Approved for Public Release; Distribution Unlimited
Case #04-0985

MTR 04B0000017

MITRE TECHNICAL REPORT

Confirmation Bias in Complex Analyses

October 2004

Brant A. Cheikes[†]
Mark J. Brown
Paul E. Lehner
Leonard Adelman[‡]

Sponsor:	MITRE Sponsored Research		
Dept. No.:	G062	**Project No.:**	51MSR114-A4

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

Approved for public release; distribution unlimited.

©2004 The MITRE Corporation. All Rights Reserved.

MITRE
Center for Integrated Intelligence Systems
Bedford, Massachusetts

[†] E-mail: bcheikes@mitre.org

[‡] George Mason University

Abstract

Most research investigating the confirmation bias involves abstract experimental tasks where subjects draw inferences from just a few items of evidence. These tasks are not representative of complex analysis tasks characteristic of law enforcement investigations, financial analysis and intelligence analysis. This study examines the confirmation bias in a more complex analysis task and evaluates a recommended procedure, called Analysis of Competing Hypotheses (ACH), designed to mitigate confirmation bias. Results indicate that participant assessment of new evidence was significantly impacted by beliefs they held at the time evidence was received. Evidence confirming current beliefs was given more “weight” than disconfirming evidence. However, current beliefs did not influence the assessment of whether an evidence item was confirming or disconfirming. ACH did reduce confirmation bias, but the effect was limited to participants without professional analysis experience.

Table of Contents

1 Introduction

Wickens and Hollands (2000, p. 312) define the confirmation bias as a tendency “for people to seek information and cues that confirm the tentatively held hypothesis or belief, and not seek (or discount) those that support an opposite conclusion or belief.” Klayman and Ha (1987) have correctly pointed out that “positive testing” may be the only strategy to obtain critical falsifications, for example, in cases where the hypothesis is not yet defined specifically enough to be falsified by one instance. However, the concern is that in cases where the hypotheses are well defined, the tendency for people to seek confirming information might result in “cognitive tunnel vision … in which operators fail to encode or process information that is contradictory to or inconsistent with the initially formulated hypothesis” (Wickens & Holland, p. 312). This “… may be dangerous because potential risks and warning signals may be overlooked and, thus, decision fiascos may be the consequence” (Jonas, Schulz-Hardt, Frey, & Thelen, 2001, p. 557). Indeed, anecdotal evidence (e.g., the Senate Intelligence Committee Report, 2004) suggests that recent intelligence analysis failures may be due in part to confirmation bias.[1]

The concept of a confirmation bias was introduced by Wason (1960), who used a “rule identification task” such as the following (from Bazerman, 2002, p. 34):

> Imagine that the sequence of three numbers (e.g., 2-4-6) follows a rule. Your task is to diagnose that rule by writing down another sequence of 3 numbers. Your instructor will tell you whether or not your sequence follows the correct rule.

The typical result in such tasks is that people tend to generate number sequences that are consistent with (or confirm) the rule that they think is the correct rule, such as 1-3-5 if one believes the rule is “numbers that go up by two.” People seldom generate sequences that try to disconfirm the rule, which in this study was “any three ascending numbers.”

In addition to rule identification tasks, three other types of conceptual tasks have routinely been used to study the confirmation bias. One has been the “trait hypothesis-testing paradigm” (Galinsky & Moskowitz, 2000, p. 398). In this paradigm, participants are given a narrative describing a person and then asked to decide whether the person described in the narrative possesses one or more traits (e.g., self control) by either (a) selecting information that confirms or disconfirms the focal hypothesis, or (b) asking participants to remember previous information, some of which confirms or disconfirms the focal hypothesis. Another conceptual task is the “pseudo-diagnosticity” task (Evans, et al., 2002, p. 32), originally proposed by Doherty, Mynatt, Tweney, and Schiavo (1979). In this task, participants are typically asked to indicate which of three pieces of data about two hypotheses they would select to answer a particular question, with the typical finding being that they request data about the focal hypothesis (confirmation bias) because it seems (incorrectly) most diagnostic in answering the question. The last type of

[1] Although we note that such “anecdotal evidence” may itself be an instance of confirmation bias.

conceptual task is the “scientific inquiry” task (Koslowski & Maqueda, 1993, p. 105), proposed by Myatt, Doherty, and Tweney (1977). In this task, participants select a small number of science tests designed to generate data confirming or disconfirming their hypothesis.

Research with all four types of tasks has used minimal data (e.g., less than 10 data items) that did not vary in interpretability or reliability, and sometimes not even in diagnosticity. This level of artificiality raises concerns as to whether confirmation bias is characteristic of more complex analysis tasks where there is substantial evidence and the evidence items vary greatly in interpretability, reliability and diagnosticity.

One effort to experimentally investigate confirmation bias in a more representative setting was Tolcott, Marvin, and Lehner (1989), who worked with Army tactical intelligence analysts. Working in teams of two, analysts were given an initial battlefield scenario and then asked to estimate the most likely avenue of approach (of three possible) for the enemy’s attack and their degree of confidence for it on a 0 to 100 scale. They were then given three rounds of incoming intelligence data. Each round contained 15 pieces of data, three supporting each of the two most likely avenues of approach, and nine being neutral. The analysts provided a new estimate and confidence level after each round. After the third round, they rated the degree to which each of the 45 pieces of intelligence data (presented during the three rounds) supported or contradicted the avenue of approach (hypothesis) they considered most likely, on a -2 to +2 scale. Tolcott et al. (1989, p. 606) found that “Regardless of initial hypothesis, confidence was generally high and tended to increase as the situation evolved. Confirming evidence was sought, and weighted significantly higher than disconfirming evidence. Contradictory evidence was usually recognized as disconfirming, but was weighted lower than supportive evidence, was often regarded as neutral, and sometimes as deliberatively deceptive.” Consistent with Wickens & Hollands (2000, pp. 311-312), we refer to the Tolcott et al. findings that participants did not change their confidence in the initial hypothesis (even given evidence inconsistent with it) as representing an anchoring effect (or heuristic) and the greater weighting of confirming evidence as representing the confirmation bias.

This paper describes an experiment (1) to replicate the Tolcott et al. result that confirmation bias is manifest in complex analysis tasks, and if so, (2) to determine whether a procedure recommended for use in the intelligence analysis community (Heuer, 1999; Jones, 1998) successfully mitigates it.

The first goal of the experiment was to see if we could replicate the Tolcott et al. findings. Although confident, we were not certain that we would do so since (1) most confirmation-bias studies have used tasks that were conceptually different than the Tolcott et al. and (2) there are a number of studies that failed to obtain (or mitigate) the supposedly ubiquitous “confirmation bias” implied in introductory texts (e.g., Bazerman, 2002). For example, Ayton (1992) reviewed studies mitigating the confirmation bias for a rule identification task; Galinsky & Moskowitz (2000) did so for a “trait hypothesis testing” task; and Evans et al. (2002) did so for a “pseudo diagnosticity” task.

Moreover, it was not clear if the Tolcott et al. differential-weight findings were at odds with recent research on "predecision distortions" in jury decision making by Carlson and Russo (2001). The latter used the same approach as Tolcott et al. of (1) presenting initial background information and obtaining a confidence rating for the most likely hypothesis (in their case, between plaintiff and defendant), then (2) presenting rounds of new evidence (three in favor of the plaintiff and three the defendant), and (3) obtaining participants' rating of the degree to which each evidence item supported the hypothesis (plaintiff or defendant). Carlson and Russo also found a significant relationship between participants' initial confidence rating ("predecison") and subsequent coding of new evidence ("distortion"). However, Carlson and Russo only measured distortion as the difference between a participant's rating of evidence and an unbiased, mean rating of the evidence. Consequently, one cannot tell from their paper if participants' initial confidence rating caused them to (a) completely reinterpret subsequent, disconfirming evidence (e.g., participants with a confidence rating favoring the plaintiff rated subsequent evidence favoring the defendant as actually favoring the plaintiff) or (b) simply gave the evidence a lower rating (e.g., one still favoring the defendant), thereby giving it less weight as Tolcott et al. found, before making their final decision. The current study distinguishes between evidence reinterpretation and weighting.

The second goal of the study was to test the effectiveness of a procedure, called Analysis of Competing Hypotheses (ACH), proposed by Heuer (1999) and Jones (1998) to minimize or eliminate characteristics of the confirmation bias. Although ACH has eight steps, their approach revolves around developing a "hypothesis testing matrix," where the rows represent the evidence, the columns the hypotheses under consideration, and the cells the extent to which each piece of evidence is consistent or inconsistent with each hypothesis. The goals of the ACH matrix are to overcome the memory limitations affecting one's ability to keep multiple data and hypotheses in mind, and to break the tendency to focus on developing a single coherent story for explaining the evidence—a tendency which Carlson & Russo (2001) hypothesized creates predecision distortions (and presumably the confirmation bias). ACH is hypothesized to offset confirmation bias by ensuring that analysts actively rate evidence against multiple hypotheses and reminding analysts to focus on disconfirming evidence. However, the only experiment testing the effectiveness of ACH found mixed results (Folker, 1999): it helped intelligence analysts identify the correct answer to one problem, but not another. (We note that both problems had fewer than 20 evidence items.) No experiment has directly tested ACH's ability to mitigate the confirmation bias.

2 Method

This section describes the participants, procedures, and information manipulation used to conduct the experiment.

2.1 Participants

Twenty-four (24) employees of a large research and development corporation volunteered to participate in an experiment evaluating structured argumentation methods. Twenty were male, four female. All participants were interested in intelligence analysis, with 12 of the participants having intelligence analysis experience (ranging from 1 to 18 years, with a median of 9.5 years). Participants' ages ranged from 27 to 63, with a median of 47.50. Of the 23 participants who indicated their education, 22 had completed college, with 12 having a masters' degree, three a Ph.D, and one an M.D. Sixteen of the participants majored in math, physics, computer science or engineering.

2.2 Procedures

The entire experiment was conducted via email, and all data was collected within a two-month period. Participants were randomly assigned so that there were 12 in the ACH condition and 12 in the non-ACH condition. All participants started with an email providing a general description of what they would do and a request to complete all materials within one sitting, which was estimated to be (and was) two hours or less. The specific procedures for the ACH and non-ACH groups are described next.

2.2.1 Procedures for ACH Group

After reading the general description of the study, ACH participants opened a file containing the ACH Tutorial, which was modeled after Folker's (1999). The tutorial described the general concepts and steps of the ACH procedure, and emphasized the construction of the hypothesis-testing matrix and the importance of disconfirming (versus confirming) evidence. The hypotheses were the columns of the matrix, and the individual evidence items were the rows. Cell entries indicated the extent to which the evidence supported a hypothesis using an integer scale ranging from -2 ("evidence strongly supports conclusion that hypothesis is false") to +2 ("evidence strongly supports conclusion that hypothesis is true"), with 0 indicating that the "evidence neither supports nor contradicts hypothesis." Consistent with the ACH presentations in Folker (1999), Heuer (1999), and Jones (1998), participants were given no guidance regarding how they might or should combine the numeric cell entries into an overall conclusion. Rather, the tutorial emphasized focusing on disconfirming evidence when assessing the overall relative likelihood of the competing hypotheses, consistent with ACH proponents. Then, the tutorial gave two increasingly complex examples, with the participants being asked to construct a hypothesis-testing matrix for the second example problem. After completing the matrix, the participants could go to the next page to see the experimenters' answers. All participants completed the matrix and we assume, but cannot verify, that they did so before seeing our answers.

After completing the tutorial the ACH participants performed the experimental task, which was to assess the relative likelihood for three hypothesized causes of the explosion on the battleship USS Iowa in April 1989. The Iowa Explosion is the principal exercise for practicing ACH in Jones (1998, pp. 209-216). Consistent with complex analysis tasks, this exercise has multiple hypotheses and potentially a great deal of evidence. We used 60 pieces of evidence, some of which were as short as one sentence but most of which were one paragraph (varying from 2 to 6 sentences) in length. The evidence covered various topics (e.g., mechanics, electronics, and psychological diagnoses) and contained conflicting expert testimony, as is typical of complex analysis tasks in intelligence and financial analyses and law enforcement investigations. (Our evidence came partly from Jones' book, and partly from open source materials in books and on the Internet.) Most of the evidence was reliable, with only three cases where later evidence contradicted earlier evidence. The evidence varied considerably in its diagnostic value, but there was no case where the same piece of evidence was intended by the experimenters to be consistent (or inconsistent) with more than one hypothesis.

Procedurally, participants first received background material describing the Iowa class battleship, the 16-inch gun turrets (with schematic) where the explosion occurred, the job of each turret crew member, how the guns were loaded and fired (with schematic), and the reasons for the Navy's three hypothesized causes of the explosion, which were:

- **An overram ignited powder (H1)** – The rammerman, Robert Backherms, was unqualified for the rammerman position and inadvertently caused the mechanical rammer to shove powder bags too fast and/or too far into the gun chamber, forcing the bags against the base of the projectile. The impact caused the powder to explode.
- **Friction ignited powder (H2)** – Friction produced a buildup of static electricity inside the gun chamber while the powder bags were being rammed into position by the rammerman. Static electricity caused a spark which then ignited the powder in the gun chamber.
- **Gun captain placed incendiary device (H3)** – Intent on killing himself and others, Clayton Hartwig discreetly hid a small incendiary device between powder bags just before ramming. Ramming detonated the device, which then ignited the powder.

Participants then received fifteen initial pieces of evidence. In addition, ACH participants were given a blank hypothesis-testing matrix with the hypothesized causes of the explosion as the columns and the evidence item numbers as the rows. As they read each piece of evidence, ACH participants scored the extent to which it supported each of the three hypotheses using the -2 to +2 scale introduced in the tutorial.

After considering the background information and the first 15 evidence items, the ACH participants were asked to indicate their level of confidence in each hypothesized cause of the explosion by allocating 100 points among the three hypotheses. Participants were told to give the most points to the hypothesis they considered most likely, the fewest points to the hypothesis they considered least likely, and to reflect the relative likelihood between hypotheses by the ratio of points between them. Three examples were provided to illustrate how their points indicated the relative likelihood of the hypotheses.

After completing their confidence ratings for the hypotheses (based on the background information and the first 15 pieces of evidence), the ACH participants received three additional rounds of evidence, with 15 pieces of evidence in each round, for a total of 60 pieces of evidence. The ACH participants continued using the matrix they had started for the first round of evidence and, after reading each piece of evidence, rated the extent to which it supported each hypothesis using the -2 to +2 scale described earlier. After reading the evidence for each round (i.e., after the 30th, 45th, and 60th piece of evidence), participants gave their confidence ratings for the hypotheses (by allocating 100 points among the three hypotheses) based on all the evidence they had received to that point.

The ACH participants completed a confidential biographical questionnaire as their last activity. The questionnaire requested their corporate location, job title, age, gender, education, major area of study, whether or not they had taken college courses or training workshops (or seminars) in argumentation, reasoning, or logic, if they had worked as an intelligence analyst (and if so, how many years), amount of prior ACH experience, amount of prior knowledge about the Iowa explosion, and asked them to provide a general description of how they made their confidence ratings.

2.2.2 Procedures for Non-ACH Group

The non-ACH participants followed the same procedures as the ACH participants with a few exceptions. First, they did not receive the ACH tutorial. Instead, they began by receiving the background description for the Iowa explosion and the first 15 pieces of evidence. Second, they were not given the hypothesis-testing matrix to complete while reading the evidence. Instead, they simply read the first 15 pieces of evidence and then provided their confidence ratings for the three hypotheses (using the same confidence rating form as used by the ACH participants). Then they received the three rounds of evidence, in turn. The non-ACH participants did not complete the hypothesis-testing matrix while reading the evidence. They simply read the evidence and provided their confidence ratings after each round of evidence, that is, after reading the 30th, 45th, and 60th pieces of evidence. After providing their fourth (and last) confidence ratings, the non-ACH participants completed the hypothesis-testing matrix for all 60 evidence items. The non-ACH participants also completed the biographical questionnaire as their last activity.

2.3 Information Manipulation

The cause of the Iowa Explosion has never been determined conclusively. Consequently, all three hypotheses are truly viable hypothesized causes of the explosion, with considerable information in the literature (Schwoebel, 1999; Thompson, 1999) supporting each one. This made it possible for us to construct the evidence to directly test for an anchoring effect and confirmation bias. This was accomplishing by having H1 and H3 receive the most confirming evidence in the first two rounds, while having H2 end up with the least amount of disconfirming evidence (and highest ratio of confirming to disconfirming evidence) by the end of round four. In addition, the background material and first round of evidence were constructed to make H1 and H3 easiest to visualize. Consequently, we expected that the mean confidence ratings would be highest for H1 and H3 after round 1.

The evidence items used in this study were selected from a larger pool of potential items on the basis of the authors' independent ratings of the items using the -2 to +2 scale described above. We discarded potential items where the authors disagreed on whether the item was confirming or disconfirming of a hypothesis, as well as any items where the authors felt that the evidence confirmed or disconfirmed more that one hypothesis. Table 1 presents the authors' mean diagnosticity for the evidence items both by round and in total, with the number of evidence items confirming or disconfirming a hypothesis shown in parentheses. Below we will refer to the authors' evidence ratings as the a priori evidence ratings.

As shown in Table 1, no disconfirming (negative) evidence was presented in the first two rounds. Moreover, the evidence in rounds 1 and 2 supported the Overram (H1) and Hartwig (H3) hypotheses over Friction (H2). However, more evidence disconfirming H1 and H3 than H2 was presented in rounds 3 and 4, such that in total, H2 was (in our opinion) more likely than H1 and H3. It should be noted that half the participants randomly received the above ordering of the hypotheses, and the other half received an ordering where Friction was the first hypothesis and Overram the second. Henceforth, however, Overram will always be H1, Friction H2, and Hartwig H3.

Table 1. Experimenters' Assessment of Degree of Support for Each Hypothesis

	Overram (H1)		Friction (H2)		Hartwig (H3)	
	Confirms	Disconfirms	Confirms	Disconfirms	Confirms	Disconf.
Round 1	5.25 (4)		2.25 (2)		5.75 (5)	
Round 2	2.5 (2)		2.75 (2)		2.5 (3)	
Round 3	1.75 (2)	-2.25 (2)	3.5 (3)	-1.25 (1)	1.75 (1)	-2.5 (2)
Round 4	2.5 (2)	-2.5 (2)	2.5 (2)	-0.75 (1)	2.5 (2)	-2.25 (3)
Total:	12.0 (10)	-4.75 (4)	11.0 (9)	-2.0 (2)	12.5 (11)	-4.75 (5)
Ratio	2.52 (2.5)		5.50 (4.5)		2.63 (2.2)	

3 Results

The results testing for an anchoring effect and confirmation bias are presented, in turn.

3.1 Anchoring Effect

As noted above, the anchoring effect is a tendency to resist change after an initial hypothesis is formed. In this experiment, such an effect would manifest itself as a tendency to maintain high confidence values for the hypothesis that received the highest value after the first round of evidence. More specifically, if there was a strong anchoring effect for the non-ACH group, one would expect the highest mean confidence ratings to remain on H1 or H3, not H2, after round 4. In contrast, if ACH was effective in overcoming an anchoring effect, then one would expect the highest mean confidence rating for the ACH group to shift to H2 after round 4. Table 2 presents the mean confidence ratings for the ACH and non-ACH groups after each round of evidence. As can be seen, the highest mean confidence rating is on H1 for both groups. Although the mean H2 confidence rating moved more for the ACH than the non-ACH group, a 2 (ACH versus non-ACH) x 2 (Round 1 versus Round 4) Analysis of Variance (ANOVA) indicated that the amount of movement was not significantly different: $F(1,22) = 0.13$, $p > 0.05$. The Round main effect [$F(1,22) = 1.06$] and Group x Round interaction [$F(1,22) = 2.34$] were also not significant.

Table 2. Mean Confidence Ratings for Both Groups by Rounds

	ACH			Non-ACH		
Round #	H1	H2	H3	H1	H2	H3
1	42.6	22.9	34.5	38.3	30.2	31.5
2	40.1	21.9	38.3	44.3	28.9	26.8
3	35.4	30.8	33.0	35.3	37.2	27.5
4	36.4	31.6	32.0	42.8	28.5	28.7

A second ANOVA was performed to test whether the confidence rating for the hypothesis that participants considered most likely after round 1 was, on average, significantly lower after round 4 for the ACH than the non-ACH group. The particular hypothesis was irrelevant in this test because we simply looked at the (round 4 – round 1) difference for whatever hypothesis a participant considered most likely after round 1. (This test was used by Tolcott et al. because they did not manipulate the information to support one hypothesis over the other.) Again, ACH failed to mitigate any anchoring effect: $\text{mean}_{\text{ach}} = 47.4$, $\text{mean}_{\text{non-ach}} = 50.7$, $F(1,22) = 0.35$. The Round main effect [$F(1,22) = 0.70$] and Group x Round interaction [$F(1,22) = 0.0$] were also not significant.

Overall, the results appear consistent with (1) the existence of an anchoring effect and (2) that ACH did not significantly reduce that effect. However there are two reasons why the support for both results should be considered weak. First, these results depend on the authors' a priori assessment of the weight of the evidence. If participants disagree with our ratings (and they did) then they may be adjusting their confidence assessments appropriately to their perception of the evidence. Second, there was considerable variation amongst subjects in their confidence ratings. For example, the round 4 minus round 1 difference in participants' confidence ratings for H2 ranged from -42 (this participant began with a confidence rating of 72 on H2 and ended with a 30 on it) to +40 (began with a confidence rating of 20 on H2 and ended with a 60). The standard deviation for the difference in the confidence ratings between rounds 4 and 1 was 17.2. This variation made it difficult to find significant ANOVA effects, given our sample size.

The confirmation bias results are far more definitive.

3.2 Confirmation Bias

Recall that a confirmation bias is a tendency to seek confirming evidence and/or to bias the assessment of available evidence in a direction that supports preferred hypotheses. This experiment focused on the latter. Specifically, if there was biased assessment of evidence, we would expect a positive correlation between participant confidence ratings in each round and their evidence ratings in the next round of evidence.

Table 3. Correlation of Confidence and Evidence Evaluation Ratings

	Rnd 1 Confid. with Rnd 2 Eval			Rnd 2 Confid. with Rnd 3 Eval			Rnd 3 Confid. with Rnd 4 Eval			Combined Analysis		
	H1	H2	H3	H1	H2	H3	H1	H2	H3	#Obs	Comb r	p
Total	.050	.162	**.585***	.128	**.401**	**.437**	.165	.141	**.575***	216	**.308***	**.000***
ACH	.126	.331	.290	-.122	**.571**	.483	-.073	.121	**.628**	108	**.282***	**.002***
Non-ACH	.027	.032	**.832***	.219	.056	**.641**	.336	.016	**.624**	108	**.359***	**.000***

Table 3 presents the correlations between (1) participants' confidence rating for each hypothesis after each of the first three rounds of evidence (e.g., after round 1) and (2) their summative rating for the evidence items in the next round (e.g., in round 2). The top row shows the correlations for all 24 participants; the last two rows for the 12 ACH and 12 non-ACH participants, respectively. For the overall results, correlations greater than 0.344 (n = 24) and 0.497 (n = 12) have a p-value (1-tail) less than 0.05 (shown in bold). For the ACH and non-ACH rows, correlations greater than 0.515 (for n = 24) and 0.661 (for n = 12) have a p-value (1-tail) less than 0.005 (shown in bold, followed by an asterisk).

The overall results (top row) show a consistent positive correlation with confidence ratings and later evidence ratings; four of the nine correlations exceeded 0.05 significance. The last column of Table 3 shows the results of a combined analysis that applies the Stoufer Combined Test against the Fisher-Z transform of the nine individual correlations (see Lipsey & Wilson, 2001; Wolf, 1986). This test uses all 216 observations (24 participants x 9 correlations) and 108 observations for the ACH and non-ACH groups. Clearly the combined test showed a significant positive correlation for both the ACH and non-ACH groups.[2] The combined correlations for the ACH (0.282) and non-ACH (0.359) groups were not statistically different.

Given that there was strong evidence of bias in the evidence ratings, we examined the nature of the bias. Specifically, we were interested in the extent to which the bias manifested itself as a distortion effect (e.g., misinterpreting evidence as confirming a preferred hypothesis when it "should" be disconfirming) or as a weighting effect (e.g., giving a +2 rating to evidence supporting a preferred hypothesis and a +1 or 0 to comparable evidence confirming a non-preferred hypothesis.) Below we examine both distortion and weighting effects in turn.

3.2.1 Distortions

Below we examine two types of possible distortions.

3.2.1.1 Analysis of Distortions of Diagnostic Evidence

For purposes of this analysis we define diagnostic evidence for a hypothesis as evidence to which the researchers a priori assigned a rating other than zero. A diagnostic distortion occurs when a participant rates the evidence in the opposite direction of the a priori rating.

Overall, diagnostic distortions were rare. Eleven of the 14 were positive distortions, where participants assigned a positive rating to an evidence item that received an a priori negative rating (p<.05, 2-tailed). In evidence rounds 2, 3 and 4, there were thirteen distortions. Nearly all of these (11 of 13) were associated with a currently preferred hypothesis (p<.01, 2-tailed), which is the hypothesis that the participant rated as the most likely in the previous round of confidence ratings. Of the distortion associated with currently preferred hypotheses, most were positive (9 of 11). While this appears significant, we note that both of the distortions associated with non-preferred hypotheses were positive (2 of 2).

In sum, participants exhibited a tendency toward positive distortion and to distort evidence associated with their preferred hypothesis. However, there is insufficient data to test whether or not the tendency to distort evidence positively is stronger for preferred hypotheses than for non-preferred hypotheses. The analysis of the non-diagnostic evidence items resolves this.

3.2.1.2 Analysis of Distortions of Non-Diagnostic Evidence

An evidence item is "non-diagnostic" of a hypothesis when the researchers' a priori rating of that evidence is zero, that is, when all four experimenters independently assessed the evidence as neither confirming nor disconfirming a hypothesis. We felt that if participants we going to exhibit an interpretation bias, it would be with these evidence items. This is because these items are, in

[2] We note here that the significance of the combined correlations was also tested using the Fisher Combined Test. This test combines one-tailed significance p-values into an overall significance test. We converted the individual correlations into their corresponding p-values and then applied the Fisher Combined Test to estimate the combined p-value. The results showed the same pattern of significance as the Stoufer Combined Test for this and all subsequent analyses reported in this paper. Therefore, only the results for the Stoufer Combined Test are presented.

effect, a tabula rasa upon which a participant could project their beliefs. Assigning either a positive or negative rating to a non-diagnostic item is defined as a non-diagnostic distortion.

Non-diagnostic distortions were more common than diagnostic distortions, occurring 247 times out of 3336 opportunities. Most distortions were positive (144 of 247, p<.01, 2-tailed). When examining the distortions that occurred in evidence rounds 2 through 4, we find that unlike diagnostic distortions, most of the non-diagnostic distortions were not associated with the currently preferred hypothesis (70 of 179). For currently preferred hypotheses, 48 of 70 (69%) were positive, while for the non-preferred hypotheses 66 out of 109 (61%) were positive. The difference between these proportions is not statistically significant.

ACH participants exhibited more non-diagnostic distortions (101 of 1668) than non-ACH participants (78 of 1668), but this difference is not statistically significant (p>.05, 1-tailed). However, non-ACH participants exhibited positive distortions (61 of 78) substantially more often than ACH participants (53 of 101). This difference is significant (p<.001, 2-tailed). Interestingly this effect was mostly due to the assessment of non-preferred hypotheses, where for ACH participants the minority of non-diagnostic distortions were positive (29 of 62) while non-ACH participants exhibited mostly positive distortions (37 of 47).

3.2.1.3 Summary of Distortion Analysis

When non-ACH participants misinterpreted evidence, they showed a strong tendency toward positive distortions over negative distortions. However, this tendency toward positive distortions was not significantly related to prior confidence ratings, and so is not evidence of confirmation bias. There was no evidence that ACH participants exhibited this tendency.

3.2.2 Weighting

Product-moment correlations (r) were used to assess the strength of the relationship between participants' confidence and evidence ratings, and are presented below. Using correlations to represent evidence weighting has a long history in judgment and decision making research; for example, see Cooksey (1996) and Hammond et al. (1975).

Table 4 shows the correlations for confidence ratings and later evidence ratings for diagnostic evidence where the 14 distortions were removed—in other words, for the 970 ratings where the ratings were all in the same direction or neutral.[3] Since all ratings are in the same direction, positive correlations indicate a tendency to shift evidence ratings toward more preferred hypotheses (e.g., rate evidence as "+2" rather than as "+1" or as "+1" rather than "0"). As Table 4 shows, there is a strong tendency to shift ratings toward more preferred hypotheses.

[3] We note that all of the correlation results are substantially the same whether or not the 14 distortions, out of a possible 984, are removed.

Table 4. Correlation of Confidence and Diagnostic Evidence Ratings with Distortions Removed

	Rnd 1 Confid. with Rnd 2 Eval			Rnd 2 Confid. with Rnd 3 Eval			Rnd 3 Confid. with Rnd 4 Eval			Combined Analysis		
	H1	H2	H3	H1	H2	H3	H1	H2	H3	#Obs	Comb Rho	p
Total	.256	.118	**.442**	.150	**.576***	**.489**	**.409**	.309	**.611***	216	**.386***	**.000***
ACH	.200	.408	.002	-.162	**.636**	.401	-.006	.491	**.511**	108	**.296***	**.001***
Non-ACH	.402	-.195	**.768***	.322	.417	**.651**	**.604**	.107	**.693***	108	**.459***	**.000***

3.2.3 Additional Analysis

3.2.3.1 Impact of Analysis Experience

Twelve of our participants had professional analysis experience (specifically intelligence analysis) and 12 did not.[4] We examined the difference between these groups.

T-tests found no difference between the groups in their mean age, amount of prior knowledge about the Iowa explosion or amount of ACH experience. Proportion tests found no difference in their gender, education or logic courses. There was no difference between the groups on their average confidence or evidence ratings, but there was substantial difference between theses groups on the variation of their ratings. For all twelve confidence ratings, the average variance of confidence ratings for the 12 experienced analysts was greater than for the participants without experience. Of the 120 evidence items, the average variance for the experienced participants was greater than for the inexperienced 84 times, while the inexperienced showed greater average variance 31 times ($p<.001$, 2-tailed).

The Stoufer Combined Test using all evidence indicated that both groups had significant correlations between their confidence and evidence ratings: with experience ($r = 0.234$, $n = 108$, $p = 0.012$) and without experience ($r = 0.468$, $n = 108$, $p < 0.001$). The confidence-evidence relationship was, however, significantly weaker for participants with intelligence analysis experience ($z = -1.95$, $n_1 = n_2 = 108$, $p = 0.05$, 2-tailed). In addition, there was no difference in the correlations for the 5 ACH and 7 non-ACH participants with intelligence analysis experience: 0.303 versus 0.176, respectively, $z = 0.67$, $n_1 = 45$, $n_2 = 63$, $p > 0.05$. In contrast, there was a significant difference in the correlations for the 7 ACH and 5 non-ACH participants without analysis experience: 0.193 versus 0.774, respectively, $z = -4.15$, $n_1 = 63$, $n_2 = 45$, $p < 0.001$, 2-tailed. Participants without intelligence analysis experience using ACH had substantially smaller confidence-evidence correlations.

[4] We note that the inclusion of 12 analysts in our study was fortuitous. Subjects were not selected on the basis of analysis experience.

3.2.3.2 When Evidence Ratings are Determined

Recall that the principal difference between the ACH and non-ACH condition is when participants rated evidence. ACH participants rated evidence when they received it, while non-ACH participants rated evidence at the end. This raises the question of the extent to which participants in the non-ACH condition may have revised their rating of evidence between the time they initially saw the evidence and the time that they finally recorded their rating. To address this issue, we attempted to test two contrasting hypotheses:

H_E: Evidence ratings were determined when participants first observed the evidence, irrespective of when they recorded that rating. A delay in recording the rating did not impact that rating.

H_R: Evidence ratings were determined when the participants recorded those ratings. A delay in recording may change the rating.

Let $R_{a\text{-}i}$ be the correlation between the initial (or prior) confidence ratings and the evidence ratings for the ACH group, and $R_{a\text{-}f}$ the correlation between the final confidence ratings and evidence ratings, for the ACH group. Let $R_{n\text{-}i}$ and $R_{n\text{-}f}$ be the same correlations for the non-ACH group. The first hypothesis predicts $R_{a\text{-}i}=R_{n\text{-}i}$ and $R_{a\text{-}f}=R_{n\text{-}f}$ (i.e., no difference in the confidence-evidence correlations between the ACH and non-ACH groups, whether using the initial or final confidence ratings) because it claims that evidence ratings are determined when the evidence is observed, not when the ratings are recorded. Unfortunately this prediction is confounded with the hypothesis that ACH reduces confirmation bias, i.e., $R_{a\text{-}i}<R_{n\text{-}i}$. This suggests that the first correlation only be examined for participants for which ACH did not reduce confirmation bias, namely, the experienced analysts. In contrast, the second hypothesis predicts that the confidence-evidence correlations will be significantly higher based on the final confidence ratings (i.e., $R_{n\text{-}f}>R_{n\text{-}i}$) because it claims that what participants believed at the time they recorded the rating determined their rating and not what they believed earlier (i.e., when they first saw the evidence).

Table 5 shows all the relevant correlations for the experienced analysts using all the data (Stoufer Combined Test). With respect to H_E, there was no difference between $R_{a\text{-}i}$ and $R_{n\text{-}i}$, or between $R_{a\text{-}f}$ and $R_{n\text{-}f}$, for any comparison. This is consistent with H_E. With respect to H_R, there was no case where $R_{a\text{-}f}$ was significantly higher than $R_{n\text{-}f}$ as would be expected if the non-ACH participants' evidence ratings were affected by their final confidence ratings. This is inconsistent with H_R. Overall the data appears consistent with the view that evidence ratings were determined when the evidence was first observed and not when it was later recorded.

Table 5. Mean Correlations Using All Data for Participants With Experience

	C1-R2	R2-Cf	sig	C2-R3	R3-Cf	sig	C3-R4	R4-Cf	sig
ACH	0.256	0.403	0.43	0.438	0.599	0.41	0.205	0.394	0.42
NON-ACH	0.101	0.222	0.43	0.050	0.227	0.40	0.364	0.488	0.41
significance	0.57	0.59		0.69	0.70		0.42	0.45	

4 Discussion

Summarizing the results, there were several key findings. There was some evidence of an anchoring effect, and no evidence that ACH mitigated any anchoring effect that might have been present. There was a strong tendency to "misinterpret" or "distort" evidence as confirming a hypothesis rather than disconfirming, but there was little evidence to suggest that such distortions were biased toward confirming hypotheses that subjects viewed as the most likely. ACH substantially reduced this tendency. There was a strong weighting effect. When participants agreed on the interpretation of the evidence, they tended to give stronger evidence ratings to hypotheses with higher confidence ratings. For participants without professional analysis experience, ACH substantially reduced this weighting effect. For participants with professional analysis experience, ACH had no impact at all on this weighting effect. In short, confirmation bias was manifested principally as a weighting bias and not as a distortion bias, and the ACH results were mixed.

These results replicate the findings reported in Tolcott et al. (1989), the only study we found examining the confirmation bias for a complex analysis task. Tolcott et al. also found an anchoring effect and a confirmation bias due to participants' weighting of the evidence, not distortions. In addition, the results clarify Carlson and Russo's (2001) findings for a legal inquiry task. Carlson and Russo found "predecision distortions," but they measured distortion as the difference between a participant's rating of evidence and the unbiased, mean rating of the evidence. This measure did not distinguish between cases where participants completely reinterpreted the diagnostic evidence item's meaning or simply gave it less weight. The current study showed that reinterpretations of diagnostic evidence are rare, and suggests that Carlson and Russo's predecision distortions were probably due to participants differentially weighting evidence depending upon whether the evidence supported hypotheses they considered more or less likely.

It was not clear what cognitive processes led to the observed anchoring and confirmation-bias effect. Consistent with Carlson and Russo's (2001) hypothesis and Pennington and Hastie's (1993) theory of explanation-based decision making, it could have been the result of participants' efforts to develop a coherent story or explanation for their prior beliefs and subsequent conclusions (confidence ratings). On the other hand, it could have represented the use of an anchoring and adjustment heuristic, as postulated by Tversky & Kahneman (1974), with insufficient adjustment away from the anchor because evidence confirming current beliefs was given more weight than disconfirming evidence. Future research, probably using protocol analysis, could address this issue.

The anchoring and weighting effects represent a primacy effect: prior information and beliefs overwhelmed the impact of later, disconfirming information. This finding is inconsistent with much of the "order effects" research (e.g., Hogarth & Einhorn, 1992) showing a recency effect, where more recent as opposed to prior information has a bigger effect on one's decision. This inconsistency may exist because most "order effects" research involved tasks with minimal amounts of evidence; for example, Hogarth & Einhorn's experiments manipulated two critical pieces of evidence. In addition, order effects research by Adelman et al. (1997) showed one could obtain primacy or recency effects depending on whether or not situation-specific features permitted people to "explain away" and thereby give less weight to (or reinterpret) recent, disconfirming information. However, even the Adelman et al. task had only five critical pieces of evidence. Therefore, future confirmation-bias and order-effects research both need to use complex analysis problems with large amounts of evidence.

ACH is intended to mitigate confirmation bias in intelligence analysts. Unfortunately, there is no evidence that ACH reliably achieves this intended effect. Beyond that, however, ACH had some interesting effects. ACH and non-ACH participants misinterpreted evidence with equal frequency, but ACH participants were balanced in whether they misinterpreted evidence as confirming or disconfirming, while non-ACH participants showed a strong tendency to misinterpret evidence as confirming. This tendency was not related to which hypotheses the participants currently believed. ACH did impact the tendency to give more weight to evidence confirming current beliefs, but this effect was limited to subjects without analysis experience. Mixed results for ACH were also found in Folker's (1999) study. Taken together these studies suggest that ACH, in its current form, does not provide a robust method for mitigating the confirmation bias. Nevertheless, in our opinion there are sufficient results to warrant a systematic investigation into variations of the ACH approach in hopes of finding a robust debiasing method.

Lastly, future confirmation-bias research using complex analysis tasks needs to systematically evaluate the effect of larger, less reliable data sets. Although the current study had 60 pieces of evidence (Tolcott et al. had 45) varying in their interpretability (because of conflicting expert testimony), complex analysis problems can have hundreds, if not thousands of such evidence items. Moreover, the evidence may vary in reliability in ways we did not manipulate in our study. And in some situations, evidence can seem to support two or more hypotheses, which was not the case in our study, and vary widely in its diagnostic value. These are task characteristics that certainly might lead one to err in seeing coherence that is not there (Hammond, 1996), as represented by the confirmation bias. Since these are the characteristics of the complex tasks facing, for example, intelligence and financial analysts and law enforcement investigators, they need to be represented in experimental tasks to more fully understand the frequency and nature of the confirmation bias and the cognitive processes that might cause it.

As noted in the Introduction, many studies have failed to replicate the confirmation bias using the four conceptual tasks typically used to study it. Complex analysis tasks need to be subjected to the same scrutiny. However, until such research is conducted, it appears that complex analysis tasks are subject to an anchoring effect and confirmation bias. Hopefully, this will not result in the decision fiascos feared by Jonas et al. (2001). Thus far, however, our experimental results and anecdotal, retrospective reports (e.g. the Senate Intelligence Committee's report on the Intelligence Community's prewar assessments of Iraq) suggest cause for concern.

References

Adelman, L., Bresnick, T.A., Christian, M., Gualtieri, J., & Minionis, D. (1997). Demonstrating the effect of context on order effects for an Army air defense task using the Patriot simulator. *Journal of Behavioral Decision Making, 10*, 327-342.

Ayton, P. (1992). On the competence and incompetence of experts. In G. Wright & F. Bolger (Eds.), *Expertise and decision support* (pp. 77-105). NY: Plenum Press.

Bazerman, M.H. (2002). *Judgment in managerial decision making* (5th Ed.). NY: Wiley.

Carlson, K.A., & Russo, J.E. (2001). Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied, 7(2)*, 91-103.

Cooksey, R.W. (1996). *Judgment analysis: Theory, method, and applications*. NY: Academic Press.

Doherty, M.E., Mynatt, C.R., Tweney, R.D., & Schiavo, M.D. (1979). Pseudodiagnosticity. *Acta Psychologica, 43*, 11-21.

Evans, J. St. B.T., Venn, S., & Feeney, A. (2002). Implicit and explicit processes in a hypothesis testing task. *British Journal of Psychology, 93*, 31-46.

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 288-299.

Folker, R.D. Jr. (1999). *Exploiting structured methodologies to improve qualitative intelligence analysis*. Masters thesis: Joint Military Intelligence College.

Galinsky, A.D., & Moskowitz, G.B., (2000). Counterfactuals as behavioral primes: Priming the simulation heuristic and consideration of alternatives. Journal of *Experimental Social Psychology, 36*, 384-409.

Hammond, K.R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, and unavoidable injustice*. NY: Oxford University Press.

Hammond, K.R., Stewart, T.R., Brehmer, B., & Steinmann, D.O. (1975). Social judgment theory (pp. 271-312). In M. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes*. NY: Academic Press.

Heuer, J.R. (1999). *The psychology of intelligence analysis.* Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency.

Hogarth, R.M., & Einhorn, H.J. (1992). Order effects in belief updating: The belief adjustment model. *Cognitive Psychology, 24*, 1-55.

Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology, 80(4)*, 557-571.

Jones, M.D. (1998). *The thinker's toolkit.* NY: Three Rivers Press.

Klayman, J., & Ha, Y-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Koslowski, B., & Maqueda, M. (1993). What is confirmation bias and when do people actually have it. *Merrill-Palmer Quarterly, 39(1)*, 104-130.

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Mynatt, C.R., Doherty, M.E., & Tweney, R.D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific influence. *Quarterly Journal of Experimental Psychology, 29*, 85-95.

Pennington, N., & Hastie, R. (1993). Reasoning in explanation-based decision making. *Cognition, 49*, 123-163.

Schwoebel, R.L. (1999). *Explosion aboard the Iowa*. Annapolis, MD: Naval Institute Press.

Thompson, C.C. (1999). *A glimpse of hell: The explosion on the U.S.S. Iowa & its cover-up*. NY: W.W. Norton & Company.

Tversky, A., & Kahneman, D. (1974). Judgments under uncertainty. Heuristics and biases. *Science, 185*, 1124-1131.

Tolcott, M.A., Marvin, F.F., & Lehner, P.E. (1989). Expert decisionmaking in evolving situations. *IEEE Transactions on Systems, Man, and Cybernetics, 19(3)*, 606-615.

Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Psychology, 12*, 129-140.

Wickens, C.D., & Hollands, J.G. *Engineering Psychology and Human Performance* (3rd edition). Upper Saddle River, NJ: Prentice-Hall, 2000.

Wolf, F.M. (1986). Meta-analysis: *Quantitative methods for research synthesis*. Thousand Oaks, CA: Sage Publications.