Mitigation of Cognitive Bias Through the Use of a Serious Game

Norah E. Dunbar[1], Claude H. Miller[1], Bradley J. Adame[1], Javier Elizondo[1], Scott N. Wilson[1], Stephanie G. Schartel[1], Brianna Lane[1], Abigail Allums Kauffman[2], Sara Straub[1], Judee K. Burgoon[3], Joseph Valicich[3], Elena Bessarabova[1], Matthew L. Jensen[1], Jeffrey Jenkins[3], Jun Zhang[4], and Dallas Morrison[1]

University of Oklahoma[1], University of Texas – Permian Basin[2], University of Arizona[3], University of Michigan[4]

Abstract: Intelligence analysts gather information from a variety of sources, process the information incrementally as it is received, and are under constant pressure for quick and accurate judgments. A serious training game called MACBETH was designed to address and mitigate cognitive biases undermining analysts' accurate collection and interpretation of intelligence. The IARPA SIRIUS program directed attention to two cognitive biases that are the focus of this experimental study—*fundamental attribution error,* and *confirmation bias.* In this experiment, 703 participants played the MACBETH game or engaged in a more traditional learning method—a video describing the same two cognitive biases. Results demonstrated the game to be more effective than the video when explicit training methods were combined with repetitive play.

Intelligence analysts gather information from a variety of sources, process the information incrementally as it is received, and are under constant pressure for quick and accurate judgments. In his book *The Psychology of Intelligence Analysis,* Heuer (2006) calls this process, "a recipe for inaccurate perception" (p. 27). To address the cognitive biases undermining accurate collection, interpretation and synthesis of intelligence by intelligence analysts, our interdisciplinary team of experts has designed, built and tested a serious training game called *MACBETH (**M**itigating **A**nalyst **C**ognitive **B**ias by **E**liminating **T**ask **H**euristics)* to improve the ability of future intelligence analysts to make decisions with greater systematic processing, and less reliance on cognitively biased heuristics. In this paper, we present the results of our first experiment testing the ability of MACBETH to mitigate two such biases: *the fundamental attribution error (FAE),* and *confirmation bias (CB).* Below, we will explain these two biases, describe the MACBETH game, and outline an experimental test designed to demonstrate the efficacy of MACBETH at mitigating these forms of flawed heuristic processing.

Confirmation bias is the tendency to search for and interpret information so as to confirm one's preconceived assumptions, expectations, or hypotheses (Nickerson, 1998). Heuer (2006) suggests at least 15 different strategies for "keeping an open mind," including identifying alternative models, thinking backwards, playing devil's advocate, using deferred judgment, and cross-fertilization of ideas. Although we could not implement all of these methods within the game, several of them were used in the MACBETH experiment. The FAE is a function of the tendency to over-emphasize stable, personality-based explanations (i.e., dispositions) for behaviors observed in others, while under-emphasizing the role and power of transitory, situational influences on the same behavior (Harvey, Town, & Yarkin, 1981). The MACBETH game was designed to make the differences between dispositional and situational cues salient, and reinforce the use of situational cues while discouraging the use of dispositional cues when making threat assessments about a source.

The details of the game mechanics and the development of the game are beyond the scope of this paper, however, they are described in detail elsewhere (Dunbar et al., 2013). In MACBETH, players assume the role of analysts who are presented with a series of fictional scenarios of impending terrorist attacks. The object is to figure out who the bad guys are, what weapons they will use, and where their attack will occur. In one turn, player-analysts have the opportunity to gather two pieces of information about the person, location, and/or weapon from various sources of intel, whereupon they can make a hypothesis, or aid another analyst with information that proves or disproves the other analyst's hypothesis. Some of the information analysts receive is vague, and therefore needs to be substantiated by first earning, then expending a chip allowing them to verify the relevant intel. Analysts earn chips by playing the mini-game "Archive," wherein they look at past case files and, based on situational and dispositional cues, decide whether a given suspect represents a threat or not. At the end of a turn, or if the players make or change

their current hypothesis, they are asked to verify their decision using information they've gathered in previous turns. Once a player-analyst has gathered enough information, s/he can submit a final hypothesis and potentially win the round. Throughout the game, analysts learn about the cognitive biases and receive both implicit and explicit feedback encouraging them to seek disconfirming information, disprove their hypotheses, rely more on situational cues, and be made aware of their susceptibility to biased decision making processes and errors. To test the efficacy of the game, we compared it to a control condition; a training video informing viewers about the nature of cognitive biases using entertaining vignettes presented by a professorial host in a lab coat. The experiment not only manipulated different versions of the MACBETH game to determine the most effective treatment conditions, it also tested the interactive internalization of information via an immersive game against the more traditional, passive learning afforded by an instructional video.

Of course, games can vary in their effectiveness depending on the nature of game mechanics introduced, and the conditions within which they function during game play. And although there is some debate as to whether explicit training is as effective as more implicit approaches allowing players to figure out on their own the best way to proceed, the literature suggests priming information (i.e., providing explicit material about biases) can generally be effective at ameliorating bias. Accordingly, the experimental manipulations included either did or did not intersperse instructions, making hasty generalizations and over-reliance on dispositional attributions salient to players throughout the game. This formed the reasoning for the following hypothesis concerning the use of explicit priming to reduce CB and FAE:

H1: MACBETH with priming increases pre- to post-game bias familiarity and knowledge and reduces biased judgments relative to MACBETH without priming.

The research literature also suggests longer playing time and repeated play opportunities should enhance a game's ability to improve knowledge—and in the present case, reduce biased judgments. To test these assumptions, we manipulated the amount of game play time available to players. They were assigned either to a short (30 minute) or long (60 minute) initial game session, and to either a single-play, repeated play in the laboratory, or a repeat play take-home condition.

H2: Relative to shorter game play sessions, longer sessions increase pre- to post- bias familiarity and knowledge and reduces biased judgments.

H3: Relative to single-shot game play, repeated play (whether in-laboratory or take-home) increases bias familiarity and knowledge and reduces biased judgments across time periods.

Method

Participants

Participants included 703 students recruited by mass emails and classroom announcements at two Universities in the Southwestern and South-central United States. The sample included 329 females (47%) ranging between 18 and 62 years of age (M = 22.03, SD = 5.34).

Procedure

Upon arriving at the lab, an experimenter randomly assigned participants to one of the 10 experimental conditions in blocks. Participants were administered pretest measures, followed by the experimental treatment (either one of the MACBETH game conditions or the control video), and then post-treatment measures. At the end of their first session, participants assigned to the repeat play condition scheduled a follow-up appointment to be completed in the lab within one week. Those assigned to the take-home condition were given login instructions for the game administered online. All participants were paid $20 for their participation in each lab session and were reminded they would be emailed a link to a follow-up survey in 8 weeks. All participants were also paid $30 upon completion of the 8-week survey.

Conditions

Priming

Participants played a version of MACBETH explicitly addressing the biases (i.e., the priming condition), or a version not explicitly addressing this information (i.e., the non-priming condition). The priming version

contained pop-up quizzes at various points in the game at which time players were given text defining the bias followed by a multiple choice question checking whether they'd learned the definition correctly. If they answer the question incorrectly, they were given the definition again, followed by a repeat of the question, and were allowed to advance only after correctly answering the question. The FAE quiz appeared the first time players entered the archive mini-game, whereas the CB quiz appeared the first time they entered the hypothesis testing portion of the game.

Duration

Players were randomly assigned to 30- or 60-minute versions of the game; and in both cases saw a clock counting down the time remaining in their game.

Repetition

Players were randomly assigned to either the in-laboratory single-play, the in-laboratory repeated play, or take-home repeated play conditions. At the end of their first visit, those in the in-lab repeat-play condition were asked to sign up for a second lab visit one week later (those in the repeat-play condition who did not return for their second play session, i.e., 21%, were considered to be single-play participants). Those assigned to the take-home condition played one session in the lab and then were given a sheet of login instructions for the game which directed them to play the game at home at least twice—but they had the option of playing as often as they liked within a two week period. Of the 102 players assigned to the take-home condition, 50 logged in at least one time from home and played an average of roughly 53 minutes per session. Take-home players who did not play at home were also considered to be single-play participants.

Measurement

We used three different measures for knowledge about the biases. First, participants rated their own perceived familiarity with the biases on a 7-point scale. Second, they completed a 6-item test matching definitions with biases. Third, they responded to three multiple choice exam-style questions in a bias application test presenting short scenarios for which participants identified the bias being illustrated.

For CB bias mitigation, six new items were developed to measure CB (modeled after Rassin, 2010) which included scenarios such as: "You find out someone borrowed your laptop last night without your permission and let the battery run all the way down. You think it's one of your brother's friends because he's always forgetting his own laptop, and he was probably around last night. If you wanted to test your suspicion, what question(s) would you ask others about this person?" Each item offered a similarly brief scenario followed by four response options, two of which indicate confirming responses (coded -1), and two of which indicate disconfirming responses (coded +1), so possible responses ranged between -2 and +2, with lower scores indicating higher levels of confirmation bias. Of the six items, three were included at the pretest and three at Posttest 1, with the three pretest items used again at Posttest 2, and the three Posttest 1 items used again at the 8-week follow-up posttest. The three items used at each test period were summed to create one scale which we called "NewCB" ranging from -6 to +6. Three Confirmation Bias Application Measures (CBAM) were developed based Watson's (1968) card flip paradigm, and used as a second CB mitigation measure.

For FAE bias mitigation, two different measures were used; one consisted of four vignettes adapted from Riggio and Garcia's (2009) "Ron's Bad Day" scenario. Each vignette presented a short scenario in which the central character experienced either positive or negative consequences due to their choices or the circumstances. The scenarios were sufficiently vague as to allow participants to build their own attributions about the causes of consequences within each narrative. After reading a scenario, participants were presented with 10 situational (e.g. broken down car, the weather, work environment) and dispositional items (e.g. personality, attitude, skills and abilities), and asked to indicate the degree to which each played a role in the consequential outcomes. The second FAE measure, adapted from Stalder (2000), required participants to watch a short video in which two participants play a trivia game with one randomly chosen to make up questions for the other. Participants were asked to evaluate both characters' knowledge, memory, and ability. Because the situation could make it appear as though the questioner is more knowledgeable than the contestant—even though each was randomly assigned their

roles—the degree to which participants rate the questioner as superior to the contestant indicates FAE through the discounting of situational cues.

Results

Analysis Overview

For all analyses reported below, to capture results from participants' posttests following the last time they played MACBETH prior to the 8-week posttest, we conducted repeated measures MANCOVA using a "last post" variable that recoded Pretest, Post 1, Post 2 and 8-week Post into three levels (pre, last post, and 8-weeks). The reason for this is a function of the way SPSS deals with missing data (i.e., using "listwise deletion" for repeated measures analyses, which removes cases that do not include data for all time periods (pre-test, last post-play test and 8-week follow up posttest). Creating this "last post" variable allows for more accurate detection of differences between repeat players and non-repeat players by including the repetition variable as a factor. The analysis used a 2 (duration; 30/60 minute) x 2 (training type; priming/no-priming) x 2 (location; University 1 or 2), x 2 (Sex of Subject; M/F) design with age, extroversion, openness, agreeableness, conscientiousness, emotional stability, need for closure, horizontal individualism, vertical individualism, horizontal collectivism, confirmation proneness, personal need for structure/personal fear of invalidity, sensation seeking, handedness, logic aptitude, computer comfort, gaming experience entered as covariates. All non-significant covariates were dropped and analyses rerun using the reduced models.

Recognition of Cognitive Biases

Familiarity

Participants rated on a 7-point scale their degree of familiarity with each of the biases, from very familiar to very unfamiliar. Every cell had the means and SDs calculated for CB and FAE. Prior to game play or control video treatments, participants reported low levels of familiarity with each bias (paired-samples t-tests assessing pre- to posttest knowledge were significant at p <.001 for all conditions).

Somewhat unexpectedly, there was a slight increase in the players' familiarity rating from their last posttest to the 8-week follow-up. When looking at each bias and cell individually, there are higher mean familiarity ratings at the 8-week posttest than at last posttest. All 8-week means were higher than the Posttest 1 means, and all the differences from the pretest to 8-week were significant: CB: t(560)= 5.53, p<.001; FAE: t(559)= -8.74, p<.001, Posttest 1 to 8-week, CB: t(559)= 5.51, p<.001; FAE: t (559)= 4.43, p<.001, and Posttest 2 to 8-week, CB: t(247)= 3.82, p<.001; FAE: t(247)= 4.73, p<.001. Participants who played the explicit training (primed) game reported higher levels of familiarity at the 8-week posttest (CB: t(515)= 6.378, p<.001; FAE: t(474)= 4.58, p<.001). However, those who watched the video reported significantly higher levels of bias familiarity (CB: t(559)= -2.51, p=.001; FAE: t(559)= -3.03, p=.002) relative to most game conditions. There were however no significant differences between the best game condition (60 minute, primed, repeat play) and the control video.

Bias Matching Test

The second measure of bias recognition directed participants to match the definitions of the six biases with the appropriate bias names in a "drag and drop" exercise. This measure was scored such that participants earned one point for each correct match, with scores ranging from 0 to 6. All differences from Posttest 1 to the 8-week posttest were significant t(714)= 11.89, p<.001 as were the differences from Last posttest to the 8-week posttest t(714)= -9.09, p<.001. There were no significant differences from the participant's last in-lab posttest to the 8-week posttest for duration, training type or number of times participants played MACBETH. However, for the matching test, the control video outperformed all game conditions t(58)=2.40, p=.02, including the best version of the game, t(73)= -2.31, p=.02.

Bias Application Knowledge Test

To test the same repeated measures ANCOVA across 3 time periods (pre, last post, and 8-weeks), duration, repetition, training type, sex of subject, and location were included as IVs, along with the above covariates. Only extroversion and logic aptitude were included as covariates in the reduced model. Findings revealed a significant between-subjects main effect for training type F (1, 527) = 8.74, p < .003,

η^2 = .008, and a significant quadratic within-subjects time period by training type interaction, *F* (1, 527) = 6.66, *p* = .01, η^2 = .01. The priming group (*M* = 1.35, SE = .04) and the video control (*M* = 1.35, SE = .04) were not significantly different from one another (*p* = .08), however, they were both significantly different from the no-priming group (*M* = 1.35, SE = .04), with both being significantly higher than the no-priming on CB knowledge. The time period by training type interaction demonstrates again that, although the video control group had greater bias knowledge at the immediate posttest, they also had greater knowledge loss at 8-weeks. Although comparisons between the video controls and the priming groups show no significant differences at 8-week post, *t* (327) = -1.29, *p* =. 20, the video control group was significantly better then the no-priming group, *t* (279) = -2.22, *p* =. 03.

For FAE, no covariates were retained within the reduced ANCOVA model. Results indicated a significant quadratic within-subjects main effect for duration, *F* (1. 527) = 13.12, *p* < .001, η^2 = .02, a between-subjects main effect for training type, *F* (1. 527) = 7.79, *p* =.005, η^2 = .02, and for location, *F* (1. 527) = 6.50, *p* = .01, η^2 = .01, along with a significant quadratic within-subjects time period by duration by sex interaction effect for time X duration X sex of subject, *F* (1. 527) = 8.83, *p* = .003, η^2 = .02. The main effect for time period suggests FAE knowledge increased from pre-test (*M* = 1.22, *SE* = .04) to the last posttest (*M* = 1.40, *SE* = .05), and then decreased again at the 8-week posttest (*M* = 1.31, *SE* = .04), however, it nevertheless remained marginally higher (p = .06) after 8-weeks than at pre-test. The training type main effect suggests priming (*M* = 1.34, *SE* = .04) led to greater FAE knowledge than no-priming (*M* = 1.19, *SE* = .05) and the control video led to more knowledge (*M* = 1.69, *SE* = .11) than both priming and no-priming conditions (all *p* <.05). The location main effect indicated University 1 participants scored better overall on the FAE Knowledge test (*M* = 1.42, *SE* = .05) than University 2 (*M* = 1.21, *SE* = .04). Examination of the simple effects suggests both males and females performed better in the video than the game conditions, however, controls had greater knowledge loss at 8-weeks compared to all game conditions. For males, the 30-minute game condition had greater knowledge loss than the 60-minute condition at the 8 week post, t (258) = -2.65, *p* =. 001, but for females this difference was not significant, t (259) = .81, *p* =. 42.

Mitigation of Confirmation Bias

To test CB mitigation, repeated measures ANCOVAs were conducted, using both the CBAM and the NewCB scores as the DVs. None of the covariates were retained for the CBAM analysis, and there were no significant results for any of the between-subjects factors, or for test period. The NewCB analysis included agreeableness (*p* = .026) and need for closure (*p* = .024) as significant covariates, and yielded a significant test period by training type (priming vs. no priming), and repetition (one-shot, repeat-play, take-home, video) as between-subjects factors. There was a significant interaction, *F*(2, 804) = 3.32, *p* = .036, η_p^2 = .01, with participants in the priming condition demonstrating a greater reduction in CB than no-priming participants (see Figure 1; those in the video condition are also shown for reference). This effect appears to be fairly robust as little decline between the last posttest and the 8-week posttest is evident.

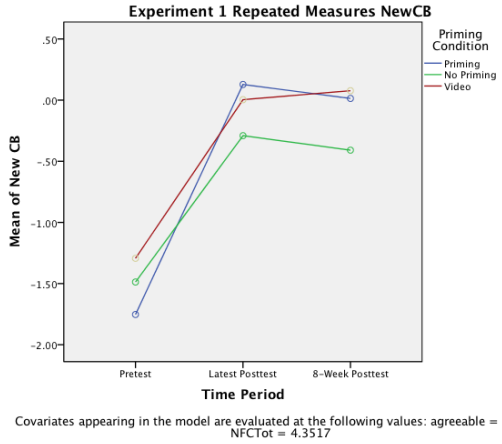

Figure 1: Repeated Measures NewCB Means by Condition

Mitigation of Fundamental Attribution Error (FAE)

To analyze the FAE data, mean scores were calculated for dispositional and situational cues, and a separate repeated measures analysis was run for each group of cues. Scores for dispositional situational cues were coded so higher scores indicated increased reliance on each cue. ANCOVA tests used the same IVs as above, for which no significant covariates were retained, revealing a significant main effect for time period, Wilks' λ = .985, $F(2, 574) = 4.42$, $p = .012$, $\eta^2 = .015$, indicating dispositional scores for all participants changed over the three time periods.

Between subjects effects revealed a significant main effect for time $F(2, 1150) = 4.06$, $p = .018$, $\eta^2 = .007$, such that participants in all conditions reduced their reliance on dispositional cues across the three time periods, pre, $M = 5.12$ ($SD = 1.60$), last post, $M = 4.71$ ($SD = 1.77$), and 8-week post, $M = 4.65$ ($SD = 1.62$). There was a significant time by duration by repetition interaction, Wilks' λ = .989, $F(2, 574) = 3.08$, $p = .074$, $\eta^2 = .011$, with between-subjects effects indicating 30-minute duration participants in the video condition reporting slightly less reliance on dispositional cues across time periods, $F(2, 1150) = 3.70$, $p = .025$, $\eta^2 = .006$ (Figure 2). Despite lower scores for each time period, repeat game players showed a trend of decreasing reliance on dispositional cues from latest post to 8-week post, whereas those in both the one-shot and control conditions showed a slight increase in reliance on dispositional cues. Participants in the 60-minute play conditions reported decreasing reliance on dispositional cues across the three time periods, with those in the one-shot condition reporting the least reliance on dispositional cues at the 8-week test time.

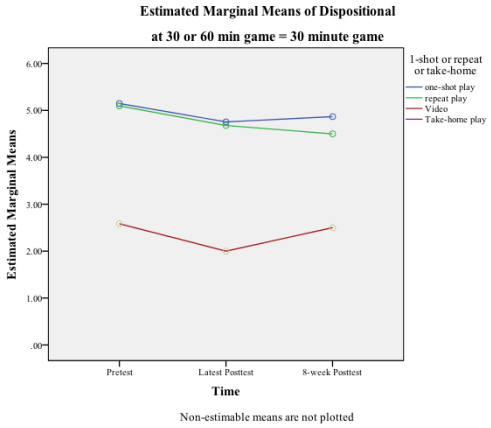

Figure 2: Repeated Measures Dispositional Cue Reliance Means by Condition

MANOVA analyses for preference of situational cue scores at pretest, last post, and 8-week post found no significant effects for any of the IVs or covariates. To examine how participants reacted to the questioner and contestant featured in the Stalder video task (Stalder, 2000), a repeated measures ANCOVA was conducted using the above IVs along with sex included as a fixed factor. A significant multivariate effect for time period, Wilks' λ = .972, $F(1, 544) = 15.45$, $p < .001$, $\eta^2 = .028$, indicated participants in all conditions reduced their bias across time periods. No significant effects were found for duration, Wilks' λ = .999, $F(1, 544) = .519$, $p = .472$, training type, Wilks' λ = 1.00, $F(1, 544) = .071$, $p = .79$, or repetition, Wilks' λ = 1.00, $F(2, 544) = .09$, $p = .914$, although the main effect for sex was marginal, Wilks' λ = .994, $F(1, 544) = 3.15$, $p = .076$, $\eta^2 = .006$, and the between-subjects effect for time period indicated participants in all conditions reduced their bias from last post, $M = 2.12$ ($SD = 1.38$) to 8-week post, $M = 1.22$ ($SD = 1.11$), $F(1, 544) = 130.51$, $p < .000$, $\eta^2 = .24$.

Discussion

Hypothesis 1 predicted, relative to no-priming, MACBETH with priming should increase pre- to posttest reduction in biased judgments, and this expectation was supported in part: There was a significant main effect for priming on the NewCB scale responses, indicating the primed game conditions produced greater CB mitigation relative to the non-primed conditions, and this effect was qualified by a marginally

significant 3-way interaction between priming, duration, and location which suggests participants at one location appear to have responded more favorably to priming in the 30-minute game, whereas those at other location appear to have responded more favorably to priming in the 60-minute game. Priming also produced greater FAE bias reduction along with a significant interaction between priming and time period. Both versions of the game successfully reduced the use of dispositional cues from pretest to posttest 1, however, the no-priming and control groups showed a reversal at the last posttest, with a significantly greater use of dispositional cues relative to the priming group, for which reduced use of dispositional cues persisted after 8 weeks.

Analyses of game duration on bias familiarity, knowledge, and judgment provides partial support for H2. Although little effect was found for duration on biased judgments in pre- vs. posttest comparisons, and there was no significant effect for duration on FAE judgments, a significant main effect was found in the NewCB judgment test, which was qualified by a duration by location interaction indicating 30-minute players performed better at one location, whereas 60-minute players performed better at the other. While longer duration was found to enhance learning and bias mitigation in some cases, most of the knowledge- and judgment-related tests revealed only minor improvements, with some showing no effect. However, several results indicate longer duration of play was very effective when combined with priming and repeat play, particularly in the optimal (60 minute, primed, repeat) experimental group. When compared to single-play, repeate play offers several clear benefits: As support for H3 suggests, repeat players performed better on judgment tests than players in the single play condition, and as the mitigation results using the NewCB scale show, repeat-play clearly outperformed single-play. Moreover, with regard the FAE measures, there was a significant repetition by time period interaction indicating increased use of situational cues by those in the repeat-play conditions across time periods.

Conclusions

Although one might expect some decay in the effects of MACBETH on bias mitigation measured 8 weeks after game play, the positive de-biasing results appear largely robust to the passage of time. Across bias mitigation tests within the optimal game conditions, our analyses show only a very slight nonsignificant drop in the two primary bias mitigation measures, implying the knowledge gained by playing MACBETH appears to be internalized and retained, It seems clear that repeated play and longer duration both work to increase the positive effects of MACBETH game play. Moreover, Relative to implicit training alone—as provided by the no-priming conditions—the priming conditions were clearly more effective, suggesting explicit training with definitions and quizzes offers the most optimal method of training within the game. Although the explicit training provided by the video appears to be more effective at teaching knowledge about the biases, knowledge alone does not translate into actual reduction in bias use.

References

Dunbar et al. (2013, June). *The Development of a Serious Game for the Mitigation of Cognitive Bias: The Case Study of MACBETH.* Paper to be presented at the ICA Convention, London, UK.

Harvey, J. H., Town, J. P., & Yarkin, K. L. (1981). How fundamental is the fundamental attribution error? *Journal of Personality and Social Psychology, 40*(2), 346.

Heuer, R. J. (2006). *The Psychology of Intelligence Analysis*. New York: Novinka.

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

Rassin, E. (2010). Blindness to alternative scenarios in evidence evaluation. *Journal of Investigative Psychology and Offender Profiling, 7*(2), 153-163. doi: 10.1002/jip.116

Riggio, H. R., & Garcia, A. L. (2009). The Power of Situations: Jonestown and the Fundamental Attribution Error. *Teaching of Psychology, 36*(2), 108-112. doi: 10.1080/00986280902739636

Stalder, D. R. (2000). Does logic moderate the fundamental attribution error? *Psychological Reports, 86*(3), 879-882. doi: 10.2466/pr0.2000.86.3.879

Acknowledgements: This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL). The views and conclusions contained 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, AFRL, or the U.S. Government.