Serious Game User Data Analysis and Visualization: Savoring the Breadcrumbs

Brandt Dargue	Dov Jacobson	John Sanders
Boeing Research & Technologies	GamesThatWork	Historical Online Learning Foundation
St. Louis, MO	Atlanta, GA	Louisville, KY
Brandt.W.Dargue@Boeing.com	Dov@GamesThatWork.com	SandersJ@HOLF.org

ABSTRACT

Following the evidence of research in critical thinking and cognitive bias training, we developed training designed to demonstrate people's decisions and actions affected by cognitive bias or elicit a bias in the player. Ideas were collected from and refined by experts with diverse backgrounds distributed geographically using an innovative solution. We then developed a video game that used multiple scenarios to teach cognitive bias recognition and mitigation. A learning and training effectiveness study indicated that the game was effective for learning although results were inconsistent for the different cognitive biases addressed by the game. The game recorded detailed data about scenes, scenarios and decisions each player made in the game. Called "breadcrumbs", this data detailed the path every player took. Traditional statistical analytic techniques tend to be clumsy instruments for breadcrumb analysis. Additionally, early aggregation and dimension reduction make the data more tractable but less meaningful.

This paper details specific examples of how the breadcrumbs – paired with the study data – provided valuable answers in pinpointing areas to improve the game's learning effectiveness. The paper provides enough background information on the subject to enable the audience to appreciate the difficulty in cognitive bias training effectiveness and understand the examples shown. The majority of the paper discusses the data, the analysis, and the innovative data visualization techniques used. We discuss approaches that may prove more appropriate to extracting useful information from breadcrumb trails than traditional statistical analytic techniques. The audience will gain an understanding of the value of testing, data collection, and data visualization in training, education, simulations, and serious games. The paper will conclude with a discussion on using the techniques to improve the small batch testing in serious game development.

ABOUT THE AUTHORS

Brandt Dargue is an Associate Technical Fellow performing research into current and future training technologies including simulations, automated performance assessment, adaptive scenarios, intelligent tutoring, virtual environments, mobile platforms, gaming concepts and gaming technologies. Employed at Boeing for 25 years, he has chaired or participated in several international standards development and study groups, and was the Program Manager and Principal Investigator for *The Enemy of Reason* serious game and training effectiveness studies.

Dov Jacobson leads a studio of talented, happy and committed workers at GamesThatWork. Dov has been the Principal Investigator in eight successful federally sponsored game research projects. Since 1981, he has led the development of 44 games, including award winning titles in both entertainment and learning genres. He focuses GamesThatWork's formidable creative resources on the problems faced by the studio's sponsors. Dov is a frequent presenter at game and learning conferences and is known for his merciless sense of humor.

John Sanders, LTC (ret) has broad experience in training and simulation for Army and Joint Service applications. At Boeing, his duties included Lead Systems Integrator for collective training of the Future Combat Systems (FCS) core program and Spin-Out 1 systems. Mr. Sanders participated in projects for serious games, adaptive training, mobile apps to train and sustain proficiency of operators in Afghanistan using the Army's Joint Recovery and Distribution System (JRaDS), and adaptation of an automated progressive training program for the Brigade Combat Team Modernization project. As a government employee, he worked in program management and operational requirements for a variety of simulators, simulations, part-task trainers, and worked on various training technology research projects including Project Scimitar (SASC project), Force XXI, and the Staff Group Trainer.

Serious Game User Data Analysis and Visualization: Savoring the Breadcrumbs

Brandt Dargue	Dov Jacobson	John Sanders
Boeing Research & Technology	GamesThatWork	Historical Online Learning Foundation
St. Louis, MO	Atlanta, GA	Louisville, KY
Brandt.W.Dargue@Boeing.com	Dov@GamesThatWork.com	SandersJ@HOLF.org

INTRODUCTION

The story of two very clever children named Hansel and Gretel has been immortalized because of their attempt to avoid getting lost in the woods by leaving a trail of crumbs from a loaf of bread. This famous technique has been improved upon and is still helpful centuries later to avoid getting lost in big data. In this paper, we describe the use of digital "breadcrumbs" to help developers discover where players in a serious video game "went astray" and discover correlations with learning effectiveness measures to pinpoint areas for improvement. This paper tells the true story of analyzing breadcrumb data recorded about every decision/action each player made in the game combined with records about every decision the player made in the test. This story has a happy ending like the old tale, as the analysis enabled the game to be improved and be more effective at transferring knowledge, skills, and attitudes (KSA), so that the players learn to be less prone to making judgment errors that are caused by cognitive biases.

Cognitive Biases

Several decades of research suggest that the human brain manages complex judgment and uncertain information by applying heuristics or cognitive "short cuts" (BR&T, 2013; Heuer, 1999). While this is necessary for our survival, it can lead to improperly influenced decisions, particularly when the information is incomplete or uncertain (e.g., (Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980). To test a theory that a serious game might be more effective than traditional training in cognitive bias recognition and mitigation, the Intelligence Advanced Research Projects Activity (IARPA) and the Air Force Research Laboratory (AFRL) sponsored the Sirius Project under which multiple teams developed various interactive computer games and conducted training effectiveness studies of those games as compared to a training video. Each team designed and developed alternative versions that tested hypotheses each team proposed to increase the training effectiveness of their game. The cognitive biases addressed by Phase 1 of the project as defined in the Broad Agency Announcement (BAA) (IARPA, 2011, p. 7) were:

- **Confirmation Bias**: The tendency to [preferentially] search for or interpret information in a way that confirms one's preconceptions. Often preceded by priming.
- **Fundamental Attribution Error (FAE)**: The tendency for people to over-emphasize personality-based explanations for behaviors observed in others while underemphasizing the role and power of situational influences on the same behavior (also called attribution bias).
- **Bias Blind Spot (BBS)**: The tendency for an individual to be unaware of their own cognitive biases, even when the individual can recognize [these] cognitive biases in others.

Cognitive biases are a result of the heuristic being processed unconsciously which is described as "System 1" thinking. (Dual process theory, 2014) The primary method of mitigating cognitive biases is to trigger "System 2" thinking. "*Heuristic judgments, which lead to biases, are associated with System 1, and analytic reasoning, which may intervene with these judgments and improve them, are linked to System 2*" (Evans, 2008). The mitigation technique for each of the cognitive biases in this program are:

- **Confirmation Bias**
  - Consciously seek and possibly emphasize evidence that disconfirms your theory or initial evidence
  - Use the Analysis of Competing Hypothesis (ACH) Structured Analytic Technique (SAT)
- **Fundamental Attribution Error**:
  - Examine situational influences
  - Put yourself in the other person's situation
- **Bias Blind Spot**:
  - Remind yourself that cognitive biases occur unconsciously – even <u>your</u> mind is suspect.

The Enemy of Reason

The National Research Council Committee on Learning Research and Educational Practice's report on "How People Learn" states that experts are not just "smart" or knowledgeable on factual information, but are able "to generate responses, with minimal cues, repeatedly over time with varied applications so that recall becomes fluent and is more likely to occur across different contexts and content domains." (Donovan, Bransford, & Pellegrino, 1999). The best way to develop this ability is by requiring students to frequently retrieve the knowledge in a variety of contexts and settings (Halpern & Hakel, 2003). Typical instruction of critical thinking skills that mitigate cognitive biases can be generalized as awareness of the biases followed by awareness of the situations where they typically appear. It is therefore important to have lots of examples and lots of practice. The examples are typically case studies and the practice that is most effective are role-playing exercises. In addition, we included an in-game SAT based on ACH that was adapted to help mitigate all three of the cognitive biases (Dargue, Jacobson, & Sanders, 2014).

The Enemy of Reason leverages video game technologies and game concepts to provide knowledge about each bias, presents a variety of examples of biased people, and provides practice exercises with instructional feedback. The exercises immerse the player in role-playing scenarios that are based on actual case studies or cognitive bias research that have been woven into *The Enemy of Reason* storyline. At first, the player has to recognize and mitigate the bias of non-player characters in the game. In advanced levels, the player must uncover and mitigate his or her own cognitive biases in interpreting information and making decisions. Because it is important to provide as many examples as possible in different contexts and settings as suggested by Halpern & Hakel (2003), the game includes 188 Cognitive Bias Interactions (CBIs) and is designed to be played in approximately five hours.

The game puts the player in the role of Agent Ian Solitaire and as other members of Ian's team trying to solve the mystery of a weapon that has dispersed a red cloud on the city. The player holds conversations with non-player characters to uncover hypotheses and gather evidence, until he or she commits to a course of action (BR&T, 2013, p. 8). The red cloud contains bias viruses that appear to make some of the cognitive biases in the population more pronounced and harder to mitigate. This plot device embedded in a classic espionage game filled with ambiguity and uncertainty provides players with many opportunities to explore cognitive bias in others and oneself. (Dargue, Jacobson, & Sanders, 2014). A considerable part of the game development involved authoring these opportunities by writing the situations and dialogs for the interactions with the characters.

EFFECTIVENESS STUDY

Research Design

We used a pre-test/post-test control group design. Post-testing included both an immediate post-test, upon completion of the training, and a retention post-test, 8 weeks later. The test instruments were on-line written tests that had both a knowledge test and a behavior test. One was designed to measure the participant's knowledge about the cognitive biases. The second section of the test was designed to measure the participant's degree of cognitive bias. An effective game would then show an improvement in the "knowledge of biases" and a reduction in "cognitive bias" from pre-test to post-test.

The Enemy of Reason study involved the manipulation of two independent variables: the use (or withholding) of a SAT; and the manipulation of training duration/repetition. These two independent variables were used to form 3 experimental groups, to which a control group is added. The resulting 4 training treatment groups in the study are as follows (BR&T, 2013):

1. Six Game Segments with SAT (SAT-6). This group receives six 30-minute segments of game-based training with access to the SAT. The 6 segments used in SAT-6 involved combinations of abbreviated forms of the 10 segments used in SAT-10.
2. Ten Game Segments with SAT (SAT-10). This group receives ten segments of game-based training with access to the SAT. Each level lasted approximately 30 minutes.
3. Ten Game Segments without SAT (No SAT). This group receives ten 30-minute segments of game-based training, but without access to the SAT.
4. Control. Training for the Control group is limited to one viewing of a government-provided, 30-minute instructional video on cognitive biases.

We selected a sample size of 30 participants per treatment group. Participants were drawn from the Mercyhurst University Institute for Intelligence Studies. The protocol required that they be at least 18 years of age at the time of recruitment. We used a stratified random sampling technique to ensure a gender ratio within the treatment groups of no more than 65:35 for either sex. Table 1 shows the sample sizes at pre-test, immediate post-test, and retention post-test.

Table 1. Sample Sizes at Pre-Test, Immediate Post-Test, and Retention Post-Test. From Perrin et al, (BR&T, 2013)

	Pre-Test	Immediate Post-Test	Retention Post-Test
Control	30	30	29
SAT-6	34	34	34
SAT-10	34	33	32
No SAT	33	31	28
Total	**131**	**128**	**123**

Study Results

The program had five metrics defined by IARPA: Improved knowledge of biases; immediate reduction in cognitive biases; persistent reduction in cognitive biases; game that trains more effectively than an instructional video; and a Game that is Engaging. IARPA provided the instructional video on DVD as a hi-definition video file for playback on computer screens. We developed a knowledge test and a bias behavior test instrument to measure the knowledge and degree of bias for each participant. This test was provided as a pre-test before playing the game, a post-test immediately after the last session playing the game, and a retention follow-up test given eight weeks after the last game play session. Eye-tracking equipment was used to measure engagement. Negative percentages below indicate where the participant's post-test score was lower than the pre-test score. This was only observed in the control condition.

- **Improved Knowledge of Biases**:
  - All of our training treatments produced significant increases in the knowledge test scores at both the immediate and retention post-test.
    - Immediate: Trial: F(1,119) = 502.3, p < .001.
    - Retention: Trial: F(1,119) = 219.34, p < .001
- **Immediate Reduction in Cognitive Biases**:
  - One or more game conditions exceeded the instructional video control. Averaged across all biases, SAT-6 was the best condition with 10.3% reduction in the immediate test and No SAT was the best condition for the retention test with 1.8% reduction. When looking at the individual biases for both immediate and retention, SAT-6 was best at FAE mitigation training, and No SAT was best for BBS mitigation training.
    - SAT-6: 24.6% reduction in FAE (Control: -8.7%)
    - No SAT: 28.6% reduction in BBS (Control: -1.0%)
    - No groups showed a reduction in CB (Control: 0.8%)
- **Persistent Reduction in Cognitive Biases:**
  - Reduction across biases (pre to retention post-test)
    - No SAT (best condition): 1.8%; however, best condition depended on the bias measured
    - Bias reduction for Control: -7.5%
  - Reduction in individual biases
    - SAT-6: 6.8% reduction in FAE (Control: -11.2%)
    - No SAT: 23.1% reduction in BBS (Control: 2.0%)
    - No groups showed a reduction in CB (Control: -19.4%)
- **Game that Trains More Effectively than an Instructional Video (Control)**
  - Effectiveness (pre to immediate posttest)

        - SAT-6 > Control: FAE; F(3, 119) = 2.71, p < .048) for Treatment by Trial interaction
        - No SAT > Control: BBS; F(3, 119) = 3.20, p < .026) for Treatment by Trial interaction
    - Effectiveness (pre to retention post-test)
        - No significant differences between Control and any version of the game
- **Game that is engaging**:
    - The mean level of engagement across all participants was 80.2%, compared to the Program objective of 75%. The minimum level of engagement for any participant was 53.1%, compared to a Program objective of 50%

Table 2 provides additional detail on the changes in average bias knowledge score, showing the pre-test, immediate post-test, and retention post-test means and standard deviations by training treatment group. Figure 1 shows these results as a graph. As shown in Figure 2 and discussed above, there was some significance differences in the results of knowledge improvement between the biases and for FAE, between study conditions. These graphs also clearly show that the game was better at persistent knowledge improvement than the control video.

Table 2. Immediate and Delayed Bias Knowledge Results. From Perrin et al, (BR&T, 2013)

	Pre-Test			Post-Test				Delayed (8-week) Post-Test			
Knowledge Measure	N	Mean	SD	N	Mean	SD	Pre-Post Difference	N	Mean	SD	Pre-Post Difference
SAT-6	34	35.3	21.5	34	80.4	18.0	*t(33) = 12.85, p < .001*	34	76.5	22.4	*t(33) = 9.72, p < .001*
SAT-10	32	33.3	20.6	32	76.0	20.4	*t(31) = 9.88, p < .001*	32	68.4	26.3	*t(31) = 6.67, p < .001*
No SAT	28	35.7	19.7	28	79.0	21.7	*t(27) = 8.96, p < .001*	28	76.6	23.7	*t(27) = 8.43, p < .001*
Control	29	30.3	20.6	29	83.1	18.7	*t(28) = 14.19, p < .001*	29	59.8	25.4	*t(28) = 5.45, p < .001*

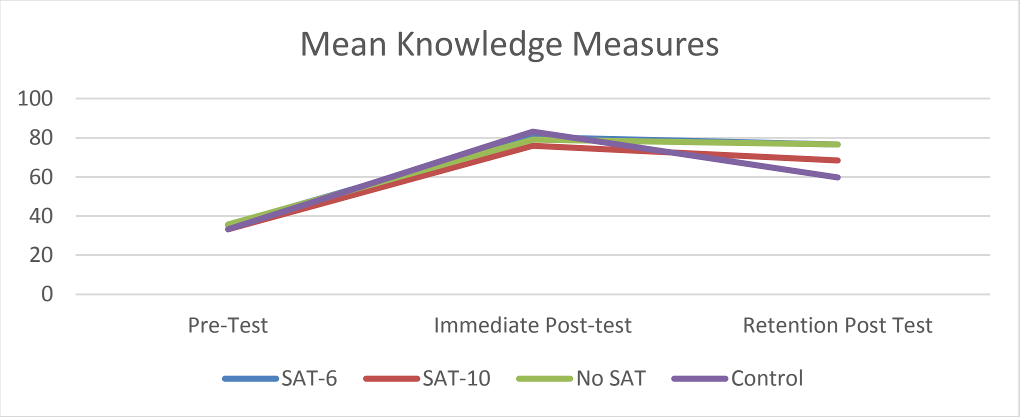

Figure 1. Immediate and Delayed Bias Knowledge Results

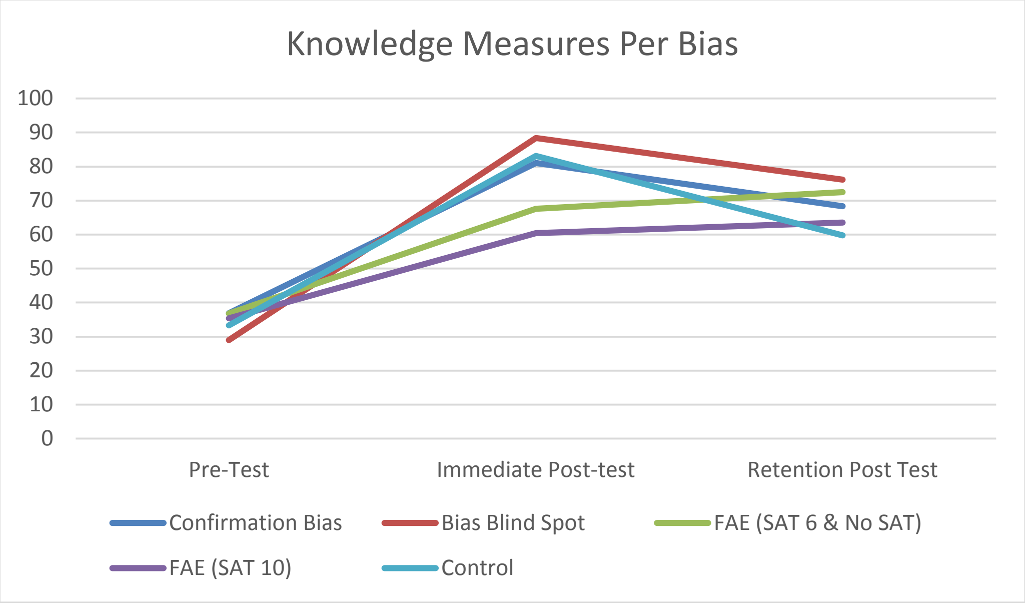

Figure 2. Immediate and Delayed Bias Knowledge Results Per Bias/Condition

DATA ANALYSIS

Traditional multivariate statistical analysis, such as ANOVA, is an extremely powerful tool for identifying the correlations between features. Given a sufficient number of samples, an analyst can reliably match outcomes with putative inputs. The analyst who detects such correlations between, for example, the achievement of learning objectives and a particular training methodology will be tempted to attribute this learning success to that particular methodology. In a perfect experiment, the inference will be true.

In our case we could control for most variables, while we examined two independent conditions. One was the length of play and the other was the presence of a Structured Analytic Tool (a job aid for cognitive bias mitigation). Because of the carefully controlled experiment, correlations between these variables and the learning outcomes can be reliably classified as causal. Although the experiments were designed to measure the instructional value of these two features, the team relied on the data from the study to create a better game for later trials. This means evaluation of features other than the two independent variables.

The Enemy of Reason presents the player with nearly two hundred individual plot points in which the player is called upon to identify and mitigate cognitive bias. The bias is manifest either in a game character or, in the advanced levels, in the player herself. Since total time of play was one of the two independent variables, the team developed a shorter edition of the game in which the cognitive bias encounters were reduced by thirty-six percent. This reduction offered the opportunity to eliminate the least valuable CBIs and ensure we retain the most potent interactions.

The identification of the value of each CBI began with statistical analysis of the breadcrumb record. Each of these CBIs offers both a learning experience and a decision point that exercises the recently learned principle. The analysis sought to measure the gameplay data to find correlations between these intended learning experiences and final outcomes. The figure of merit was contribution to the program’s goals – improved knowledge and skills. "Improvement" was defined as the reduction in error between the pre-treatment test and the post-treatment test. Using anonymously assigned identifiers, these improvement scores were calculated for two dimensions of learning: knowledge acquisition and behavior change.

Breadcrumb data is a record of the path a player takes through the game. In web page design, breadcrumbs are exposed to a site's visitors to offer both an orienting device and a set of incremental steps back to the website's home page. These website breadcrumbs do not really record the user's path from the home page to the current page. They simply show the shortest path back. The user arrived at the current page after numerous distractions, dead ends and fruitless loops. These would not appear in the "breadcrumb" navigational device employed by website users. However, every misstep, backstep and repetition is potential gold to the prospectors panning the clickstream of player experience for the paydirt of serious games and adaptive training efficacy.

Applying ANOVA techniques against this richer but unfiltered record of player choices, good and bad, yields interesting insights. Evidence shows what game designers and classroom teachers have long known: more learning is earned by those who fail and struggle to correct themselves than those who achieve correct answers easily. This is shown in the "Learning Traffic" maps of the game's level 4 (Figure 3). This visualization was inspired by Edward Tufte's data presentation for Charles Minard's map of Napoleon's Russian campaign (Morais, 2014). For the full version of the game, half the people who began the level learned the lesson. Two thirds of those who finished it learned. This visualization helped determine where the learning occurred and where the attrition occurred. (Jacobson, 2013)

For early challenges in the game, more learners make incorrect decisions. In the figure below, each scene with a decision in level 4 of the game is shown as a box. The number and type of players that made each decision is shown next to the paths out of the box. The optimum path is displayed left to right along the horizontal. Branching above the horizontal are the result of incorrect or "less than optimal" decisions. The blue path and numbers represent test subjects (players) for whom the game was substantially effective – i.e. their learning was significant as measured by comparing test scores before and after playing the game. As can be seen by the size of the traffic and the numbers, the majority of (three times more) players in level 4 did not make the optimum choice for the first decision regardless of whether they were learners or not (23+25 versus 8+7). For the second decision made in this level (made at the mess area), the incorrect choices were almost evenly split among the learners (11 versus 15) but the majority of non-learners choose the choice that led them to the campfire (22) as compared to the choice that would have led them to the money tent (8).

By applying the traffic map style of analysis to the breadcrumb data gathered in the initial study the developmental team was able to refine user choices to pinpoint where the embedded decision tree did not support the learning design. The utility of the breadcrumb data was found to provide the conclusive evidence that solved major debates among the design team on how to enable the game to better support user learning. For example, this analysis showed that the mess area needed to better prepare the learner; the fighting scene should be removed as it was not critical and the mistakes in that scene led to major attrition; and the money tent scene should be improved.

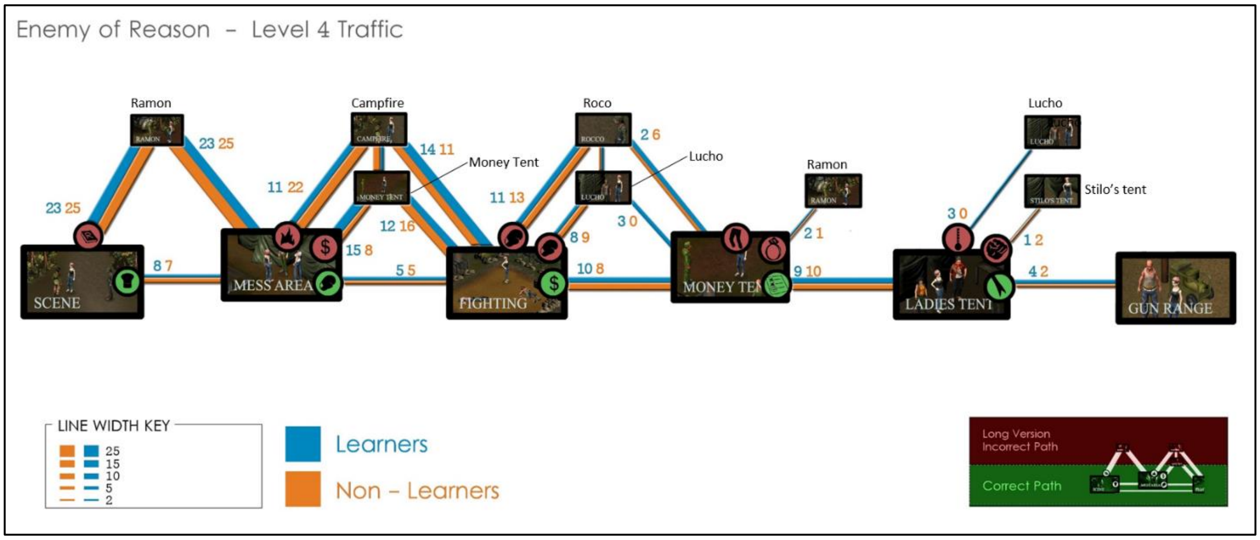

Figure 3. Traffic Map Data Visualization

Discovering Shopping Cart Mom

When we had less than desired learning effects for FAE and BBS, we applied data analysis techniques on the breadcrumb data looking for any correlation. We did not find any except where the measures were made in the test. Because the breadcrumb data was recorded for every decision the player made, we concluded that, as the analysis showed, there were no optional paths that were at fault. We concluded that it had to be a branchless area of the game present in both the short and long version that was causing people to get confused. That focused our attention on the common tutorial, and then we used human review to isolate down to a scenario titled *Shopping Cart Mom*. This scenario was authored as a complex example of multiple biases and was made to be an effective teaching moment only when used with carefully authored instructional feedback. However, during the study, the scenario was used in the tutorial when introducing the biases. We believe that without the instructional feedback, it probably confused the player more than it helped. The focused hunt enabled us to rapidly identify where to make changes to the version that was delivered for IV&V testing, such as moving *Shopping Cart Mom* from the tutorial to where it was used with instructional feedback. Improvements like this resulted in a game that resulted in higher IV&V effectiveness measures than the version we tested.

INDEPENDENT VALIDATION OF RESULTS

We incorporated changes identified by the breadcrumb analysis and provided the game for independent testing of the Phase 1 games conducted by The Johns Hopkins University Applied Physics Laboratory (JHU/APL) on behalf of the Government. Figure 4 shows combined bias scores for all three biases, higher (100) is better. The percentages reflect percent improvement relative to pretest scores. The modified *Enemy of Reason* game is shown as the third set of columns from the left and the control condition using the video is the last set of columns on the right. This IV&V study used the same protocol, however, the results are based on a different set of test instruments, so a comparison is not conclusive. However, we are confident that the changes we made based on the data analysis such as moving *Shopping Cart Mom* to an advanced level did indeed improve the game.

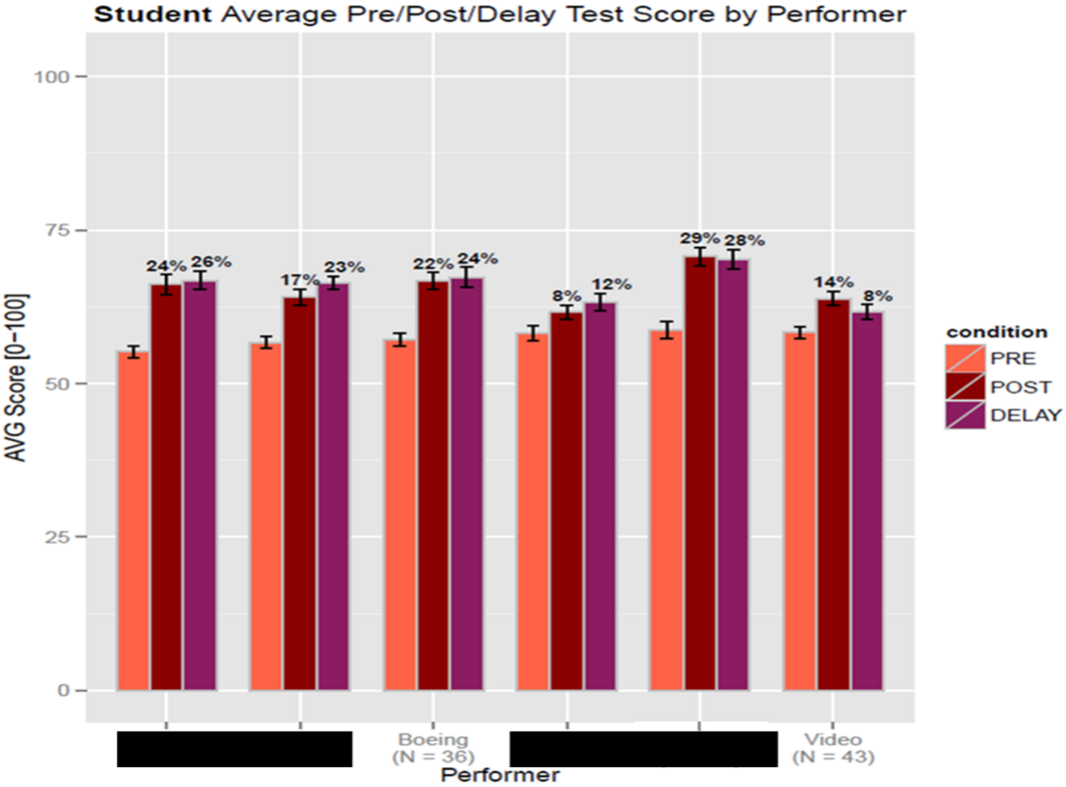

Figure 4. Independent testing of the Sirius Phase 1 games conducted by JHU/APL

SUMMARY

The original plan to use online survey services and apply traditional trend analysis on pre- and post-test on knowledge and behavioral adaptation actually increased work load on the developmental team and led to false conclusions due to a limited data set (only one user trial). The inconclusive results caused from initial trend data forced the design team to consider alternative analytical methods.

An expanded analytics of the breadcrumb data, once understood, provided very insightful analysis of the learning being accomplished by the test audience. It proved to be more pragmatic, objective, and precise than using the competing "poll numbers" method of online surveys, SME murder boards, and the analysis of trends from early user tests and the initial large group trial (the analysis of the trial results were not aimed at learning analysis within the game nodes; the analysis was to gauge and measure progress toward program objectives rather than answering why trends were occurring within the game). After applying changes identified by the breadcrumb analysis, immediate bias reduction increased from 10% to 22% in the immediate post-test and from 2% to 24% in the retention post-test.

The traffic map graph was a valuable tool for analyzing specific learning experiences - but it could be improved. Next time, instead of measuring aggregate traffic at each edge of the graph, we will consider each individual's path through the entire level. By comparing these breadcrumb trails, we can determine the most constructive paths of success and failure to ensure that little failures by the learners in the game lead to fewer failures by the learners in their daily work.

ACKNOWLEDGEMENTS

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory contract number FA8650-12-C-7234. 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.

The team that designed and developed *The Enemy of Reason* was led by Boeing Research & Technology. Boeing Training Systems and Government Services provided psychologists and statisticians to design the studies, design and develop the test instruments, and perform analysis of the game's training effectiveness. Boeing Kestrel provided experience in U.S. Department of Defense and non-DoD intelligence, gaming and simulation; Toffler Associates provided experts in cognitive bias recognition and mitigation instruction for the intelligence community; GamesThatWork (GTW) designed and produced the game and provided Emmy-nominated script writers; and Mercyhurst University Institute for Intelligence Studies, provided expertise in game design to teach intelligence analysts, graduate students, and access to 350 student analysts for the studies.

REFERENCES

BR&T. (2013). *The Enemy of Reason Phase 1 Final Report, Contract # FA8650-12-C-7234.* St. Louis, MO: Boeing Research & Technology (BR&T).

Dargue, B., Jacobson, D., & Sanders, J. (2014). Transmedial and Paramedial Serious Game Deployment. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).* Orlando, FL.

Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (1999). *How people learn: Bridging research and practice.* Washington, D.C.: National Academy Press.

Dual process theory. (2014, May 15). Retrieved June 17, 2014, from Wikipedia, The Free Encyclopedia: http://en.wikipedia.org/w/index.php?title=Dual_process_theory&oldid=608744330

Evans, J. S. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology 2008, 59*, 255-278.

Halpern, D. F., & Hakel, M. D. (2003). Applying the science of learning to the university and beyond. *Change, 35*(4), pp. 36-41.

Jacobson, D. (2013). Learning by Losing. *Serious Play Conference.* Redmond.

Morais, C. D. (2014, May 29). *Mapping Time: A Detailed Look at Minard's Flow Map*. Retrieved from GIS Lounge: http://www.gislounge.com/mapping-time-detailed-look-minards-flow-map/