Authors' Submitted Version

Jennifer Stromer-Galley, Patricia Rossini, Kate Kenski, Brian McKernan, Benjamin Clegg, James Folkestad, Carsten Østerlund, Lael Schooler, Olga Boichak, Jordan Canzonetta, Rosa Mikeal Martey, Corey Pavlich, Eric Tsetsi & Nancy McCracken (2021). Flexible versus structured support for reasoning: enhancing analytical reasoning through a flexible analytic technique, Intelligence and National Security, 36:2, 279-298, DOI: 10.1080/02684527.2020.1841466

Flexible versus Structured Support for Reasoning: Enhancing Analytical Reasoning Through a Flexible Analytic Technique

Jennifer Stromer-Galley, corresponding author, Syracuse University, jstromer@syr.edu, https://orcid.org/0000-0001-6079-8788
Patricia Rossini, Syracuse University, https://orcid.org/0000-0002-4463-6444
Kate Kenski, University of Arizona
Brian McKernan, Syracuse University
Benjamin Clegg, Colorado State University, https://orcid.org/0000-0001-6026-5076
James Folkestad, Colorado State University, https://orcid.org/0000-0003-0301-8364
Carsten Østerlund, Syracuse University, https://orcid.org/0000-0003-0612-1551
Lael Schooler, Syracuse University, https://orcid.org/0000-0003-2964-2886
Olga Boichak, Syracuse University, https://orcid.org/0000-0001-6547-2015
Jordan Canzonetta, Syracuse University
Rosa Mikeal Martey, Colorado State University
Corey Pavlich, University of Arizona
Eric Tsetsi, University of Arizona
Nancy McCracken, Syracuse University, https://orcid.org/0000-0002-1267-3443

Title: Flexible versus Structured Support for Reasoning: Enhancing Analytical Reasoning Through a Flexible Analytic Technique

Abstract

Structured analytic techniques (SATs) have been developed to help the intelligence community reduce flaws in cognition that lead to faulty reasoning. To ascertain whether SATs provide benefits to reasoning we conducted an experiment within a web-based application, comparing three conditions: 1) unaided reasoning, 2) a prototypical order-based SAT and 3) a flexible, process-based SAT that we call TRACE. Our findings suggest that the more flexible SAT generated higher quality reasoning compared to the other conditions. Consequently, techniques and training that support flexible analytical processes rather than those that require a set sequence of steps may be more beneficial to intelligence analysis and complex reasoning.

Acknowledgements

The authors would like to thank Sarah Taylor and Roc Myers for their subject matter experience of intelligence analysis and structured analytic techniques. The experimentation software was developed by SRC., Inc. We wish to thank Deb Plochocki, Lou Nau, Andrew Whalen, Laura Simonetta, Ryan Conner, and the rest of the SRC team. We also acknowledge the contributions of several additional team members including: Yatish Hegde, Niraj Sitaula, Sarah Bolden, Erin Bartolo, Jerry Robinson, Jun Fang, Priya Harindranathan, Marcia Morales, Thomas Gallegos, Paige Odegard, Gregory Russel, Rhema Zlaten, Cayla Dorsey, Sophie Estep, Audrey Lew, Quincy Nolan, Sweeney Pandit, and William Wang. This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior

National Business Center contract number 2017-16121900004. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

Introduction

Analytical reasoning involves carefully considering sources or problems and coming to a judgement, and it is a commonplace activity in many settings—from the corporate boardroom to government intelligence analysis to policy making. This type of reasoning is demanding because it entails the evaluation of large quantities of complex information. It can also be susceptible to different types of cognitive biases, where reliance on heuristics to reach conclusions can lead to problematic decision-making,[1] or failures due to cognitive overload that affects working memory.[2]

The challenging nature of analytical reasoning and concerns about critical errors have led to the development of techniques aimed at improving such complex reasoning. Richards Heuer, a CIA officer and pioneer who used cognitive psychology research to inform intelligence analysis,[3] along with Randolph Pherson, a former intelligence analyst, offered several structured analytic techniques (SATs) with the intent of improving intelligence reasoning.[4]

One of the central SATs that Heuer advanced was Analysis of Competing Hypotheses (ACH).[5] Originally developed to help with deception analysis, ACH has come to be one of the most widely known SATs within the intelligence community (IC).[6] As an indication of its prominence, Jones notes that ACH is one of the select few SATs that the United States Government included in its *Tradecraft Primer*.[7] ACH is designed to reduce cognitive biases and other errors by aiding analysts in the identification of all relevant hypotheses and in evaluation of supportive and non-supportive evidence for each hypothesis using a matrix with the goal of ranking the identified hypotheses based on how much they are supported by available information. Despite the awareness of ACH, there is little empirical research as to the effectiveness of order-based analytic techniques, and the few studies that do such tests provide mixed results.[8] Along with other SATs, ACH is often

the subject of criticism for being too burdensome and inefficient, and as such it is not widely adopted by the IC.[9]

To address the lack of rigorous testing of SATs, and especially the underlying presupposition that engaging in a set of prescribed steps can improve reasoning outcomes, we conducted experiments with novice analysts using two types of SATs, one that was a prototypical ordered SAT similar to ACH, and another that provided a more flexible and wider array of tools designed to aid complex reasoning and reduce cognitive load called Trackable Reasoning and Analysis using Collaboration and Evaluation (TRACE).

ACH and other SATs are devised with prescribed processes and steps, but that structure may be a source of trouble for analysts. Heuer and Pherson proposed that doing an analysis in a specific sequence with a mandated set of steps would be more effective at reducing errors, especially cognitive biases, than unaided reasoning.[10] In interviews we conducted with members of the IC, analysts identified the sequenced nature of these techniques as a problem because they disrupt fluid reasoning and can even add to cognitive load and possibly unintentionally induce cognitive biases.[11]

To set up our experimental conditions, we consulted with subject matter experts to design a simplified analogue of ACH, which we refer to as Evidence Alignment Assessment (EAA). EAA includes core ACH features, such as the creation of a matrix to connect hypotheses to evidence and to rank their likelihood of correctness but reduces the required steps to complete the analysis.

In contrast to EAA's mandated sequence of steps, the TRACE condition provides a more flexible analytic environment where analysts have a large set of tools designed to aid different types of analysis at different stages of the analysis process. It does not impose a structure, an order, or a set of requirements for the analytical process. Analysts are thus left to choose what tools and

aids to use as the problem and their own reasoning needs mandate. These two analytic approaches (EAA and TRACE) were compared to a control condition of unaided reasoning, in which participants did not have access to any software to help problem-solving.

This study conducted a controlled experiment to test the efficacy of SATs. Our results suggest that the more flexible analytic technique, TRACE, significantly outperforms the structured EAA, and that EAA only modestly outperforms the control condition of unaided reasoning. The findings indicate that a more flexible analytical environment in which individuals can engage analytical tools as they need them—tools that promote critical thinking and aid in memory and recall–-can be more effective than a sequenced structure of analysis with a limited set of analytical aids.

Structured analytic techniques in the intelligence community

The IC has long been interested in the development of techniques and technologies to improve the quality and accuracy of reasoning and decision-making.[12] Following the acknowledged failure of the IC in concluding that Iraq had weapons of mass destruction (WMD) capabilities, the Robb-Silberman Commission reported that the most significant failure was in poor overall analytic process, also called tradecraft, and analytical shortcomings.[13] The Commission noted that "the most damaging tradecraft weakness we observed was the failure of analysts to conclude – when appropriate – there was not enough information available to make a defensible judgement."[14] As an example, the Duelfer report (also known as the Iraq Survey Group report) concluded that analysts did not understand Saddam Hussein and his motives and looked for patterns of behaviour and strategy when in fact there were none.[15] The Commission recommended several items, including: more strategic intelligence; greater transparency in reporting and more detailed information about sources that inform analysis; using more supportive aids for reasoning and decision-making, such as red teaming; improving information searching and knowledge extraction; and continual training to establish better standards for analysis. This led the IC to further invest in tools and training to support tradecraft techniques that help reduce the types of errors that were on display in the Iraq WMD analysis.

To facilitate the integration of SATs with analysts' routines, there have been efforts to develop software to implement them and encourage their use. For instance, argument mapping software enables analysts to create a visual representation of key argumentative components, such as premises, warrants, and evidence, to aid the construction of sound argumentation and the identification of strengths and weaknesses in reasoning. Research suggests that argument mapping software can help improve reasoning and argumentation skills.[16] The caveat, however, is that

argument mapping software places heavy cognitive demands on users that make them unappealing and difficult to use.[17] There have been several attempts to create a software-based version of ACH.[18] These include: simple solutions such as a spreadsheet-based matrix, which only partially aids the ACH process; to more complex solutions, such as a software developed by the proponent of the method, Richards Heuer and the Palo Alto Research Center,[19] which allows analysts to work on multiple projects and provides a ranking of hypotheses based on three different algorithms.[20]

Despite the efforts to develop SATs to aid reasoning and decision-making, most of these techniques are rarely adopted by the IC. For instance, a recent study found that only 18 reports from a sample of 63 reports from the CIA, the National Intelligence Council, and the Defense Intelligence Agency showed evidence of SAT use.[21] Additionally, in a survey of eighty analysts working in the US State Department's Bureau of Intelligence and Research (INR), Coulthart found that one third of the analysts never used SATs and the remainder used SATs only rarely to sometimes.[22]

Some scholars suggest that a lack of training is the cause of low uptake of SATs.[23] Others suggest that even with training analysts view structured techniques as burdensome, inefficient, and time-consuming because they interrupt the natural analytical process that analysts have traditionally been trained on and are used to.[24] These troubles, however, might be mitigated with more effective training.[25] Moreover, some scholars have suggested that new analysts, who tend to be younger, may be more receptive to training and using SATs.[26] Research by Coulthart, however, suggests that years of IC experience is not related to whether an INR analyst uses SATs and that receiving training on these techniques is the most influential factor.[27]

Other scholarship suggests that analysts do not consider SATs to be suitable for their line of work. Based on an ethnographic study of the IC, Johnston finds that analysts consider more

formal and ordered techniques to be inefficient given the difficult time constraints of their work.[28] On the other hand, Coulthart finds that time pressure was not related to whether INR analysts use SATs and consequently suggests that the time pressure analysts face may vary by agency.[29] Marrin's research indicates that many intelligence scholars believe that more rigorous or scientific techniques such as SATs are inappropriate for intelligence analysis given that analysts work with incomplete and often deceptive data.[30] As well, there is some criticism that SATs may negatively affect the quality of reasoning because they disrupt fluid intelligence and analogical reasoning.[31]

Along with the lack of adoption of SATs within the IC, judgment and decision science researchers have argued that there is not enough research that systematically tests the efficacy of SATs in meeting these goals,[31] and the studies that have examined them have so far found mixed support.[32] In a more general critique to the use of SATs, Chang et al. argue that these techniques are unlikely to effectively mitigate cognitive biases because they propose unipolar solutions when most cognitive biases are bipolar.[33] For example, efforts to counter an analyst's overconfidence in their judgement may cause the analyst to become under-confident.

SATs also do not address another problem for the IC, which is the writing of the analytical product. As noted earlier, they can lack clarity or fail to provide support for their judgements. This makes it challenging for consumers of these products, which are decision-makers or other analysts, to fully understand the analytical process that led to a particular judgement, thus missing one of the recommendations of the Robb-Silberman commission.[34]

Designing and implementing structured techniques

Our team set out to devise a potentially more effective SAT to test using social science experimentation methods, that we call TRACE (Trackable Reasoning and Analysis using Collaboration and Evaluation). We experimentally tested TRACE against EAA and a control

condition of unaided reasoning. In this section, we review the research on ACH, describe the version of EAA we designed for this experiment, and introduce TRACE.

Analysis of competing hypotheses

ACH was originally developed by Heuer and involved eight sequenced steps, including a thorough identification of all validly possible hypotheses, a listing of significant evidence, and an evaluation of the supportiveness of the evidence for each hypothesis [See Figure 1].[35]

[Figure 1 here]

According to Heuer, ACH provides an analytical process that is markedly different from what he describes as the unaided, conventional approach most analysts follow.[36] Whereas analysts often start by first identifying what they consider to be the most likely answer or hypothesis and then examining the available information, ACH instructs analysts to instead begin by identifying the "full set of alternative possibilities."[37] Furthermore, ACH directs analysts to evaluate the significance of each piece of evidence and argument for every hypothesis. During this step, the evidence that supports several hypotheses are of little diagnostic value and should be discarded. Finally, whereas analysts often base their final judgements on the hypothesis that has the most supporting evidence, ACH instead instructs analyst to select the hypothesis with the least amount of disconfirming evidence. Heuer explains that an analyst can never fully "prove" that a single hypothesis is true, since much of the supporting evidence may support alternative hypotheses as well. However, a single piece of disconfirming evidence may suffice to reject a hypothesis. Overall, Heuer argues that ACH's process of identifying all of the hypotheses at the outset, mapping all of the information to every hypothesis, and prioritizing disconfirming evidence helps analysts avoid several critical errors, such as confirmation bias and anchoring bias.

Although ACH has received prominent endorsements and is one of the most widely studied SATs in the field of intelligence analysis, it has also received serious critiques from practicing analysts and scholars. For instance, both Jones and Chang et al. criticize ACH for failing to provide clear instructions for key steps in the ACH process.[38] Jones argues that ACH does not specify how to correctly individuate evidence.[39] For example, if two different witnesses claim to have seen the same event, Jones claims that ACH is unclear if this should be treated as one piece of evidence or two separate pieces of evidence. Similarly, Chang et al. argue that ACH does not provide clear and concrete instructions for how to correctly evaluate the evidence.[40] For both Jones and Chang et al., ACH's lack of clear instructions inserts a degree of subjectivity into the process that can result in two different analysts deriving drastically different conclusions.[41]

These obstacles to ACH's adoption by working analysts may not be insurmountable if proponents for the technique could point to research findings that demonstrate ACH's efficacy. There have been a few empirical studies on ACH, however, and the results are mixed.[42] Lehner et al. found that ACH did not reduce anchoring bias,[43] which is the tendency for future reasoning to be inordinately influenced by exposure to past information,[44] but was effective in reducing confirmation bias, which is the tendency to seek information that confirms hypotheses and discount disconfirming information among participants without professional analysis experience. In a small pilot study, Kretz et al. examined the effect different SATs have on confirmation bias and anchoring bias among participants with no analytical experience.[45] Although participants in the ACH condition generated more hypotheses in certain phases of the experiment, there was no universal pattern across the experiment, indicating that none of the SATs were superior at bias reduction. Billman et al. experimentally tested a collaborative software-based version of ACH they developed called CACHE.[46] They found that CACHE significantly reduced confirmation

bias among participants in groups whose members possessed different initial beliefs and participants that worked by themselves relative to groups whose members possessed similar initial beliefs.[47] This study did not include a control condition, and thus their results do not provide an indication of CACHE's efficacy relative to unaided reasoning.[48] Whitesmith found that participants trained and instructed to use ACH did not exhibit lower levels of confirmation bias relative to participants in a control condition that received no such guidance to complete an analytical task.[49] Mandel et al. found that intelligence analysts trained and instructed to use ACH to complete an analytical task performed slightly worse on evaluating the probability of each hypothesis than analysts not trained and not instructed to use ACH (the control) to complete the same analytical task.[50] Finally, Dhami et al. analyzed the results of Mandel et al.'s experiment and found that participants in the ACH condition were less likely to consider base rate information as well as less likely to evaluate evidence consistently relative to participants in the control group.[51] Moreover, the majority of the participants in the ACH condition did not correctly follow all of the steps despite receiving training in ACH.[52]

Overall, the current findings from the experimental studies on ACH do not provide a clear indication of potential efficacy. Moreover, all of the experimental research on ACH thus far has focused solely on how ACH impacts *judgement accuracy*, often employing problems for which the sensitivity, reliability, and validity are unknown and thus may not provide ideal assessments.[53] None of these prior studies have examined whether or not analysts that use ACH compose clearer, more complete analytic products that better justify the judgment, articulate assumptions, and note gaps in information that might affect the judgment. Due to these limitations in the existing literature, more systematic empirical investigation of ACH and other

SAT approaches is warranted, especially given that the IC devotes resources to training analysts in ACH and implementing it as part of the standard and expected analytic tradecraft.

The implementation of a simplified analogue of ACH: Evidence Alignment Assessment (EAA)

To begin to address the dearth of rigorous experimentation, we designed a more simplified analogue of ACH – Evidence Alignment Assessment (EAA) – to examine the effect of SATs on both accuracy of reasoning and quality of analytical report writing. EAA was designed to provide the core elements of ACH, such as the generation of multiple hypotheses and the evaluation of supportive and non-supportive evidence, but also to reduce the analytical steps to avoid unnecessary burden and repetition. The design of this simplified version was informed by the following sources. First, prior research suggests that ACH may be confusing and challenging to use, as we described previously. Second, we conducted interviews with nine current or former intelligence analysts to learn their workflows and processes, and our EAA design was based on the interview findings. Third, our experimentation was constrained to untrained or novice analysts. Consequently, we designed EAA with the goal of providing participants with a SAT that kept the core elements of ACH yet was accessible to people with minimal to no prior professional analytical experience.

Thus, we developed a simplified version that provided the most central elements of ACH, such as hypothesis generation, and identifying key evidence that might confirm and disconfirm those hypotheses. This approach enables examination of whether these foundational elements of ACH are associated with improvements in reasoning. EAA has six steps. The first is an orientation step that introduces the analyst to the problem statement. The second step is a hypothesis generation step where they are prompted to generate several alternative hypotheses about the problem. This step comes before a review of the source materials to help encourage

creative thinking, reduce the chances that the source materials restrict the potential hypotheses generated, and also helps mitigate cognitive biases, such as confirmation bias **(see Figure 2)**. The third step encourages a critical reading of each source, including assessing the credibility of information and the relevance of it to the problem. The fourth step shows the hypotheses again which were generated in step two and encourages another critical review of the hypotheses given a deeper understanding of the problem space and the related source materials. The fifth step is similar to ACH in that the analyst considers each hypothesis and the source materials and how supportive or contradictory the information is to the problem (**see Figure 3**). The last step shows the analyst each hypothesis ranked by how supportive the source materials are for that hypothesis, using a formula that weights the relative supportive evaluation, along with the analysts' scoring of the source's relevance and credibility. The analyst can choose to further re-order or reconsider the source materials based on this ranking. Although EAA is a simplified version of ACH, it still instructs users to follow an order to conduct their analysis, although we designed EAA such that users could skip steps or go back to a prior step if they so chose.

[Figure 2 here]

[Figure 3 here]

From structure to flexibility: TRACE

To address some of the shortcomings of SATs identified via our interviews and the existing literature, we designed the TRACE application to provide a flexible reasoning approach that supports users through the entire analytical process: from evaluating source materials, to critically reasoning through the problem and rendering a judgement, to writing the analytical report. We also designed TRACE to reduce users' over-reliance on (faulty) memory, to reduce

the likelihood that cognitive biases were driving the analysis, and to help users manage cognitive load. Thus, TRACE offers mechanisms to support the process of reasoning while not demanding a prescribed order to activities as SATs do.

Within the TRACE application, users can "understand the task" by answering a short series of questions about the problem and receive recommendations as to how to analyse it and what tools in TRACE to use. They can review the analytical task, what we call the "background," examine all of the source materials and rank, rate, and annotate them, and use a variety of tools to aid critical thinking about the problem, and compose and review their report. The core of the TRACE application is the report. The TRACE report provides users with a template that breaks down the final report into sections. These correspond with key aspects of a successful reasoning and analytical process, including information evaluation, final judgement, justification, assumptions, alternative hypotheses considered, and gaps in information, as identified in the Intelligence Community Directive 203, which spells out key features of high-quality analytical products **(See Figure 4).** We designed the TRACE report to benefit both the user conducting the analysis as well as consumers of the user's final analytic product. Additionally, by breaking the report template into different sections that correspond to these key reasoning steps, the analytic products users generate through TRACE should provide readers of the analytic products with a clearer sense of the analyst's reasoning process.

[Figure 4 here]

TRACE's other tools are optional and intended to help users complete the report. TRACE provides an annotation tool for sensemaking and critical evaluation of source materials. Prior scholarship suggests that critical source annotation promotes a more active and analytical form of reading.[54] The annotation tool in TRACE enables the user to highlight information that they

consider important, place descriptive tags and comments on those highlights, and add date-based information directly to a timeline tool. TRACE stores the contents of the user's annotations in an easy-access repository for later retrieval. The annotation tool is meant to help users focus on identifying and evaluating key information and to document it easily so that the analyst does not need to keep track of key information in working memory.

TRACE also provides analysis tools to assist users with key aspects of the analytical process that we modelled after prominent SATs, including structured debate,[55] key assumptions check,[56] and ACH. TRACE provides them as aids that the analyst can choose to use as the problem and the circumstances dictate. For example, the key assumptions check tool asks the user to identify the assumptions underlying what they consider to be their best hypothesis and consider situations where these assumptions could be wrong as well as the effect this would have on their conclusions. Users enter their answers to these questions in text boxes that they can then export into the assumptions section of the TRACE report.

TRACE's optional hypothesis evaluation tool provides a modified version of ACH. This tool asks users to rate the supportiveness of each piece of evidence they've tagged using the annotation tool on a scale from "Highly Contradictory" to "Highly Supportive" for each hypothesis they have identified **(see Figure 5).** Based on this evaluation, the tool then generates a score for each hypothesis to help users determine which of the hypotheses was the strongest given the currently available information. Participants can then select what they considered to be the best hypothesis and add this hypothesis into the final judgement report section.

[Figure 5 here]

TRACE uses context-aware nudging to offer guidance and provide feedback at key moments during the reasoning process. Research suggests that nudging can be effective at

helping people do a task without significantly disrupting flow.[57] TRACE provides nudges in multiple forms, including text suggestions in the tools, pop-up notifications, and visual cues. For example, a pop-up notification encourages use of the annotation tool if they have not done so after viewing a source for three minutes. Similarly, each report section includes links to specific tools that can aid users to complete that section.

Overall, TRACE provides tools and techniques that users can adopt to augment their natural reasoning process. To do so, TRACE offers a report template, guidance, and optional tools and techniques intended to help users process and evaluate information, offload their memory and reasoning work to the application,[58] as well as to mitigate cognitive biases and other errors in reasoning and decision-making.

Methods

We designed an experiment to test the effectiveness of TRACE in improving quality of reasoning when compared to Evidence Alignment Assessment (EAA), and to an external control in which participants were given an analytical problem in PDF format and a blank text box on Qualtrics, an online survey software, to report their findings. Although the ultimate questions around the effectiveness of SATs need to be addressed in real world settings, initial investigations of core underlying properties, such as this one, can begin with an exploration of the impact on more general reasoning.

We developed three complex problems to use in testing. These problems were designed to simulate situations that intelligence analysts encounter, such as determining an evacuation route during a riot ("Cambria Escape Route"), identifying a suspect based on limited information ("Which Lovell?"), or determining the motivation behind a violent attack ("Unusual Suspects"). Each problem had between five and eight pieces of evidence or source materials associated with

them. For example, the problem about determining the motivation behind a violent attack, "Unusual Suspects," included newspaper reports of the event, a Wikipedia entry about actors involved in the attack, a Facebook page of the possible attacker, and the front page of a website for a watchdog organization that contained detailed information about the attack. We designed these problems with input from former open source intelligence analysts. The problems thus were designed to have face validity relative to the types of problems open source analysts would tackle in their actual daily work, including misleading and irrelevant source materials, and hence were intended to tap into similar types of reasoning and judgment in non-expert participants. Moreover, we designed the problems to activate common cognitive biases, such as confirmation bias and fundamental attribution error. The three problems used in this experiment required different types of reasoning, work-flows, and also were of varied difficulty levels, as ascertained in pre-testing of these materials using accuracy of the judgment as the metric. Because of this, we considered them as conditions in the experimental design for the purposes of sample size and power calculation. This experiment was pre-registered with the Open Science Foundation (OSF),[59] following best practices in psychology and the need for more transparency in experimental designs.[60]

We hypothesized that both TRACE and EAA participants would demonstrate higher quality of reasoning in their analytical products than those assigned to an external control. Thus:

H1a: Subjects randomly assigned to TRACE will demonstrate higher quality reasoning skills than will subjects assigned to a control condition.

H1b: Subjects randomly assigned to EAA, a modified version of Analysis of Competing Hypotheses, will demonstrate higher quality reasoning than will subjects assigned to a control condition.

Given that TRACE provides a flexible approach relative to EAA's approach, we hypothesized that participants assigned to the TRACE condition would perform better in terms of quality of reasoning than those in the EAA condition.

H2: Subjects randomly assigned to TRACE will demonstrate higher quality reasoning than will subjects assigned to modified ACH (EAA).

Finally, given that the problems varied in their difficulty, we hypothesized that the quality of reasoning would be associated with the assignment to different problems.

H3: Reasoning quality will vary by problem, with some problems being more difficult than others.

Participants were assigned to conditions via a block randomization design. Within each block, participants were randomly assigned to conditions by sampling without replacement from the set of all conditions. Participants for this study were recruited from Amazon Mechanical Turk (MTurk). We recruited 396 participants,[61] based on a power analysis considering 9 conditions, an effect size f = .25, $\alpha = 0.05$, and power = 0.95, which indicated that we needed a minimum of 378 participants.[62] After solving a problem in their assigned condition and submitting their analytical report, participants were asked to answer a 20-minute survey. The survey asked a series of questions about demographics. The final numbers of participants for each case were: Cambria Escape Route, $n = 127$; Unusual Suspects, $n = 130$; Which Lovell?, $n = 139$.

Measures:

Quality of Reasoning: We used systematic content analysis as suggested by Neuendorf and Krippendorff to determine the quality of reasoning, on several dimensions, in analytic products.[63] The analysis was informed by the Intelligence Community Directives (ICD) 203. Unlike some

prior studies that only score the probability estimates of the judgment, we aimed for a more holistic evaluation. Analytic products are read by decision-makers who are making determinations of the soundness of the product on several elements, including considerations of gaps in the information that would change the judgment and detailing key assumptions. For this reason, our evaluation criteria included several dimensions. All dimensions were then constructed as categorical variables. We operationalize the following dimensions of reasoning:

a) Quality of Justification, defined by a set of dimensions of argumentation and reasoning, such as the use of evidence-based argumentation, offering complete answers and making direct connections between arguments and evidence;
b) Evaluation of Resources, defined as a critical evaluation of usefulness and credibility of available resources;
c) Evaluation of Assumptions, which considers the extent to which reasoners are able to identify underlying assumptions in their conclusion and reasoning, as well as to evaluate their potential impact;
d) Identification of Alternative Hypotheses, evaluated as the extent to which reasoners present and justify alternative plausible answers and whether they offer reasons to discard them in favor of their selected answer;
e) Gaps in Knowledge, defined as reasoners' ability to identify relevant gaps in the available evidence and the extent to which they may affect the selected hypothesis.

Coders were trained to evaluate cases on these dimensions based on a set of directions that were further tailored for each specific problem. These additional directions outlined examples of correct answers, presented examples of high and low quality of reasoning, and provided

problem-specific guidelines or rules to judge analytical reports in the general categories of the codebook. For reliability, we used pairwise agreement scores over 80% before coders were able to analyze reports. All coders achieved this level of agreement. After coding analytical reports individually, each pair of coders discussed any remaining disagreements in their scoring so as to create a final dataset, which was then used for analysis.

Other Measures: To measure system usability, we administered the System Usability Scale (SUS) during the post survey. [64] The SUS includes a set of questions that measures the system functionality of a given digital platform, such as the web-based applications that participants in each condition needed to use to solve the problem. The scale provides options, such as "I find the application very cumbersome to use" and "I felt very confident using the application."

Results

Main results

The experiment was conducted on 396 subjects: EAA (n = 120), TRACE (n = 114), and the external control (n = 162). Owing to a structural error in the external control condition, we discarded 40 cases.[65] The final dataset for analysis had 356 valid participants. **Table 1** shows the number of cases by assigned condition and problem.

[Table 1 here]

The quality of reasoning scores ranged from -8 to +17, and the average quality of reasoning scores, considering the three problems, was 6.4 (SD = 4.8). **Table 2 and Figure 6** show the differences in quality of reasoning between conditions. TRACE's 95% confidence interval lower bound was much higher than the upper bounds for EAA and the external control, demonstrating that TRACE produced comparably higher levels of quality of reasoning.

[Table 2 here]

[Figure 6 here]

The first set of hypotheses compared TRACE and EAA to the external control condition. H1a posited that participants assigned to TRACE, our flexible analytic technique, would demonstrate greater quality of reasoning skills when compared to the external control condition. The results support this hypothesis, as TRACE participants had significantly higher quality of reasoning scores than those assigned to the control condition. Participants assigned to EAA, however, did not demonstrate greater quality of reasoning skills compared to the external control, rejecting H1b. As demonstrated in Figure 6, the confidence intervals for quality of reasoning scores overlap, suggesting that the differences in quality of reasoning scores were not significant (EAA had a lower bound of 4.2, and the external control had an upper bound of 4.7).

The second hypothesis compared the two structured techniques and stated that TRACE would yield greater quality of reasoning than would EAA. This hypothesis was supported. When looking at confidence intervals around the average quality of reasoning scores, the upper bound of EAA was 5.8, while the lower bound of TRACE was 9.4.

The third hypothesis predicted a significant variance in quality of reasoning by analytical problem. The differences in quality of reasoning between problems were not statistically meaningful: The average score for Cambria Escape Route was 6.5 (SD = 4.6), the average for Which Lovell? was 6.5 (SD = 4.7), and the average for Unusual Suspects was 6.1 (SD = 5.1). An ANOVA revealed that the differences by problem were not statistically different from one another (F (2, 353) = 0.38, p = 0.69), rejecting our third hypothesis.

We further ran an Ordinary Least Squares (OLS) regression that examined the extent to which the structured technique predicted quality of reasoning controlling for problem to test the

robustness of our findings, as shown in **Table 3.** TRACE yielded higher quality of reasoning than did the control condition (beta = 0.61, $p < 0.001$). EAA also yielded higher quality of reasoning than did the control condition (beta = 0.10, $p < 0.05$). These results provide some support for H1b, when the problem is taken into consideration. There is some evidence to suggest that TRACE provided the greatest benefit to quality of reasoning as demonstrated by the confidence intervals around the TRACE and EAA coefficients, thus providing further support for our second hypothesis.

[Table 3 here]

Post-hoc analysis

We further tested the robustness of our findings by controlling for demographic characteristics and word count. We ran an OLS regression taking demographics into account. We also included the word counts of the reports submitted. It is possible that the structured techniques, TRACE and EAA, might prompt people to write more than the control condition. We wanted to ensure that positive benefits of these conditions were more robust than mere word count boosters.

The number of words generated in the reports varied by assignment to condition. The word count averages were 728.52 (SD = 370.24) for TRACE, 461.33 (SD = 279.42) for EAA, and 435.72 (SD = 289.66) for the control condition. ANOVA results indicated that there were significant differences between conditions (F (2, 353) = 30.92, $p < 0.001$, $\eta^2 = 0.149$). This suggested that it was important to control for word count to determine whether TRACE made important contributions to quality of reasoning once word count was controlled.

As shown in **Table 4**, even if demographic characteristics and report word count are taken into consideration, both TRACE and EAA outperformed the control condition on quality

of reasoning scores (beta = 0.43, $p < 0.001$ and beta = 0.09, $p < 0.05$ respectively). The increases in quality of reasoning were not merely the result of encouraging participants to write more than they did in the control condition. The number of words in the reports was positively associated with quality of reasoning ($r = 0.619$, $p < 0.001$), and in the regression, the word count produced a beta = 0.47, $p < 0.001$. Importantly, TRACE and EAA made important, statistically significant contributions to reasoning even when word count was controlled.

[Table 4 here]

To better understand what features participants were most and least likely to use in TRACE and what that might suggest about the efficacy of TRACE, we analyzed log data. All actions in the software were logged for each individual. This allowed us to see when and what tools people used. We found that participants who used the hypothesis generation tool to generate at least two hypotheses had statistically significantly higher reasoning quality than those who did not use the tool ($t(40) = -3.08$, $p = .004$). Using the tagging tool to highlight and annotate source materials also had a strong, positive association with judgment accuracy, $t(112) = -1.99$, $p = 0.04$, and quality of reasoning, $t(112) = -3.32$, $p = .001$, when we compared participants who used the tagging tool and made at least ten annotations or highlights (N=40), to those who made nine or fewer uses of the tool (N = 72).

We analysed the sequences of actions in TRACE to further understand how high quality reasoning was produced. For this sequence analysis, we binned the quality of reasoning scores into three categories of low, medium, and high and analysed when in their progress they used tools and features of TRACE. As Figure 7 shows, users who wrote the highest quality analytical products were: more likely to spend time with the “understand the task” tool; to use the timeline tool early in their analysis; to evaluate and rate the relevance and credibility of the source

materials early in the analysis; to generate hypotheses earlier in their analysis; and to review the background document, which detailed the problem that they were to reason through. By contrast, those who wrote analytic products that were scored low were likely to use the timeline tool late in their analysis process, to rate sources much later than those who scored higher, and to use the key assumptions check and/or hypothesis generation tool generally later in their process.

[Figure 7 Here]

Discussion

This paper investigated the effectiveness of TRACE, a flexible analytic technique that offers analysts a variety of optional tools to assist with memory and recall as well as guidance for each step of the analytical process, as compared to a modified version of ACH—which we refer to as *Evidence Alignment Assessment* (EAA)—and to an external control of unaided reasoning. Our analysis finds that TRACE was the most effective approach to assisting quality of reasoning in analytical problem-solving when compared to both the external control condition and to EAA. These results suggest that novice analysts can benefit from a flexible analytic technique with a light structure and several different tools to aid problem-solving as the analyst dictates, and that these benefits are not only noticeable when compared to an external control condition, in which there was minimal guidance, but also in relation to an ordered structured technique of the same basic nature as ACH. The benefits of TRACE were consistent in both the bivariate analysis and the regression analysis controlling for analytical cases.

Another goal of this study was to assess the effectiveness of the foundational elements of ACH, Although there were no meaningful reasoning benefits in the bivariate analysis for EAA—that is, without taking the analytical problems into consideration—when it was compared

to the control condition, we found some modest benefits to improving quality of reasoning in participants using EAA once the case was included as a control. These results are consistent with some of the criticism towards traditional structured techniques and their limited applicability to a broad array of intelligence problems. Following a set of predefined steps to generate and evaluate hypotheses, the core of ACH can be helpful to solve specific types of problems. Consistent with arguments other scholars have put forward, we suggest that future research should focus on understanding precisely which types of intelligence problems are suitable for which types of SAT.[66] The current results cannot speak to whether any of the omitted elements of ACH itself would have provided substantial improvements in reasoning, but given its complexity any future research seeking to examine that is likely to need to employ highly trained participants.

It is worth noting that there were no major differences in quality of reasoning by problem. This is contrary to what we had hypothesized. This finding suggests that approaches like TRACE are not limited to a single type of reasoning problem. The more flexible analytical technique can be applied to support reasoning problems with varying levels of difficulty and of different types of reasoning.

Taken together, the results of our experiment suggest that TRACE provided a markedly more effective approach for aiding reasoning relative to a more prototypical, ordered SAT. Although we did not directly measure the probable use of cognitive biases in the analysis, we designed the problems to activate them, especially confirmation bias and fundamental attribution error. Incorrect judgments and faulty reasoning were measured in the evaluation of analytical products, and weak reasoning was likely caused by reliance on cognitive biases. Future

experiments need to tease out more directly which cognitive biases might be less likely to be relied upon when using TRACE as compared with other SATs.

We do not claim these findings directly reflect either the SATs used by analysts, nor how they are used in practice, and they may not capture the performance of trained analysts on real analytic problems. Additional experiments are necessary to determine the generalizability of these results. Keeping in mind that the population for this study was not restricted to professional intelligence analysts, it may be the guidance TRACE offered was particularly beneficial for those that do not have any experience conducting analyses of this nature. It could be that trained analysts gain greater benefit from a more structured analytical technique than TRACE provided. Consequently, it is important for future research to examine TRACE's impact on the quality of reasoning among trained intelligence analysts and to do further comparisons with more structured approaches. Additionally, a longitudinal study where novices complete several analyses would help determine if TRACE's efficacy endures after participants have become more familiar with this type of analytical task.

Given these findings, we suggest that the IC needs to more carefully consider training and techniques in tradecraft to support high quality analysis. ACH and other SATs are part of the core training that analysts are given. Our results add to the small but growing body of scholarly studies that have empirically tested aspects of SATs and find that they may not have the desired effects initially predicted by Heuer and Pherson. Our results suggested that EAA, a prototypical structured technique, is less effective than TRACE, which provides a set of tools for analysis but enables more flexibility as to which tools and structures to use to support their specific analytical needs. The training and recommendations to use ACH in the *Tradecraft Primer* may need reconsideration as the evidence mounts that it fails to meet the needs it was designed to address.

Our results suggest that complex reasoning requires a variety of tools to aid in the analysis process. Moreover, this study finds that greater support for writing the analytical product is needed to help detail more fully the sources that influenced the final judgment, as well as assumptions that may have driven the analysis, and gaps in the available information that might be affecting the final judgment.

Conclusion

Heuer and Pherson substantially influenced the nature of intelligence tradecraft by pushing for greater awareness of cognitive biases and developing SATs to address them.[67] Their work over the past two dozen years has helped the IC to be rigorous and reflective about the flaws in human cognition that can lead to faulty reasoning. These SATs, though, need rigorous social scientific experimentation to ascertain whether they indeed provide the benefits to analysis. In this paper, we conducted such an experiment, examining a prototypical SAT (EAA), and compared that to a flexible SAT (TRACE). Our findings suggest that the more flexible work environment, clear template for generating a high-quality analytical product, and a suite of optional analytical tools generated higher quality reasoning than the more typical SAT approach, and as compared with unaided reasoning. There is a hint that EAA can improve reasoning when the forms of problems are taken into account, offering some limited evidence that the elements underlying common practices like ACH can be beneficial. However, the combined results suggest future support for analytical reasoning in intelligence contexts may provide more benefits if approaches are aimed at the processes of reasoning in a flexible, report-focused format as found in TRACE.

Notes

1. Gilovich, Griffin, and Kahneman, *Heuristics and Biases.*

2. Gilhooly, "Working Memory and Reasoning."

3. Heuer, *Psychology of Intelligence Analysis*.

4. Heuer, and Pherson, *Structured Analytic Techniques for Intelligence Analysis*.

5. See note 3 above.

6. Artner, Girven, and Bruce, "Assessing the Value of Structured Analytic Techniques"; Jones, "Critical Epistemology for Analysis of Competing Hypotheses"; and Wastell, "Cognitive Predispositions and Intelligence Analyst Reasoning."

7. Jones, "Critical Epistemology for Analysis of Competing Hypotheses."

8. Chang, Berdini, Mandel, and Tetlock, "Restructuring Structured Analytic Techniques"; Coulthart, "An Evidence-Based Evaluation of 12 Core Structured Analytic Techniques"; Whitesmith, "The efficacy of ACH in mitigating serial position effects and confirmation bias in an intelligence analysis scenario"; and Mandel, Karvetski, and Dhami, "Boosting intelligence analysts' judgment accuracy: What works, what fails?"

9. Wastell, "Cognitive Predispositions and Intelligence Analyst Reasoning"; and Artner, Girven, and Bruce, "Assessing the Value of Structured Analytic Techniques."

10. Heuer, *Psychology of Intelligence Analysis*; and Heuer, and Pherson, *Structured Analytic Techniques for Intelligence Analysis*.

11. We relied on convenience sampling to recruit and interview members of the intelligence community. We interviewed 9 current and former analysts with a range of experiences. The majority of our interviewees have a background in the Air Force, but we also interviewed

analysts with experiences in a number of other Department of Defense and national level intelligence agencies. The goal for these interviews was not to provide generalizable findings based on a representative sample of IC analysts, but to gather initial design ideas and feedback for us to keep in mind while developing early versions of TRACE. Each interview followed a semi-structured protocol asking the interviewee to walk us through how they would answer a fictional question. The fictional question was meant to mimic a question that intelligence analysts encounter – determining the motivation underpinning a foreign leader's recent actions, and we included source materials to help the interviewees answer the question. During the interview, we asked follow-up questions probing for details, including the interviewees' analytical process, structured analytic techniques, collaboration, the composition and reception of analytic products, the key challenges that analysts in the IC face, and IC culture.

12. Moon, and Hoffman, "How Might 'Transformational' Technologies and Concepts Be Barriers."

13. Robb, Silberman, Levin, McCain, Rowen, Slocombe, Studeman, et al., *The Commission on the Intelligence Capabilities.*

14. Ibid., *408.*

15. Duelfer, *Report on the Commission on the Intelligence Capabilities*

16. Carrington, Chen, Davies, Kaur, and Neville, "The Effectiveness of a Single Intervention"; Hitchcock, "The Effectiveness of Computer-Assisted Instruction"; Kunsch, Schnarr, and van Tyle, "The Use of Argument Mapping"; and Twardy, "Argument Maps Improve Critical Thinking."

17. Rowe, Macagno, Reed, and Walton, "Araucaria as a Tool"; and Gelder, "Cultivating Deliberation for Democracy."

18. Billman, Convertino, Shrager, Pirolli, and J Massar, “Collaborative Intelligence Analysis”; Gustavi, Karasalo, and Mårtenson, “A Tool for Generating”; and Wilson, Brown, and Biddle, “ACH Walkthrough: A Distributed Multi-Device Tool.”

19. Good, Shrager, Stefik, Pirolli, Card, and Heuer, “ACH1.1: A Tool for Analyzing.”

20. http://competinghypotheses.org/.

21. Artner, Girven, and Bruce, “Assessing the Value of Structured Analytic Techniques.” Artner et al. reviewed twenty-nine Central Intelligence Agency analytic products posted on the World Intelligence Review during a two-week period in July 2014, fourteen National Intelligence Council analytic products posted on Intelink in July 2014, and a random sample of twenty Defense Intelligence Agency and Central Intelligence Agency products published in 2013 on four specific issues. The authors do not specify the four issues they chose to examine in their random sample.

22. Coulthart, “Why Do Analysts Use Structured Analytic Techniques? An In-depth Study of an American Intelligence Agency.”

23. Moon, and Hoffman, “How Might ‘Transformational’ Technologies and Concepts Be Barriers”; and Wastell, “Cognitive Predispositions and Intelligence Analyst Reasoning.”

24. Wastell, ibid.

25. Coulthart, “Why Do Analysts Use Structured Analytic Techniques? An In-depth Study of an American Intelligence Agency.”

26. Immerman, “Transforming Analysis: The Intelligence Community’s Best Kept Secret”; Fingar, “Analysis in the US Intelligence Community: Missions, Masters, and Methods.”

27. Coulthart, “Why Do Analysts Use Structured Analytic Techniques? An In-depth Study of an American Intelligence Agency.”

28. Johnston, “Analytic Culture in the US Intelligence Community: An Ethnographic Study.”

29. Coulthart, “Why Do Analysts Use Structured Analytic Techniques? An In-depth Study of an American Intelligence Agency.”

30. Marrin, “Is Intelligence Analysis an Art or Science?”

31. Dhami, Mandel, Mellers, and Tetlock, “Improving Intelligence Analysis with Decision Science”; and Mandel, and Tetlock, “Correcting Judgment Correctives in National Security Intelligence.”

32. Artner, Girven, and Bruce, “Assessing the Value of Structured Analytic Techniques”; and Coulthart, “An Evidence-Based Evaluation of 12 Core Structured Analytic Techniques.”

33. Chang, Berdini, Mandel, and Tetlock, “Restructuring Structured Analytic Techniques.”

34. See note 13.

35. See note 3.

36. Ibid.

37. Ibid., *108.*

38. Jones, “Critical Epistemology for Analysis of Competing Hypotheses”; and Chang, Berdini, Mandel, and Tetlock, “Restructuring Structured Analytic Techniques.”

39. See note 7.

40. See note 31.

41. See note 36.

42. See note 8.

43. Lehner, Adelman, Cheikes, and Brown, “Confirmation Bias in Complex Analyses.”

44. Nickerson, “Confirmation Bias: A Ubiquitous Phenomenon.”

45. Kretz, Simpson, and Graham, “A Game-Based Experimental Protocol.”

46. Billman, Convertino, Shrager, Pirolli, and J Massar, “Collaborative Intelligence Analysis.”

47. Ibid.

48. Ibid.

49. Whitesmith, “The Efficacy of ACH in Mitigating Serial Position Effects and Confirmation Bias in an Intelligence Analysis Scenario.”

50. Mandel, Karvetski, and Dhami, “Boosting Intelligence Analysts’ Judgment Accuracy: What Works, What Fails?”

51. Dhami, Belton, and Mandel, “The ‘Analysis of Competing Hypotheses’ in Intelligence Analysis.”

52. Ibid.

53. Heuer (Dec. 8, 2009) noted in a lecture he gave to the National Academies of Science that measuring judgment accuracy as the criterion to determine the efficacy of ACH is problematic. Accuracy is conditional based on current knowledge and circumstances necessarily will change accuracy. Moreover, accuracy is necessarily probabilistic, and so determining what is accurate in a probability estimate is challenging. He argues that assessing ACH and other SATS based on quality of analysis is a better metric to determine efficacy.

54. Bradley, and Vetch, “Supporting Annotation as a Scholarly Tool”; Glover, Xu, and Hardaker, “Online Annotation – Research and Practices”; Gomez, Gomez, Cooper, Lozano, and Mancevice, “Redressing Science Learning through Supporting Language”; Herman, Perkins, Hansen, Gomez, and Gomez, “The Effectiveness of Reading Comprehension Strategies”; Simpson, and Nist, “Textbook Annotation: An Effective and Efficient Study Strategy”; and Zywica, and Gomez, “Annotating to Support Learning.”

55. Boyd, Baliko, and Polyakova-Norwood, “Using Debates to Teach”; Freeley, and Steinberg, Argumentation and Debate; and Mercier, and Landemore, “Reasoning Is for Arguing.”

56. See note 4.

57. Thaler, and Sunstein. *Nudge: Improving Decisions about Health.*

58. Hutchins, *Distributed cognition*

59. The Open Science Pre-Registration can be found here: https://osf.io/2u9nh. All experimental materials, including the case problems and the evaluation approach (codebook and case keys) are available upon request to the first author.

60. Gonzales, and Cunningham, “The Promise of Pre-Registration in Psychological Research.”

61. We used MTurk's customized qualifications to block participants that participated in previous studies and user testing of TRACE. Participants must have resided in the U.S. and have a minimum of a bachelor’s degree. Based on our analyses there were no other systematic demographic differences or other factors that set them apart from the general population.

62. We used the R package pwr to calculate the statistical power[52] we would need to ascertain discernable differences between conditions and cross-checked the results using the software G*power. See: Champely, Ekstrom, Dalgaard, Gill, Weibelzahl, Anandkumar, Ford, et al., “Package ‘Pwr.’”

63. Neuendorf, *The Content Analysis Guidebook*; and Krippendorff, *Content Analysis: An Introduction.*

64. Brooke, “SUS- A ‘quick and Dirty’ Usability Scale.”

65. After data collection, we identified that a few respondents were able to add comments in the source material, a PDF hosted in Google Docs, which changed the nature of the external control.

This problem affected 36 cases in the Cambria Escape Route external control condition and 4 cases in the Unusual Subjects condition.

66. Artner, Girven, and Bruce, "Assessing the Value of Structured Analytic Techniques"; Chang, Berdini, Mandel, and Tetlock, "Restructuring Structured Analytic Techniques"; and Dhami, Mandel, Mellers, and Tetlock, "Improving Intelligence Analysis with Decision Science."

67. We thank a reviewer for pointing out that Richards Heuer and Randolph Pherson's work on SATs was greatly influenced by prior work by Jack Davis and Roger George on "alternative analysis" techniques. For summaries, see: George, "Fixing the Problem of Analytical Mind-Sets: Alternative Analysis"; and Davis, "Improving CIA Analytic Performance: Strategic Warning."

Bibliography

Artner, Stephen, Richard S. Girven, and James Bruce. "Assessing the Value of Structured Analytic Techniques in the U. S. intelligence community." *Rand Corporation*, 2016.

Billman, Dorrit, Gregorio Convertino, Jeff Shrager, P Pirolli, and J Massar. "Collaborative Intelligence Analysis with CACHE and Its Effects on Information Gathering and Cognitive Bias." In *Human Computer Interaction Consortium Workshop*, 2006.

Boyd, Mary R., Beverly Baliko, and Vera Polyakova-Norwood. "Using Debates to Teach Evidence-Based Practice in Large Online Courses." *Journal of Nursing Education* 54, no. 10 (2015): 578–82. https://doi.org/10.3928/01484834-20150916-06.

Bradley, John, and Paul Vetch. "Supporting Annotation as a Scholarly Tool—Experiences from the Online Chopin Variorum Edition." *Literary and Linguistic Computing* 22, no. 2 (June 1, 2007): 225–41. https://doi.org/10.1093/llc/fqm001.

Brooke, John. “SUS- A ‘quick and Dirty’ Usability Scale.” In *Usability Evaluation in Industry*, edited by Patrick W Jordan, Bruce Thomas, Bernard A Weerdmeester, and Ian L McLelland, 189–94. London: Taylor & Francis, 1996.

Chang, Welton, Elissabeth Berdini, David R. Mandel, and Philip E. Tetlock. “Restructuring Structured Analytic Techniques in Intelligence.” *Intelligence and National Security* 33, no. 3 (April 16, 2018): 337–56. https://doi.org/10.1080/02684527.2017.1400230.

Cacioppo, J. T., R. E. Petty, and C. F. Kao. “Need for Cognition Scale.” Measurement Instrument Database for the Social Sciences, 2013. Retrieved from http://www.midss.org/content/need-cognition-scale. https://doi.org/10.13072/midss.650.

Carrington, Michal, Richard Chen, Martin Davies, Jagjit Kaur, and Benjamin Neville. “The Effectiveness of a Single Intervention of Computer-aided Argument Mapping in a Marketing and a Financial Accounting Subject.” *Higher Education Research & Development* 30, no. 3 (2011): 387–403. https://doi.org/10.1080/07294360.2011.559197.

Champely, Stephane, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, Helios De Rosario, and Maintainer Helios De Rosario. “Package ‘Pwr.’” *R Package Version*, 2018, 1–2.

Coulthart, Stephen J. “An Evidence-Based Evaluation of 12 Core Structured Analytic Techniques.” *International Journal of Intelligence and CounterIntelligence* 30, no. 2 (April 3, 2017): 368–91. https://doi.org/10.1080/08850607.2016.1230706.

Coulthart, Stephen. “Why Do Analysts Use Structured Analytic Techniques? An in-Depth Study of an American Intelligence Agency.” Intelligence and National Security 31, no. 7 (November 9, 2016): 933–48. https://doi.org/10.1080/02684527.2016.1140327.

Davis, Jack. “Improving CIA Analytic Performance: Strategic Warning.” *The Sherman Kent Center for Intelligence Analysis Occasional Papers* 1, no.1 (September 2002). https://www.cia.gov/library/kent-center-occasional-papers/vol1no1.htm

Dhami, Mandeep K., David R. Mandel, Barbara A. Mellers, and Philip E. Tetlock. “Improving Intelligence Analysis with Decision Science.” Perspectives on Psychological Science 10, no. 6 (November 1, 2015): 753–57. https://doi.org/10.1177/1745691615598511.

Dhami, Mandeep K., Ian K. Belton, and David R. Mandel. "The ‘Analysis of competing hypotheses’ in intelligence analysis." *Applied Cognitive Psychology* (2019). https://doi.org/10.1002/acp.3550.

Duckworth, Angela Lee, and Patrick D. Quinn. “Development and Validation of the Short Grit Scale (Grit–S).” *Journal of Personality Assessment* 91, no. 2 (February 17, 2009): 166–74. https://doi.org/10.1080/00223890802634290.

Duelfer, Charles. *Report on the Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction*. Central Intelligence Agency, 2004. https://www.cia.gov/library/reports/general-reports-1/iraq_wmd_2004/.

Fingar, Thomas. “Analysis in the US Intelligence Community: Missions, Masters, and Methods.” In Intelligence Analysis: Behavioral and Social Scientific Foundations, edited by B Fischhoff and C Chauvin, 3–27. Washington, D.C.: National Academies Press, 2011.

Freeley, Austin J, and David L Steinberg. *Argumentation and Debate: Critical Thinking for Reasoned Decision Making, 12th Edition*. Boston, MA: Wadsworth Cengage Learning, 2009.

Gelder, Tim van. “Cultivating Deliberation for Democracy.” *Journal of Public Deliberation* 8, no. 1 (2012): Article 12.

George, Roger Z. “Fixing the Problem of Analytic Mind-Sets: Alternative Analysis.” *International Journal of Intelligence and CounterIntelligence* 17, no. 3 (2004): 385-404. https://doi.org/10.1080/08850600490446727

Gigerenzer, Gerd. *Gut Feelings: The Intelligence of the Unconscious*. New York: Viking Press, 2007.

Gilhooly, Kenneth J. “Working Memory and Reasoning.” In *The Nature of Reasoning*, edited by Robert J. Sternberg and Jacqueline P. Leighton, 49–67. Cambridge, UK: Cambridge University Press, 2004.

Gilovich, Thomas, Dale W. Griffin, and Daniel Kahneman, eds. *Heuristics and Biases: The Psychology of Intuitive Judgement*. Cambridge, UK: Cambridge University Press, 2002.

Glover, Ian, Zhijie Xu, and Glenn Hardaker. “Online Annotation – Research and Practices.” *Computers & Education* 49, no. 4 (2007): 1308–20. https://doi.org/10.1016/j.compedu.2006.02.006.

Goldstein, Susan B. “Construction and Validation of a Conflict Communication Scale.” *Journal of Applied Social Psychology* 29, no. 9 (1999): 1803–1832. https://doi.org/10.1111/j.1559-1816.1999.tb00153.x.

Gomez, Kimberley, Louis M. Gomez, Benjamin Cooper, Maritza Lozano, and Nicole Mancevice. “Redressing Science Learning through Supporting Language: The Biology Credit Recovery Course.” *Urban Education*, November 15, 2016, 0042085916677345. https://doi.org/10.1177/0042085916677345.

Gonzales, Joseph, and Corbin A. Cunningham. “The Promise of Pre-Registration in Psychological Research: Encouraging A Priori Research and Decreasing Publication Bias.” American Psychological Association, August 2015. https://www.apa.org/science/about/psa/2015/08/pre-registration

Good, Lance, Jeff Shrager, Mark Stefik, Peter Pirolli, Stuart Card, and Richards J Heuer. “ACH1.1: A Tool for Analyzing Competing Hypotheses.” *Palo Alto Research Center*, 2005. http://www.pherson.org/PDFFiles/ACHTechnicalDescription.pdf.

Gustavi, Tove, Maja Karasalo, and Christian Mårtenson. “A Tool for Generating, Structuring, and Analyzing Multiple Hypotheses in Intelligence Work.” *2013 European Intelligence and Security Informatics Conference*, 2013, 23–30.

Herman, Phillip, Kristen Perkins, Martha Hansen, Louis M. Gomez, and Kimberley Gomez. “The Effectiveness of Reading Comprehension Strategies in High School Science Classrooms.” In *Proceedings of the 9th International Conference of the Learning Sciences - Volume 1*, 857–864. ICLS ’10. Chicago, Illinois: International Society of the Learning Sciences, 2010. http://dl.acm.org/citation.cfm?id=1854360.1854470.

Heuer, Richards J. *Psychology of Intelligence Analysis*. Washington, D.C.: Center for the Study of Intelligence, Central Intelligence Agency, 1999.

Heuer, Richards J. “The Evolution of Structured Analytic Techniques.” Presentation to the National Academy of Sciences, Washington, DC, Dec. 8, 1999.

Heuer, Richards J, and Randolph H. Pherson. *Structured Analytic Techniques for Intelligence Analysis*. Washington, D.C.: CQ Press, 2011.

Hitchcock, David. “The Effectiveness of Computer-Assisted Instruction in Critical Thinking.” *Informal Logic* 24, no. 3 (2004): 183–217.

Hutchins, E. (2001). Distributed cognition. In: N. J. Smelser; P. B. Baltes (Eds), International Encyclopedia of the Social and Behavioral Sciences. New York: Elsevier Science.

Johnston, Rob. Analytic Culture in the US Intelligence Community: An Ethnographic Study. Washington, D.C.: Center for the Study of Intelligence, 2005.

Jones, Nicholaos. "Critical Epistemology for Analysis of Competing Hypotheses." *Intelligence and National Security* 33, no. 2 (February 23, 2018): 273–89. https://doi.org/10.1080/02684527.2017.1395948.

Kretz, Donald R, BJ Simpson, and Colonel Jacob Graham. "A Game-Based Experimental Protocol for Identifying and Overcoming Judgement Biases in Forensic Decision Analysis." In *2012 IEEE Conference on Technologies for Homeland Security (HST)*, 439–444. IEEE, 2012. https://doi.org/10.1109/THS.2012.6459889.

Krippendorff, Klaus H. *Content Analysis: An Introduction to Its Methodology. 2nd Edition*. Thousand Oaks, CA: SAGE, 2003.

Kunsch, David W., Karin Schnarr, and Russell van Tyle. "The Use of Argument Mapping to Enhance Critical Thinking Skills in Business Education." *Journal of Education for Business* 89, no. 8 (2014): 403–10. https://doi.org/10.1080/08832323.2014.925416.

Lehner, Paul E., Leonard Adelman, B. A. Cheikes, and M. J. Brown. "Confirmation Bias in Complex Analyses." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 38 (2008): 584–92. 10.1109/TSMCA.2008.918634

Immerman, Richard H. "Transforming Analysis: The Intelligence Community's Best Kept Secret." Intelligence and National Security 26, no. 2–3 (April 1, 2011): 159–81. https://doi.org/10.1080/02684527.2011.559138.

Mandel, David R, Christopher W Karvetski, and Mandeep K Dhami. “Boosting Intelligence Analysts’ Judgment Accuracy: What Works, What Fails?” Judgment and Decision Making 13, no. 6 (2018): 607–621.

Mandel, David R., and Philip E. Tetlock. “Correcting Judgment Correctives in National Security Intelligence.” Frontiers in Psychology 9 (December 2018): 2640. https://doi.org/doi: 10.3389/fpsyg.2018.02640.

Marrin, Stephen. “Is Intelligence Analysis an Art or a Science?” International Journal of Intelligence and CounterIntelligence 25, no. 3 (2012): 529–45. https://doi.org/10.1080/08850607.2012.678690.

Mercier, Hugo, and Hélène Landemore. “Reasoning Is for Arguing: Understanding the Successes and Failures of Deliberation.” *Political Psychology* 33, no. 2 (April 1, 2012): 243–58. https://doi.org/10.1111/j.1467-9221.2012.00873.x.

Moon, Brian M., and Robert R. Hoffman. “How Might ‘Transformational’ Technologies and Concepts Be Barriers to Sensemaking in Intelligence Analysis.” In *Presentation at the Seventh International Conference on Naturalistic Decision Making*, edited by J.M.C. Schraagen, 2005.

Neuendorf, Kimberly A. *The Content Analysis Guidebook. 1st Edition*. Thousand Oaks, CA: SAGE, 2002.

Nickerson, Raymond S. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.” *Review of General Psychology* 2, no. 2 (1998): 175–220. https://doi.org/10.1037/1089-2680.2.2.175.

Roets, Arne, and Alain Van Hiel. "Item Selection and Validation of a Brief, 15-Item Version of the Need for Closure Scale." *Personality and Individual Differences* 50, no. 1 (2011): 90–94. https://doi.org/10.1016/j.paid.2010.09.004.

Rowe, Glenn, Fabrizio Macagno, Chris Reed, and Douglas Walton. "Araucaria as a Tool for Diagramming Arguments in Teaching and Studying Philosophy." *Teaching Philosophy* 29, no. 2 (2006): 111-124.

Robb, Charles S., Laurence H. Silberman, Richard C. Levin, John McCain, Henry S. Rowen, Walter B. Slocombe, William O. Studeman, Charles M. Vest, Patricia Wald, and Lloyd Cutler. *The Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction: Report to the President of the United States*. Government Printing Office, 2005.

Simpson, Michele L, and Sherrie L Nist. "Textbook Annotation: An Effective and Efficient Study Strategy for College Students." *Journal of Reading* 34, no. 2 (1990): 122–129.

Stephen, Artner, Richard S Girven, and James B Bruce. "Assessing the Value of Structured Analytic Techniques in the US intelligence community." RAND National Defense Research Institute Santa Monica United States, 2016.

Thaler, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth and Happiness*. New York: Penguin, 2008.

"The Open Source Analysis of Competing Hypotheses Project," 2010. http://competinghypotheses.org/.

Tversky, Amos, and Daniel Kahneman. "Judgement under Uncertainty: Heuristics and Biases." *Science* 185, no. 4157 (September 27, 1974): 1124. https://doi.org/10.1126/science.185.4157.1124.

Twardy, Charles. “Argument Maps Improve Critical Thinking.” *Teaching Philosophy* 27, no. 2 (2004): 95–116.

US Government. *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*. Central Intelligence Agency, Center for the Study of Intelligence, 2009.

Wastell, Colin A. “Cognitive Predispositions and Intelligence Analyst Reasoning.” *International Journal of Intelligence and CounterIntelligence* 23, no. 3 (June 8, 2010): 449–60. https://doi.org/10.1080/08850601003772802.

Whitesmith, Martha. “The Efficacy of ACH in Mitigating Serial Position Effects and Confirmation Bias in an Intelligence Analysis Scenario.” Intelligence and National Security 34, no. 2 (February 23, 2019): 225–42. https://doi.org/10.1080/02684527.2018.1534640.

Wilson, Timothy D, Christopher E Houston, Kathryn M Etling, and Nancy Brekke. “A New Look at Anchoring Effects: Basic Anchoring and Its Antecedents.” *Journal of Experimental Psychology: General* 125, no. 4 (1996): 387.

Wilson, Jeff, Judith M. Brown, and Robert Biddle. “ACH Walkthrough: A Distributed Multi-Device Tool for Collaborative Security Analysis.” In *Proceedings of the 2014 ACM Workshop on Security Information Workers*, 9–16. ACM, 2014.

Zywica, Jolene, and Kimberley Gomez. “Annotating to Support Learning in the Content Areas: Teaching and Learning Science.” *Journal of Adolescent & Adult Literacy* 52, no. 2 (2008): 155–65. https://doi.org/10.1598/JAAL52.2.6.

Figure 1: Heuer's Original Version of ACH

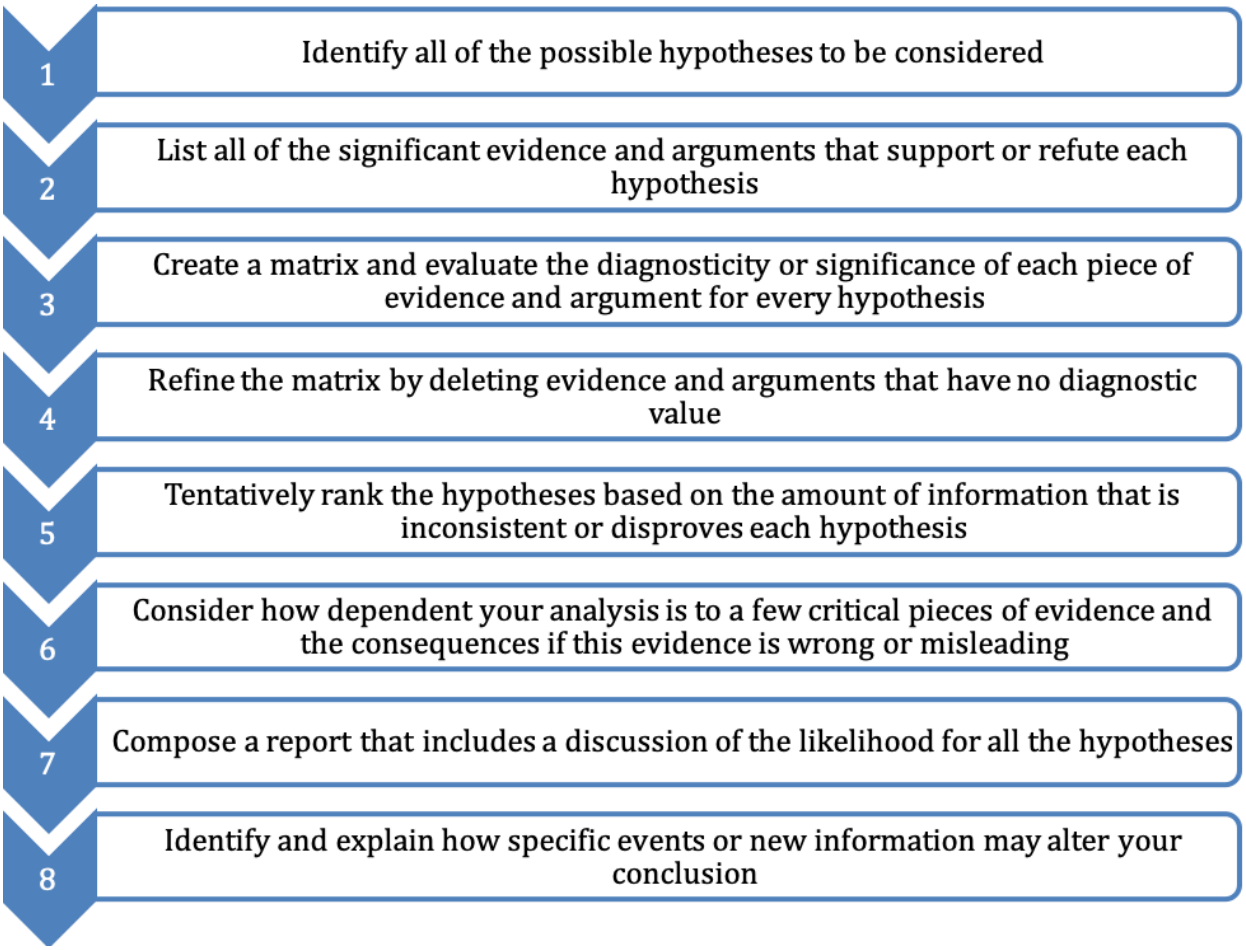

Figure 2: The Hypothesis Generation Step in EAA

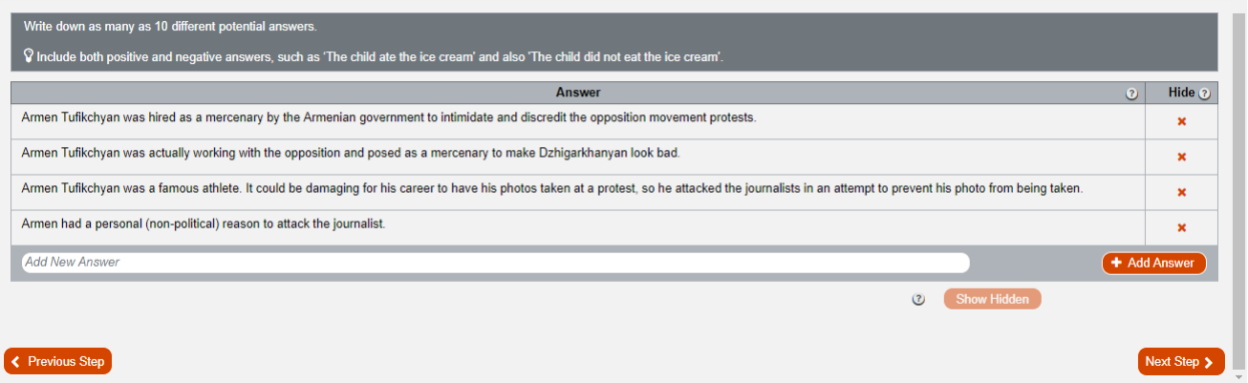

Figure 3: The Assess the Supportiveness of the Resources Step in EAA

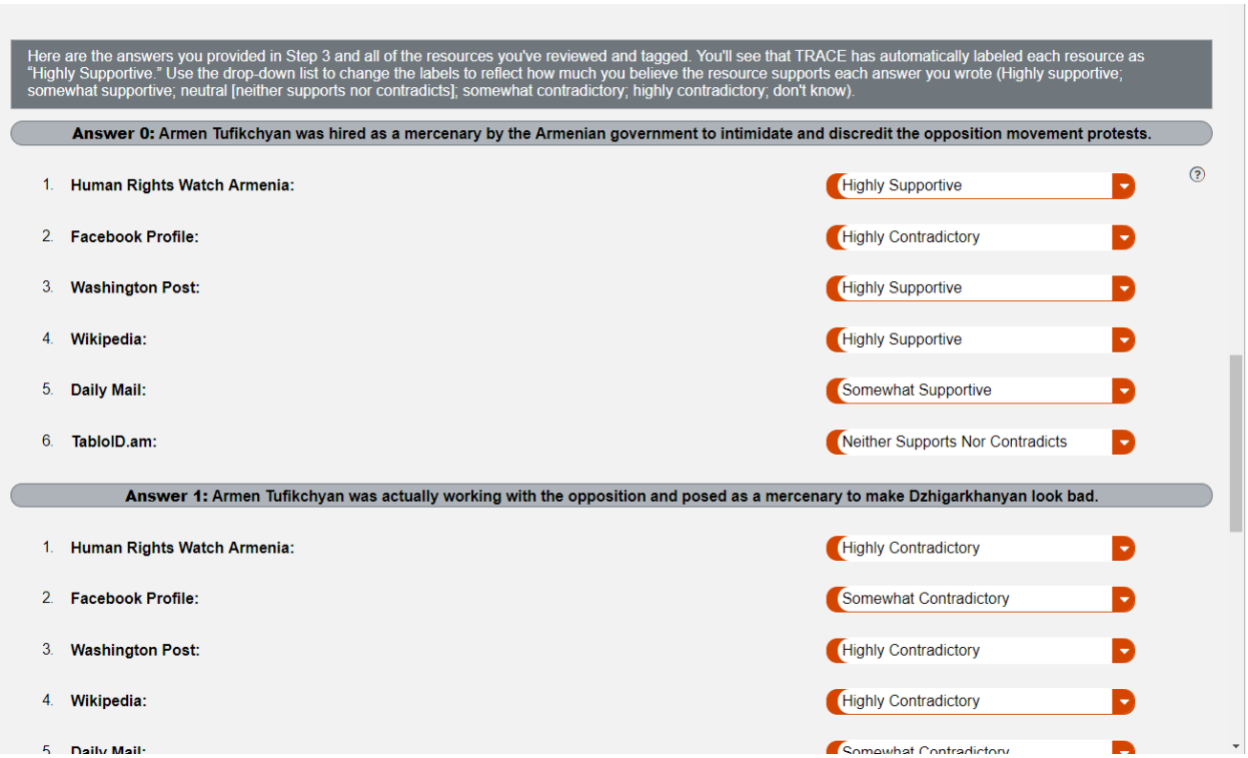

Figure 4: The TRACE Main Page

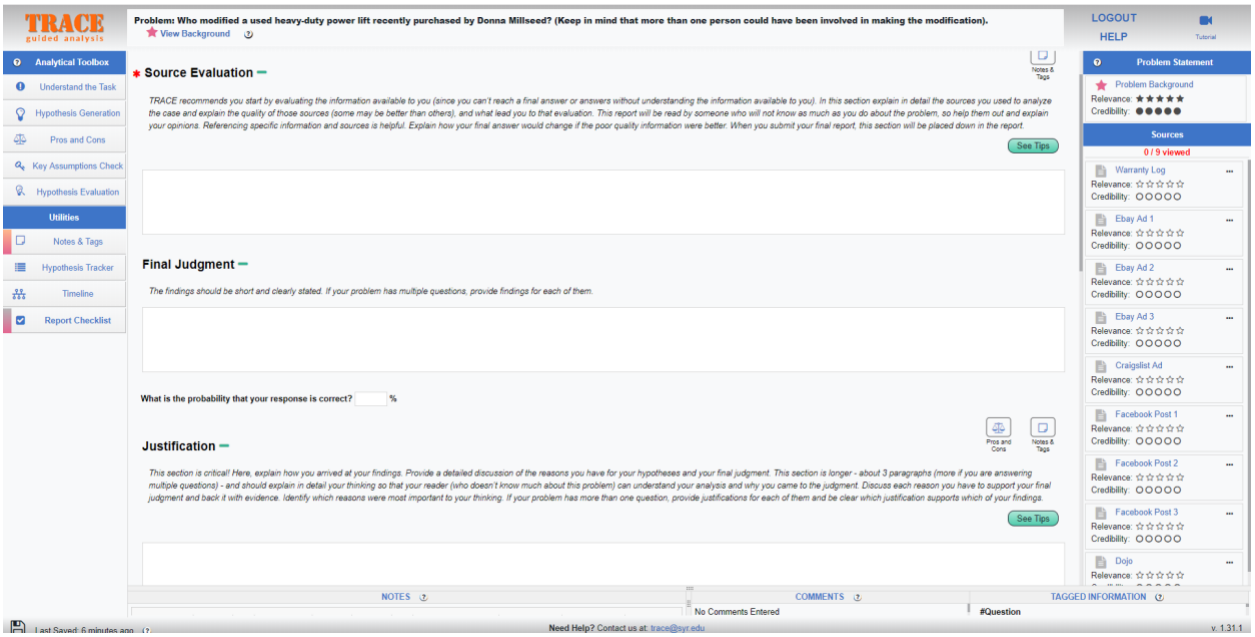

Figure 5: Hypothesis Evaluation Tool in TRACE

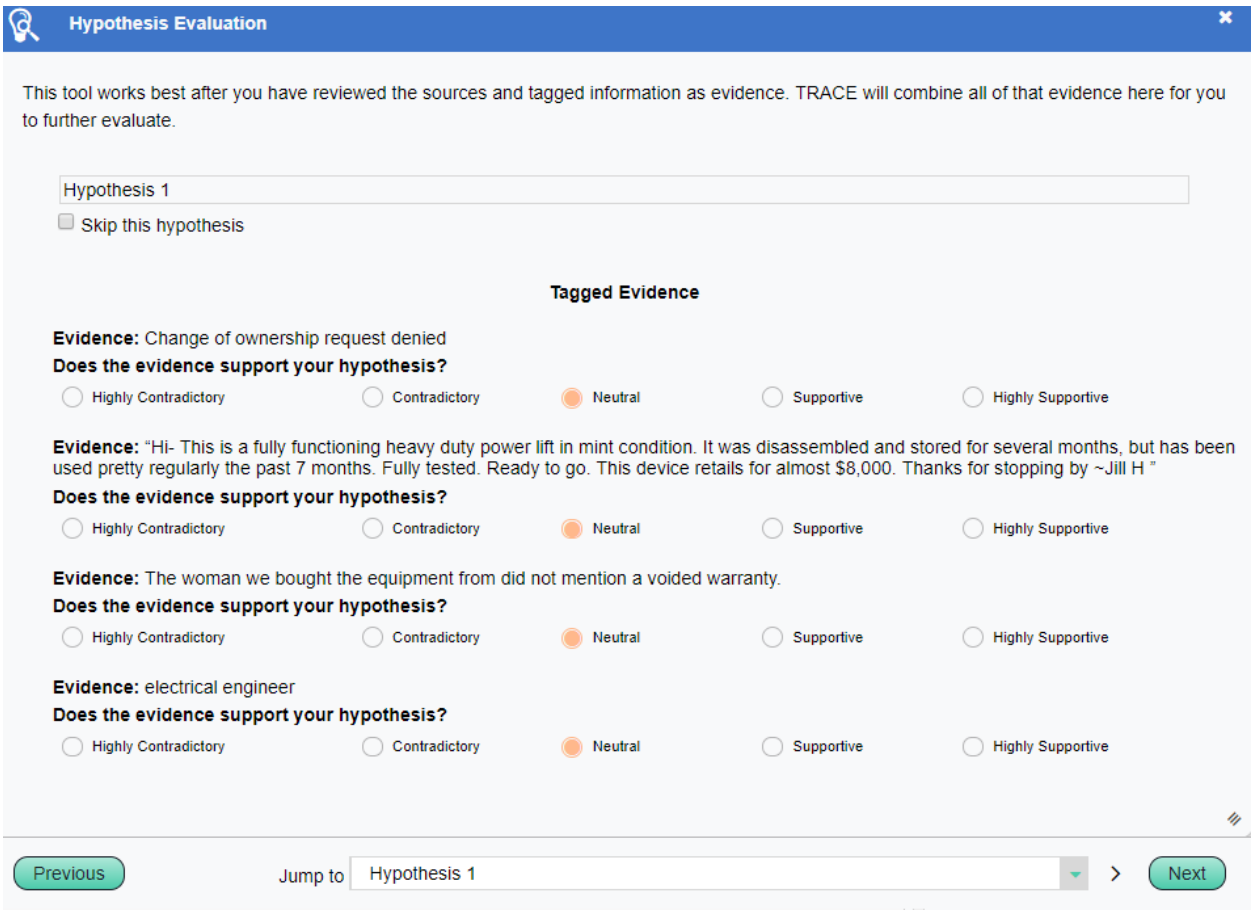

Figure 6: Mean Quality of Reasoning Scores by Condition

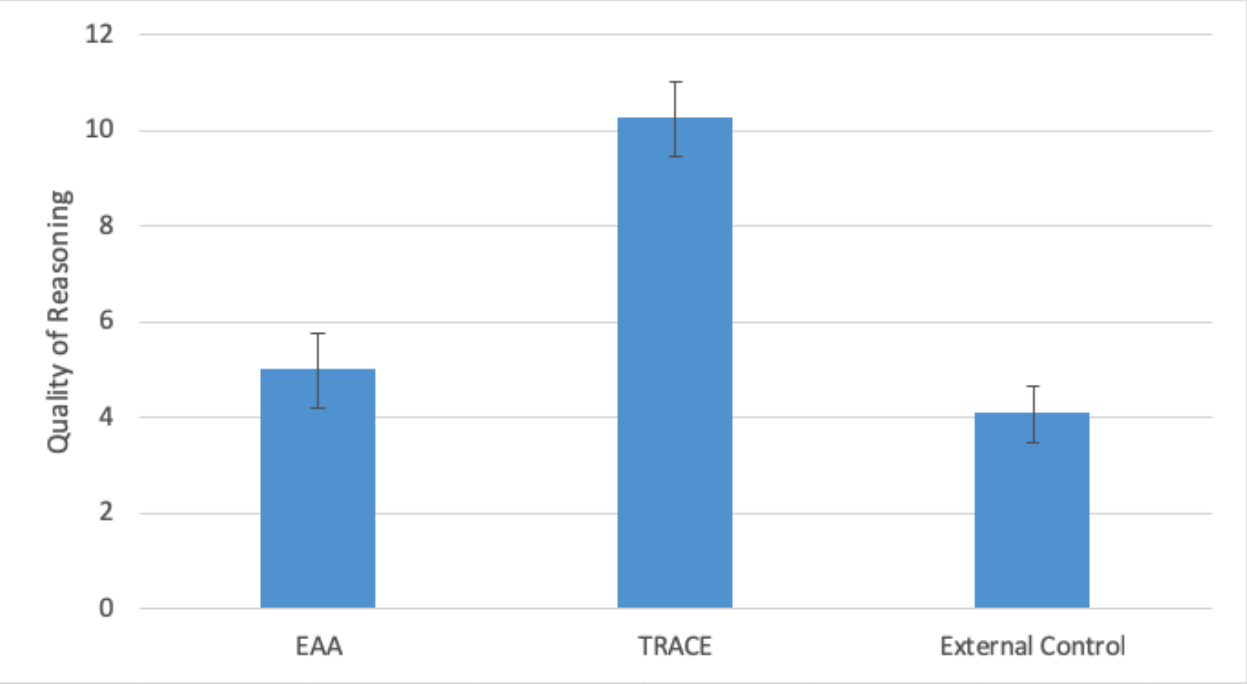

Figure 7: Logged Actions by Quality of Reasoning

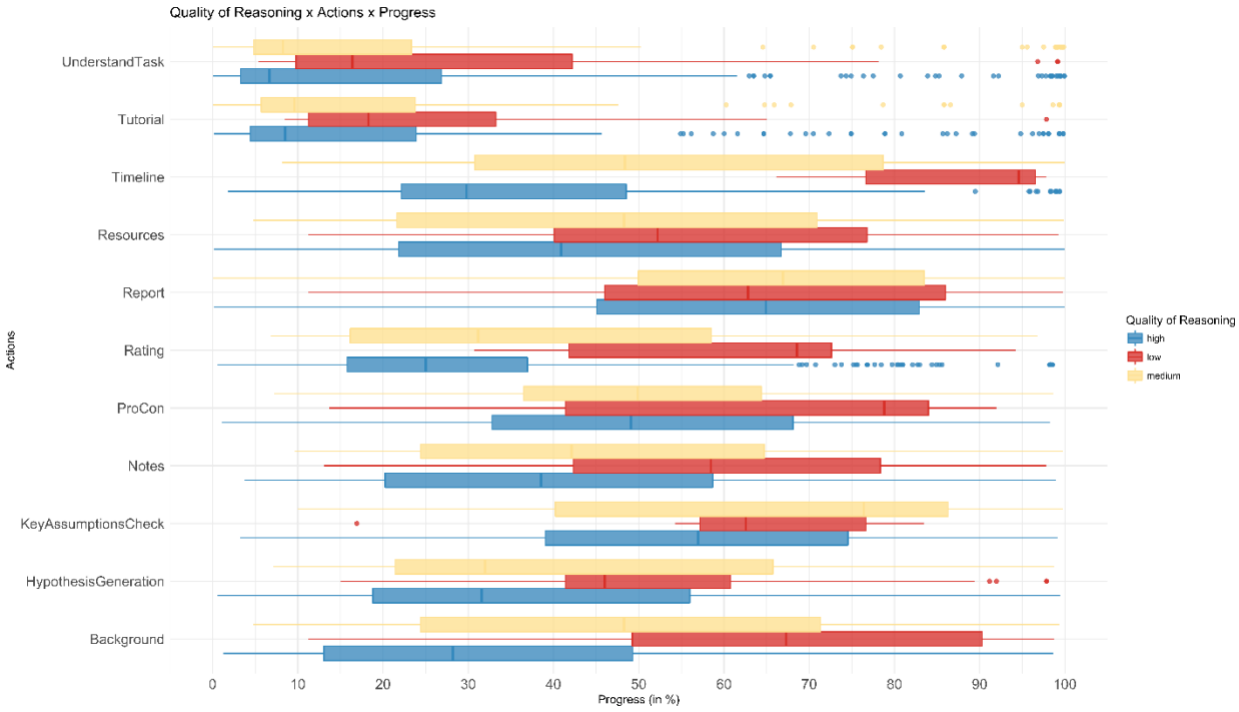

Table 1. Sample Size by Condition and Problem Assignment for Experiment 3

	EAA	TRACE	External Control	Total
Cambria Escape Route	37	38	16	91
Unusual Suspects	41	37	48	126
Which Lovell?	42	39	58	139
Total	120	114	122	356

Table 2. Mean Differences in Quality of Reasoning by Structured Technique

	Mean	SD	95% CI Lower Bound	95% CI Upper Bound	N
EAA	5.0	4.4	4.2	5.8	120
TRACE	10.2	4.3	9.4	11.0	114
External Control	4.1	3.3	3.5	4.7	122
TOTAL	6.4	4.8	5.8	6.9	356

Table 3. OLS Regression Predicting Quality of Reasoning

	Unstandardized Coefficients		Standardized Coefficients	Sig.	95% Confidence Interval for B	
	B	**Std. Error**	**Beta**		**Lower Bound**	**Upper Bound**
Structured Report	6.294	0.531	0.611	0.000	5.250	7.338
EAA	1.036	0.522	0.102	0.048	0.009	2.063
Case: Lovell	0.548	0.492	0.056	0.266	-0.420	1.516
Case Cambria	-0.373	0.558	-0.034	0.504	-1.472	0.725
Constant	3.862	0.455		0.000	2.967	4.757
R Square = 0.318						

Table 4. OLS Regression Predicting Quality of Reasoning with Controls

	Unstandardized Coefficients		Standardized Coefficients	Sig.	95% Confidence Interval for B	
	B	**Std. Error**	**Beta**		**Lower Bound**	**Upper Bound**
Constant	-3.36	2.05		0.102	-7.38	0.66
TRACE	4.41	0.48	0.43	0.000	3.47	5.36
EAA	0.91	0.44	0.09	0.042	0.03	1.78
Case: Lovell	0.27	0.42	0.03	0.513	-0.55	1.09
Case Cambria	-0.84	0.48	-0.08	0.077	-1.78	0.09
Female	0.44	0.38	0.04	0.244	-0.30	1.18
Education	1.85	1.01	0.07	0.067	-0.13	3.82
Age	0.00	0.02	0.00	0.996	-0.03	0.03
White	0.76	0.47	0.06	0.105	-0.16	1.68
Report Word Count	0.01	0.00	0.47	0.000	0.01	0.01
R Square = 0.521						