

Our first set of analyses measured the association between judgement process and group (ACH vs. untrained). Several aspects of the judgement process were examined, following the steps of ACH (see Table A1).

5.1.1 | Task understanding

ACH requires analysts to identify all possible, mutually exclusive hypotheses and evidence items relevant for testing these. Most of the analysts in the ACH group (i.e., 92%; $n = 23$) and all ($n = 25$) of the untrained group identified the four specific hypotheses in the scenario, χ^2 (1, $N = 50$) = 2.08, $p = .149$, $\phi = -.20$.

A statistically significantly greater proportion of the ACH group (i.e., 92%, $n = 23$) compared with the untrained group (i.e., 68%, $n = 17$) identified all 12 evidence items relevant for testing the alternative hypotheses, χ^2 (1, $N = 50$) = 4.50, $p = .034$, $\phi = .30$. There were no observable order effects in the evidence items that were ignored by analysts in each group.

5.1.2 | Task representation

ACH requires analysts to represent the task in terms of a matrix with hypotheses as columns and evidence as rows, and all of the ACH group did this. Eighty per cent ($n = 20$) of the untrained group also reformatted the data. The difference between groups was statistically significant, χ^2 (1, $N = 50$) = 5.56, $p = .018$, $\phi = .33$. A closer examination of data from the 20 analysts in the untrained group who reformatted the task revealed that 16 drew a matrix (i.e., 14 with hypotheses as columns and evidence as rows, and two with evidence as columns and hypotheses as rows) and four made a list.

5.1.3 | Evidence assessment

All of the ACH group applied the scoring rule for assessing evidence in relation to each hypothesis, as instructed. Eighty per cent ($n = 20$) of the untrained group used some form of scoring rule. The difference between groups was statistically significant, χ^2 (1, $N = 50$) = 5.56, $p = .018$, $\phi = .33$. Of those 20 analysts in the untrained group who used a scoring rule, eight added up the evidence likelihood percentages for each hypothesis[3] or performed a similar calculation, whereas 12 attached points for matching evidence in different ways (four of these divided the scale in half so that >50% = 1 point, six divided the scale into several intervals so that ≥75 = 3 points, ≥50% = 2 points, and ≥25% = 1 point, and the remaining two gave a point to the hypothesis that was the best match for each evidence item in terms of having the highest likelihood for that item).

5.1.4 | Evidence diagnosticity

In order to evaluate how well analysts assessed evidence diagnosticity, we examined the ACH group and untrained group separately. Only 11 analysts produced, as instructed, an amended ACH matrix (see Step 4 of the ACH process in Table A1) with evidence items reordered from the original (Step 3) matrix based on their diagnosticity. A further nine analysts removed one or more items but did not reorder their matrix. For each of these 20 analysts, we compared the rankings of the evidence items with the ranking computed using "information gain," an information utility measure that gauges reduction in Shannon entropy (see Nelson, 2005).[4] Mean Kendal's tau b was .63 ($SD = 0.16$). The tau b correlations were statistically significant for eight analysts ($p < .05$), indicating a degree of accuracy in these analysts' assessments of evidence diagnosticity.

Individual analysts in the untrained group did not list enough evidence items as diagnostic to compute individual correlations between their rankings and an objective measure. Therefore, the correlational

[3]Six correctly transformed the negative values first, one did not, and it was difficult to determine what the last one did.

[4]Shannon entropy is a well known, general purpose measure of information in a communication (or for present purposes, a measure of the diagnosticity of an evidence item).

analysis was computed across the whole group by comparing the percentage of analysts that identified the evidence items as diagnostic with the items' ranking using the Shannon entropy reduction measure. Kendal's tau b was .44, $p = .07$.

5.1.5 | Taking account of diagnosticity and base rates

ACH does not provide any guidance on the use of base rates, and a statistically significantly smaller proportion (i.e., 12%, $n = 3$) of analysts in the ACH group used base rate information compared with 52% ($n = 13$) of the untrained group, χ^2 (1, $N = 50$) = 9.19, $p = .002$, $\phi = -.43$.

Eighty per cent ($n = 20$) of the ACH group took some account of evidence diagnosticity, as instructed (i.e., by deleting some evidence items in their revised matrix and/or reordering their matrix based on diagnosticity). Thirty-two per cent ($n = 8$) of the untrained group ranked evidence items in some way based on diagnosticity or stated that they took account of diagnosticity in reaching their conclusion. The difference between groups was statistically significant, χ^2 (1, $N = 50$) = 11.69, $p = .001$, $\phi = .48$.

5.1.6 | Evidence integration

When selecting the most likely hypothesis, ACH requires analysts to add up only evidence inconsistent with each hypothesis, ignoring evidence consistent with it, and to consider the hypothesis with the lowest number of inconsistent ratings as most likely. We found a statistically significant difference between the two groups in how they selected the most likely hypothesis, χ^2 (2, $N = 50$) = 6.58, $p = .037$, $V = .38$. Post hoc analyses were conducted to further explore the source of this difference. We found that despite their training, only 20% ($n = 5$) of analysts in the ACH group relied solely on inconsistent evidence, and none of the untrained group[5] did so. This difference between groups was statistically significant, χ^2 (1, $N = 50$) = 5.14, $p = .023$, $\phi = .33$. A small minority of analysts in both groups added up only evidence consistent with each hypothesis (i.e., ACH: 4%, $n = 1$ and untrained: 22%, 4 out of $n = 20$), χ^2 (1, $N = 50$) = 2.88, $p = .090$, $\phi = -.25$. Finally, the majority of analysts in both groups added up both consistent and inconsistent evidence (i.e., ACH: 76%, $n = 19$ and untrained: 78%, 16 out of $n = 20$), χ^2 (1, $N = 50$) = 0.01, $p = .954$, $\phi = -.01$.

5.1.7 | Sensitivity analysis and indicators for future observation

Finally, ACH requires analysts to assess the sensitivity of their conclusions and identify indicators for future observation that would support or contest their conclusion. A statistically significantly greater proportion of analysts in the ACH group (i.e., 60%, $n = 15$) checked the sensitivity of their conclusions to a change in assumptions compared with 4% ($n = 1$) of the untrained group, χ^2 (1, $N = 50$) = 18.02, $p < .001$, $\phi = .60$.

Seventy-two per cent ($n = 18$) of analysts in the ACH group provided at least one indicator. A total of 68 indicators were provided by these analysts, with 22 indicators potentially confirming their conclusion, 19 disconfirming it, and 27 being neutral (i.e., that could either confirm or disconfirm their conclusion depending on the circumstances). A Friedman test found no statistically significant difference in the type of indicators provided by those 18 analysts in the ACH group who provided indicators, χ^2 (2, $N = 18$) = .13, $p = .936$, Kendall's $W = .004$.

5.2 | Within-individual consistency in judgement processes

Our next set of analyses measured the association between measures of within-individual consistency in judgement processes and group (ACH vs. untrained). The consistency variable was defined in three different ways.

5.2.1 | Consistency of evidence assessment

A statistically significantly smaller proportion of analysts in the ACH group (i.e., 4%, $n = 1$) applied their scoring rule consistently across evidence items compared with 44% (4 out of $n = 9$[6]) of the untrained group who used a scoring rule, χ^2 (1, $N = 34$) = 8.63, $p = .003$, $\phi = -.50$.

5.2.2 | Consistency of evidence integration

Seventy-six per cent ($n = 19$) of the ACH group applied their evidence integration strategy consistently across hypotheses compared with 80% (12 out of $n = 15$[7]) of the untrained group who used an evidence integration strategy, χ^2 (1, $N = 40$) = 0.09, $p = .769$, $\phi = -.05$.

5.2.3 | Consistency of judgements

The final conclusion reached by 64% ($n = 16$) of the ACH group matched the judgements made in their revised matrix (preceding judgements). By contrast, where it was possible to evaluate, the final conclusion presented by all of the analysts in the untrained group (out of $n = 16$[8]) was consistent with their preceding judgement process. This difference between groups was statistically significant, χ^2 (1, $N = 41$) = 7.38, $p = .007$, $\phi = -.42$.

[5]Here, $n = 20$ because we can only make this calculation for analysts who made a matrix or a list of evidence.

[6]Here, $n = 9$ because we can only make this calculation for analysts who used a scoring rule ($n = 12$). Of those, there were three analysts whose consistency could not be determined.

[7]Here, $n = 15$ because the calculation included the nine analysts who used a scoring rule and added up their scores plus the six analysts that added up the likelihood percentages for each hypothesis.

[8]This refers to the 16 analysts who used a scoring rule of some kind.

5.3 | Analysts' judgement and choice accuracy

Our final set of analyses measured the association between accuracy and group (ACH vs. untrained). Analysts' accuracy (i.e., the most likely tribe membership of the target individual) was evaluated in two different ways.

One way of measuring accuracy was on an ordinal scale (i.e., correctness of analysts' ranking of tribe membership from most to least likely). Here, only one (4%) of the 25 analysts in the ACH group produced a correct rank ordering of the four hypotheses compared with two (4.9%) of 16[9] analysts in the untrained group. This difference between groups was not statistically significant, χ^2 (1, N = 41) = .31, p = .308, ϕ = −.16. A further examination of the data revealed that a statistically significantly greater proportion (i.e., 80%, n = 20 of 25) of analysts in the ACH group had one or more tied ranks between hypotheses compared with 19% (n = 3 of 16) of the untrained group, χ^2 (1, N = 41) = 14.86, p < .001, ϕ = .60.

The other way of measuring accuracy was on a categorical/binary scale (i.e., whether analysts chose the correct tribe as the most likely). Thirty-six per cent (n = 9) of analysts in the ACH group and 33% (n = 8) in the untrained group[10] chose the correct hypothesis (i.e., Conda tribe), and this difference between groups was not statistically significant, χ^2 (1, N = 49) = .04, p = .845, ϕ = .03.

Finally, we also explored the relationship between within-individual consistency and accuracy across both groups. A McNemar test revealed that a statistically significantly greater proportion of analysts in both groups (i.e., 80%, n = 4 out of 5) who applied their evidence assessment rule consistently across evidence items were accurate in their choice of most likely tribe, compared with 31% (n = 9 out of 29) of analysts who were inconsistent, p = .021, odds ratio = 8.89. Similarly, a statistically significantly greater proportion of analysts in both groups who applied their evidence integration rule consistently across hypotheses (i.e., 38.7%, n = 12 out of 31) chose the correct tribe, compared with 22.2% (n = 2 out of 9) of those who were inconsistent, p < .001, odds ratio = 2.21.

6 | DISCUSSION

The intelligence community believes that ACH helps analysts to think critically and avoid "confirmation bias." The present study examined ACH in practice. We found that most analysts trained (and instructed) to use ACH deviated from one or more of the steps prescribed by this technique. In particular, they departed from ACH's Step 5, which refers to evidence integration (see Table A1). Past research on ACH has not measured the extent to which participants fully applied ACH. However, Trent, Voshell, and Patterson (2007) reported that army intelligence officers resisted using ACH after being trained and repeatedly instructed to do so. In fact, intelligence organizations also find themselves deviating from some of the steps prescribed by ACH. For instance, in its manual describing ACH, UK Defence Intelligence (UK Ministry of Defence, 2013, p. 15) asks analysts to consider "If this hypothesis were true, how likely would this evidence be?" Analysts must enter a score of 0 to 4, where 0 represents less than 10%, 1 represents 10–25%, 2 represents 50–75%, and 4 represents more than 75%. Then, they must add up the scores for each hypothesis. These are significant departures from ACH, and yet both analysts and their organizations would believe they are applying ACH. Clearly, future research ought to examine the efficacy of ACH as designed and if it is found to be useful then more needs to be done to persuade analysts and intelligence organizations to use it. Meanwhile, our discussion of the present findings below focuses on how ACH is used in practice.

Before we discuss the present findings, we highlight potential concerns some may raise about their external validity, given the nature of the analytic task used in the present study. Although intelligence analysts seldom face such neat problems (i.e., where all hypotheses are provided and are mutually exclusive, and where all relevant evidence is available and precisely quantified), we do not believe this implies that analysts would perform better when faced with real intelligence problems. This is because real problems are murky, unlike the present task—there may be not enough relevant data or there may be large volumes of data, the credibility of data sources may vary, the data may be formatted in different ways (e.g., structured/unstructured, textual/visual/audio), it may be ambiguous, unreliable and sometimes intentionally misleading, and there may be time pressure and high-stakes involved. We see no reason why ACH should help under these conditions when it does not help hypothesis evaluation under the more modest conditions of the present experimental task where the information available to analysts could be easily subjected to the consistency tests that ACH requires. We would expect that ACH would perform better in the simple analytic task used in the present study than in the much more complex tasks encountered by analysts in practice. Nevertheless, it would be useful to conduct future research on ACH involving a diverse set of tasks. Indeed, this could help to identify some of the conditions under which ACH does better or worse.

6.1 | Confirmation bias, consistency, and accuracy

In the context of our experiment, confirmation bias was conceptualized as follows: (a) not considering all alternative hypotheses; (b) only evaluating evidence based on whether it is consistent with each hypothesis under consideration; (c) not adjusting belief in a hypothesis in accordance with evidence diagnosticity; (d) selecting the most likely hypothesis based solely on evidence that is consistent with it; and (e) identifying indicators that will only confirm a hypothesis in the future. We found that analysts in the ACH group were no more likely than their untrained counterparts to identify the four alternative hypotheses in the present experiment. On the other hand, the ACH group were more likely to rate evidence as being either inconsistent or consistent with each hypothesis (as opposed to simply more or less consistent) and to take account of evidence diagnosticity. Both the ACH

[9]Here, n = 16 because we can only perform this analysis on those who provided written conclusions regarding all four hypotheses.

[10]One analyst did not provide any conclusion.

and untrained groups were equally likely to focus solely on consistent evidence when selecting the most likely hypothesis, although the majority of analysts in both groups selected the most likely hypothesis based on an integration of consistent and inconsistent evidence. Finally, analysts in the ACH group who provided indicators for future observation were no more likely to provide indicators that would disconfirm (as opposed to confirm) the hypotheses.

Taken from the perspective of untrained analysts, these findings reiterate that they do not all suffer from such bias, like participants in other psychological studies on "confirmation bias" (e.g., Beattie & Baron, 1988). Nevertheless, it seems apt that analysts may need explicit instructions to differentiate between evidence that is consistent versus inconsistent with a hypothesis and to remove nondiagnostic information from their "working out" (Kemmelmeier, 2004), especially when it may lead to the "dilution" of judgements based on diagnostic information (Shelton, 1999). The fact that ACH is vague on these two issues means that it has limited value in this regard.

Taken from the perspective of analysts trained to use ACH, the above findings highlight that not only may analysts resist applying its evidence integration rule but also they prefer to use (like their untrained counterparts) a cognitively more complex strategy (i.e., adding up both consistent and inconsistent evidence for each hypothesis). The strategy used by most of the present analysts is beneficial because there is no "loss" of relevant information and the credibility of all available evidence (rather than just the disconfirming evidence) can be taken into account.

Perhaps one benefit of any sort of structured analytic technique such as ACH is that it can make the analytic process more transparent and easier to manage and audit by increasing within-individual inconsistency. However, we found that the ACH group demonstrated significantly less consistency in terms of evidence assessment and the match between final conclusions and preceding judgements, compared with their untrained counterparts. A large proportion of analysts in both groups also applied their evidence integration strategy inconsistently across hypotheses. Inconsistency in evidence assessment may be partly explained by the fact that, although ACH asks analysts to distinguish between evidence that is highly inconsistent or inconsistent (vs. highly consistent or consistent) with a hypothesis, it does not specify how this should be done. These results support recent warnings about how structured analytic techniques, in general, can foster inconsistency in assessments (Chang et al., 2018; Mandel & Tetlock, 2018). Decision-support tools may be useful in this domain because they can reduce the cognitive burden on analysts. Reducing inconsistency is important because it is difficult to identify the source of error if an analyst is behaving inconsistently. Increasing consistency is also important because, as we found (across both groups), it was associated with the accuracy of conclusions reached.

Indeed, one could argue that the ultimate goal of analysts is to arrive at an accurate conclusion about a current or future situation. However, we found that only one of the ACH group correctly ranked the four hypotheses from most to least likely and two of the untrained group did so. Analysts in the ACH group were significantly more likely than their untrained counterparts to produce tied ranks between hypotheses, partly because ACH encourages analysts to reduce probabilistic (continuous) data regarding consistency or inconsistency to a 5-point ordinal scale. Unsurprisingly, the ACH group was no more likely than the untrained group to choose the correct hypothesis (also see Mandel, Karvetski, & Dhami, 2018).

6.2 | Other findings on how analysts test competing hypotheses

Several other findings emerged that shed some light on how analysts may solve a hypothesis testing task. First, the majority of untrained analysts reformatted the data in the task. Over half of this group drew an ACH-style matrix with hypotheses as columns and evidence as rows. It is unclear if this format is helpful. Psychological research suggests that the way in which information is formatted can aid or hinder information processing in a range of cognitive tasks (e.g., Garcia-Retamero & Dhami, 2011, 2013; Gigerenzer & Hoffrage, 1995). Future research ought to systematically examine the effects of ACH's recommended matrix format on analysts' hypothesis testing compared with alternative information formats. Cook and Smallman (2008) found that graphical information displays reduced the attention that naval personnel paid to confirming evidence.

Second, although ACH is unclear about how analysts should assess evidence diagnosticity, we observed a correlation between an objective measure of information diagnosticity and judgements of diagnosticity made by individual analysts in the ACH group as well as across analysts in the untrained group. It would, however, be premature to suggest that people may have some "intuitive" capacity to judge diagnosticity since a variety of strategies can be correlated with objectives measures such as the one we used here (i.e., information gain). Future research could more fully explore analysts' strategies for judging information diagnosticity against other existing measures (see Nelson, 2005).

Finally, as mentioned earlier, ACH does not take account of base rate information, and unsurprisingly, we found that analysts in the ACH group were significantly less likely to do so compared with their untrained counterparts. Nevertheless, only around half of the untrained group used base rate information. Base rate neglect is common (e.g., Kahneman & Tversky, 1973; Tversky & Kahneman, 1982). Base rate information is useful because it provides an indication of the priori probability of a hypothesis being true before being presented with any evidence. In the present study, such information was useful for arriving at the correct conclusion because of the inequality in base rates for the four hypotheses.

Some believe that ACH may be particularly useful for collaborative analysis, where it can provide analysts a better understanding of differences of opinion, depersonalize issues, and guide discussion (Heuer, 2007). However, there is, as yet, little empirical evidence to support this view. In Convertino et al.'s (2008) study, reviewed earlier, all groups used a collaborative version of ACH, and yet "confirmation bias" remained evident in all groups (i.e., here, evidence initially favoured one hypothesis but then balanced out across hypotheses

later on). Clearly, more research is needed to test the benefits of ACH when applied in a collaborative context.

6.3 | Alternatives to ACH

Given the paucity of research supporting the efficacy of ACH, it may be prudent for the intelligence community to consider alternatives. Some have suggested that ACH may be improved by supplementing it with other methods (e.g., Karvetski, Olson, Gantz, & Cross, 2013; Murukannaiah et al., 2015; Wheaton & Chido, 2006). For instance, Murukannaiah et al. (2015) added argumentation schemes, in what they call Arg-ACH, to elicit users' conclusions, underlying premises and critical questions for assessing the argument. In a study of 20 students who were trained to use tools that implemented either ACH or Arg-ACH, it was found that the latter group performed better in terms of, for example, the completeness and coverage of their belief search, the explicitness of the assumptions they made, and the repeatability of their reasoning. However, it is unclear how this or other ACH "add ons" would reduce confirmation bias or improve judgement accuracy. Given that one of the rationales for developing ACH was the desire to reduce confirmation bias, we suggest that there are more psychologically informed and better empirically tested alternatives to ACH for reducing confirmation bias, as well as statistically based alternatives to hypothesis testing more generally.

Several variations of the "consider-the-opposite" strategy have been reported to reduce confirmation bias. For instance, Lord, Lepper, and Preston (1984) found that instructing individuals to imagine their response if specific evidence pointed in the opposite direction reduces the tendency to discount conflicting evidence. In addition, presenting individuals with conflicting evidence in advance of their search for information reduces search for supporting evidence. Similarly, Williams and Mandel (2007) found that probability judgements were more coherent and accurate if queries made the complement of the judged event explicit (i.e., probability of *x* rather than *not x*).

Computer-based tools such as serious games have been shown to reduce confirmation bias. Morewedge et al. (2015) reported that a single training session involving playing an interactive video game led to debiasing effects immediately afterwards and for at least 2 months later. The game measured player's degree of confirmation bias (e.g., by gathering and interpreting evidence in a manner confirming rather than disconfirming the hypothesis being tested), provided them with information explaining the bias along with examples, and provided opportunities for practice and personalized feedback.

Bayesian reasoning has previously been recommended to the intelligence community (e.g., Burns, 2015; Karvetski et al., 2013; Svenson et al., 2010). Analysts can update their belief in the prior probability of a hypothesis being true (on a 0 to 1 scale) based on incoming evidence and compute a posterior probability. The prior may be an objective base rate, .5, or it may be based on subjective knowledge. The updating is done using Bayes' rule, which states that the posterior is the product of a prior and a likelihood (i.e., the probability of some evidence being observed if the hypothesis is true). When applied iteratively (or when updated), the posterior becomes the new prior. Bayesian reasoning enables analysts to take account of all of the available evidence as it emerges in a precise way and avoid base rate neglect (Mandel, 2015; see also Sedlmeier & Gigerenzer, 2001). However, Bayesian reasoning can be complex and may require decision support.

Regardless of the alternatives to ACH that may be pursued, the present study shows the importance of developing and using an evidence-base to inform decisions about the best analytic practice. An evidence-based approach not only enhances the performance of individual analysts and consequently the organizations in which they work but also supports more effective decision making to tackle security threats.

ACKNOWLEDGEMENTS

The work presented in this paper was supported by funding provided to the first author by HM Government and to the third author by Canadian Department of National Defence (project no. 05da) and Canadian Safety and Security Program (project no. CSSP-2016-TI-2224). The work also contributes to NATO System Analysis and Studies Panel Research Technical Group 114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making. We thank Jonathan Nelson for his advice on aspects of the work and Jeremy Brown and Shoshannah Harper for their research assistance.

ORCID

Mandeep K. Dhami https://orcid.org/0000-0001-6157-3142

REFERENCES

Bar-Hillel, M. (1980). The base-rate fallacy in probabilistic judgments. *Acta Psychologica*, *44*(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Beattie, J., & Baron, J. (1988). Confirmation and matching biases in hypothesis testing. *The Quarterly Journal of Experimental Psychology*, *40*(2), 269–297. https://doi.org/10.1080/02724988843000122

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

Burns, K. (2015). Bayesian HELP: Assisting inferences in all-source intelligence. Paper presented at the AAAI 2015 Fall Symposium on Cognitive Assistance in Government and Public Sector Applications, Arlington, VA. Retrieved fromhttp://www.aaai.org/ocs/index.php/FSS/FSS15/paper/view/11681

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

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 (CSCW)*, *17*(4), 353–393. https://doi.org/10.1007/s10606-008-9080-9

Cook, M. B., & Smallman, H. S. (2008). Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(5), 745–754. https://doi.org/10.1518/001872008X354183

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

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science*, *10*(6), 753–757. https://doi.org/10.1177/1745691615598511

Garcia-Retamero, R., & Dhami, M. K. (2011). Pictures speak louder than numbers: On communicating medical risks to immigrants with non-native language proficiency. *Health Expectations*, *14*(suppl. 1), 46–57. https://doi.org/10.1111/j.1369-7625.2011.00670.x

Garcia-Retamero, R., & Dhami, M. K. (2013). On avoiding framing effects in experienced decision makers. *Quarterly Journal of Experimental Psychology*, *66*(4), 829–842. https://doi.org/10.1080/17470218.2012.727836

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Heuer, R. J. (1999). *The psychology of intelligence analysis*. Washington, DC: CQ Press.

Heuer, R. J. (2005). How does analysis of competing hypotheses (ACH) improve intelligence analysis? Retrieved from http://www.pherson.org/wp-content/uploads/2013/06/06.-How-Does-ACH-Improve-Analysis_FINAL.pdf

Heuer, R. J. (2007). The future of 'alternative analysis'. Retrieved http://www.pherson.org/wp-content/uploads/2013/06/04.-Future-of-Alternative-Analysis_FINAL.pdf

Heuer, R. J., & Pherson, R. H. (2014). Structured analytic techniques: A new approach to analysis. *Analyzing Intelligence*, 231–248.

Jervis, R. (2006). Reports, politics and intelligence failures: The case of Iraq. *The Journal of Strategic Studies*, *29*(1), 3–52. https://doi.org/10.1080/01402390600566282

Jones, N. (2017). Critical epistemology for analysis of competing hypotheses. *Intelligence and National Security*., *33*, 273–289. https://doi.org/10.1080/02684527.2017.1395948

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/10.1037/h0034747

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*(3–4), 205–231. https://doi.org/10.1007/s40070-013-0001-x

Kemmelmeier, M. (2004). Separating the wheat from the chaff: Does discriminating between diagnostic and nondiagnostic information eliminate the dilution effect? *Journal of Behavioral Decision Making*, *17*(3), 231–243. https://doi.org/10.1002/bdm.473

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32*, 385–418. https://doi.org/10.1016/S0079-7421(08)60315-1

Kretz, D. R., & Granderson, C. W. (2013). An interdisciplinary approach to studying and improving terrorism analysis. In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, 157–159. https://doi.org/10.1109/ISI.2013.6578808

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. In *Homeland Security (HST), 2012 IEEE Conference on Technologies for*, 439–444. IEEE. https://doi.org/10.1109/THS.2012.6459889

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*(3), 584–592. https://doi.org/10.1109/TSMCA.2008.918634

Lehner, P. E., Adelman, L., DiStasio, R. J., Erie, M. C., Mittel, J. S., & Olson, S. L. (2009). Confirmation bias in the analysis of remote sensing data. *IEEE Transactions on Systems Man and Cybernetics - Part a Systems and Humans*, *39*(1), 218–226. https://doi.org/10.1109/TSMCA.2008.2006372

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231–1243. https://doi.org/10.1037/0022-3514.47.6.1231

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology*, *6*, 387. https://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, D.C.: Georgetown University Press.

Mandel, D. R., Karvetski, C., & Dhami, M. K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision making*, *13*(6), 607–621.

Mandel, D. R., & Tetlock, P. E. (2018). Correcting judgment correctives in national security intelligence. *Frontiers in Psychology*, *9*, 2640. https://doi.org/10.3389/fpsyg.2018.02640

Marrin, S. (2008). Training and educating U.S. intelligence analysts. *International Journal of Intelligence and Counterintelligence*, *22*, 131–146. https://doi.org/10.1080/08850600802486986

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights From the Behavioral and Brain Sciences*, *2*(1), 129–140. https://doi.org/10.1177/2372732215600886

Murukannaiah, P. K., Kalia, A. K., Telang, P. R., & Singh, M P. (2015). Resolving goal conflicts via argumentation-based analysis of competing hypotheses. IEEE 23rd International Requirements Engineering Conference (RE), 156–165.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–999. https://doi.org/10.1037/0033-295X.112.4.979

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Pherson Associates, LLC (n.d.). ACH. Retrieved from http://www.pherson.org/ach/

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

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*(3), 380–400. https://doi.org/10.1037//0096-3445.130.3.380

Shelton, S. W. (1999). The effect of experience on the use of irrelevant evidence in auditor judgment. *The Accounting Review*, *74*(2), 217–224. https://doi.org/10.2308/accr.1999.74.2.217

Svenson, P., Forsgren, R., Kylesten, B., Berggren, P., Fah, W. R., Choo, M. S., & Hann, J. K.Y. (2010). Swedish-Singapore studies of Bayesian modelling techniques for tactical intelligence analysis. In 13th Conference on Information Fusion (FUSION) 2010. 1–8.

Trent, S., Voshell, M., & Patterson, E. (2007). Team cognition in intelligence analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(4), 308–312. Sage CA: Los Angeles, CA: SAGE Publications, https://doi.org/10.1177/154193120705100434

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty:*

Heuristics and biases (pp. 153–160). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.011

UK Ministry of Defence. (2013). *Quick wins for busy analysts*. UK: Published by Defence Intelligence.

U.S. Government. (2009). A tradecraft primer: Structured analytic techniques for improving intelligence analysis. Retrieved from: https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/Tradecraft%20Primer-apr09.pdf

Wheaton, K., & Chido, D. (2006). Structured analysis of competing hypotheses: Improving a tested intelligence methodology. *Competitive Intelligence Magazine*, *9*(6), 12–15.

Williams, J. J., & Mandel, D. R. (2007). Do evaluation frames improve the quality of conditional probability judgment? In D. S. McNamara, & J. G. Trafton (Eds.), *Proceedings of the 29th annual meeting of the cognitive science society* (pp. 1653–1658). Mahwah, NJ: Lawrence Erlbaum Associates.

How to cite this article: Dhami MK, Belton IK, Mandel DR. The "analysis of competing hypotheses" in intelligence analysis. *Appl Cognit Psychol*. 2019;1–11. https://doi.org/10.1002/acp.3550

APPENDIX A

TABLE A1 ACH steps

1	Identify all possible, mutually exclusive hypotheses.
2	List evidence relevant for evaluating each hypothesis.
3	Prepare a matrix with hypotheses as columns and evidence as rows, and rate each evidence item as consistent, inconsistent or not applicable to each hypothesis.[a] Record any assumptions underlying these ratings. Sort the evidence by diagnosticity[a] (or credibility and relevance).[b]
4	Refine the matrix by excluding nondiagnostic or insufficiently diagnostic evidence and reconsidering the hypotheses. If new hypotheses are added then restart the process.
5	Draw tentative conclusions on the relative likelihood of each hypothesis by adding up the number of inconsistency ratings for each hypothesis.[c] The hypothesis with the most inconsistent rating is least likely whereas the hypothesis with the fewest is most likely.
6	Analyse the sensitivity of the conclusions to critical evidence items and consider the consequences of any assumptions underlying the evidence.
7	Report the conclusions.
8	Identify indicators for future observation that would support or contest the conclusion.

[a]Evidence is rated as "highly consistent," "consistent," "inconsistent," "highly inconsistent," or "not applicable."[abc]
[ab]This refers to a consideration of how useful the evidence is for discriminating among alternative hypotheses.
[bc]In ACH, the credibility and relevance of the evidence can each be rated as "low," "medium," or "high."
[cd]ACH requires analysts to ignore the consistency ratings. An evidence item that is inconsistent with a hypothesis is attributed 1 point and an item that is highly inconsistent is attributed 2 points. Evidence credibility is taken into account by adjusting the inconsistency scores so that evidence of low credibility has .3 or .6 deducted from it depending on if it is inconsistent or highly inconsistent with the hypothesis, respectively. Evidence of high credibility has .4 or .8 added to it depending on if it is inconsistent or highly inconsistent, respectively. Evidence of medium credibility has no adjustment made to it.

TABLE A2 Analytic task properties

	Percentage of likelihood of feature in tribe (base rate)				
Features	**Acanda (5)**	**Bango (20)**	**Conda (30)**	**Dengo (45)**	**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

TABLE A3 Response sheet used for ACH group

Step 1	**Identify all possible hypotheses. These should be mutually exclusive.**
Step 2	Make a list of significant information/evidence that is relevant for evaluating the hypotheses, including assumptions and the absence of things one might expect if the hypothesis were true.
Step 3	Create a matrix with all the hypotheses across the top and all items of relevant information down the left side. Then, analyse each piece of information by asking if it is Consistent or Inconsistent with the hypothesis or if it is Not Applicable or irrelevant. This can be done by filling each cell of the matrix row-by-row with "C," "I," or "NA." You can put two "CCs" or two "IIs" if the information is particularly compelling. The ratings will likely depend on some assumptions, and if so, then record those assumptions in another column, row-by-row.
Step 4	Think about how the matrix may need revising. To do this, sort the information for diagnosticity (i.e., which items of information are most helpful in comparing hypotheses). Consider how much confidence you have in the assumptions for the highly diagnostic Inconsistent ratings, and readjust the ratings accordingly. Delete the rows with nondiagnostic information. Reconsider the hypotheses and decide if any need combining or any if new ones need to be added. Finally, rate the information for the combined or new hypotheses, again making note of any assumptions. You will need to redraw and update the matrix.
Step 5	Draw tentative conclusions about the relative likelihood of each hypothesis based on the diagnosticity of each item of information. Do this by adding up the number of Inconsistent ratings for each hypothesis to give an "Inconsistency Score" for each hypothesis. Then, rank the hypotheses so that the highest rank is given to the one with the lowest inconsistency score. The hypothesis with the lowest inconsistency score is tentatively the most likely hypothesis and the hypothesis with the highest inconsistency score is usually the least likely.
Step 6	Analyse the sensitivity of your tentative conclusion to a change in the interpretation of a few critical items of relevant information. If one or more of these items were wrong, misleading or subject to a different interpretation will your conclusion need to change? If so, then go back and double-check the accuracy of your interpretation.
Step 7	Report your conclusions. 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. Also say why alternative hypotheses were rejected.
Step 8	Identify indicators or milestones for future observation. Create two lists—one focusing on future events or access to additional information that would support your conclusion, and one list focusing on events and information that would suggest your conclusion is less likely to be correct or that the situation has changed.