Measuring Training Effectiveness of Lightweight Game-based Constructive Simulation & using Statistical Software Available to your Students

Dr. Jonathan Stevens
Osceola Campus
Agenda

• A Little About Me

• Part 1: Training Effectiveness Evaluation
  • Introduction
  • Background
  • Method
  • Results
  • Conclusion

• Part 2: StatCrunch Use
  • StatCrunch Background
  • StatCrunch Real-World Use
  • StatCrunch Classroom Use
  • StatCrunch Demo/Orientation
A Little About Me

• Valencia tenure-track Math Professor
  • STA 2023
  • MAC 1105
  • MAT 1033c
  • MGF 1106

• Part-time Principal Investigator/Research Scientist with UCF Institute for Simulation & Training
  • DoD training simulation/human performance-related research
  • USG-sponsored research
Part 1

Training Effectiveness Evaluation
Measuring Training Effectiveness of Lightweight Game-based Constructive Simulation

Jonathan Stevens, Ph.D. University of Central Florida
Latika Eifert, United Stated Army Research laboratory
jonathan.stevens@knights.ucf.edu
Introduction

- **Research Purpose:** examine the training effectiveness of the Linguistic Geometry Real-time Adversarial Intelligence and Decision-making (LG-RAID) simulation tool in an operationally relevant environment

- **LG-RAID:**
  - Light-weight software tool
  - Generates intelligent, predictive and tactically-correct Courses of Action (COAs) for Soldiers
  - Echelon: company and below
  - Employs Linguistic Geometry (LG)
Introduction

• LG-RAID:
  • Proven to be an effective simulation driving other constructive simulations
  • May also be utilized as a game-based trainer

• However,
  • No empirical assessment has been conducted on its training effectiveness and usability as a stand-alone game
  • Minimal empirical evidence exists regarding the effectiveness of game-based training
LG-RAID Background:

- Scalable approach based on the avoidance of tree-based searches typical to other gaming systems
- Defense Advanced Research Projects Agency (DARPA) - initiated project
- Currently provides a low-overhead COA generation and wargaming capability
  - Tactical applications
  - Training applications
LG-RAID as a Training Application:

• User can employ LG-RAID to plan, evaluate, visualize and iterate alternative COAs for an assigned mission:
  • User Input: friendly forces, enemy forces & scheme of maneuver
  • Simulation Execution: LG-RAID executes the scenario, generating tactically valid COAs
  • Output: results presented to user as an animated movie & statistical BDA output
  • Iterate: execute alternative COAs to qualitatively and quantitatively compare outcomes
Game-Based Training (GBT):

- Game-like qualities are being incorporated into computer-based simulation training systems
  - Scores and points
  - Earning trophies/rewards
  - Avatar personalization, etc.
- GBT training effectiveness research highlights:
  - Most beneficial when used in combination with other traditional forms of instruction
  - Capable of promoting high levels of effective learning in simulated training systems
  - Game-like features of GBT increase user engagement and motivation for learning - positive influence on learning
Method

• **Research Purpose:** examine the training effectiveness of the LG-RAID simulation tool in an operationally relevant environment

• **Participants:**
  - Ten U.S. Soldiers participated in this experiment
  - Each participant was a recent graduate of the Maneuver Captain's Career Course (MCCC) located at Fort Benning, GA
  - Participants were randomly chosen from the course:
    - Average Age (N = 10, M = 29.1, SD = 2.9)
    - Average Years in Service (N = 10, M = 5.7, SD = 2.2)
Method

- **Method**: empirical assessment of the degree of training transfer of the LG-RAID condition in comparison to a control treatment
  - Kirkpatrick's model for evaluating training programs
  - Focused on Level II (Learning Criteria): whether or not there was an increase or decrease in the student's knowledge or capability as a consequence of using the simulation.
  - Training effectiveness: whether or not there was a qualitative difference discovered in the tactical plans produced between treatment groups
Design of Experiment:

- Experiment conducted at the Maneuver Center of Excellence (MCoE)
  - Sample: recent graduates of the MCCC
- Students were assigned a tactical mission by the MCCC Small Group Leader (SGL), in a classroom setting
  - Given five hours to develop a tactically sound plan
- One independent variable - training condition
  - Experimental group: used LG-RAID simulation software
  - Control group: employed traditional method of planning (mapboard)
Design of Experiment (continued):

• Dependent variables:
  • Performance: measured by the SGL's evaluation and assessment of the student's tactical plan, using the SGL Evaluation Survey
    • 11 distinct performance measures, 7-point Likert scale
  • Survey responses: System Usability Questionnaire & Trainee Feedback Survey
Method

Design of Experiment:

Day 1:
Training
0900 - 1600

Day 2:
Receive & Plan Mission
0800 - 1400

Day 2:
Assessment
1400 - 1600

Soldiers

Inbrief
- LG-RAID block of instruction
- Questionnaires

Experimental Group
LG-RAID

Treatments

Control Group
Traditional

Simulation may be employed by participants in any manner

Brief Plan

- Random assignment
- Receive mission
- Plan mission

- Brief plan to SGL
- SGL evaluates tactical plan
Method

Day 1 Training

Day 2 Experimental (LG-RAID) Treatment
Method

Simulation Output
Method

Design of Experiment:

• Experimental Group:
  • Utilized the LG-RAID simulation software
  • Rapidly wargame and iterate their plan development

• Both treatments: the student's objective was to create the best tactical plan

• Tactical scenario:
  • From MCoE's MCCC existing Program of Instruction (POI)
  • Stryker company, with combined arms enablers, attacking enemy forces in an urban environment
  • Developed and delivered by certified U.S. Army instructors

• Instructor (SGL) evaluation:
  • Single-blind
    • The student briefed the SGL on his plan
    • Instructor evaluated the student's plan
Results

• The SGL Evaluation Survey was composed of 11 distinct performance measures, each of which was scored on a 7-point Likert scale.

• The 11 performance measures were then aggregated into five performance areas:

<table>
<thead>
<tr>
<th>Performance Area</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tactical Plan</td>
<td>Proper employment of tactics, techniques and procedures</td>
</tr>
<tr>
<td>Mission Command</td>
<td>Student's exercise of authority and direction as the commander using mission orders to enable disciplined initiative within the commander's intent</td>
</tr>
<tr>
<td>Holistic Plan</td>
<td>Whether the student's plan utilized all available resources, was complete, and included appropriate contingency plans</td>
</tr>
<tr>
<td>Analysis</td>
<td>Whether the student's plan addressed tactical risk and incorporated proper terrain and enemy analyses</td>
</tr>
<tr>
<td>Total Performance</td>
<td>Student's aggregated score of the four performance areas</td>
</tr>
</tbody>
</table>
# Results

<table>
<thead>
<tr>
<th>Treatment</th>
<th>LG-RAID</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance Area</strong></td>
<td><strong>Mean</strong></td>
<td><strong>SD</strong></td>
</tr>
<tr>
<td>Tactical Plan</td>
<td>5.27</td>
<td>0.72</td>
</tr>
<tr>
<td>Mission Command</td>
<td>4.40</td>
<td>0.55</td>
</tr>
<tr>
<td>Holistic Plan</td>
<td>4.67</td>
<td>0.41</td>
</tr>
<tr>
<td>Analysis</td>
<td>5.20</td>
<td>1.17</td>
</tr>
<tr>
<td><strong>Total Performance</strong></td>
<td>4.93</td>
<td>0.63</td>
</tr>
</tbody>
</table>

*Mean* and *SD* values for each performance area comparing LG-RAID and Traditional treatments.

**Performance Results**

- Tactical Plan: LG-RAID (5.27 ± 0.72) vs. Traditional (5.42 ± 1.10)
- Mission Command: LG-RAID (4.40 ± 0.55) vs. Traditional (4.13 ± 0.48)
- Holistic Plan: LG-RAID (4.67 ± 0.41) vs. Traditional (5.00 ± 1.22)
- Analysis: LG-RAID (5.20 ± 1.17) vs. Traditional (4.75 ± 0.88)
- Total Performance: LG-RAID (4.93 ± 0.63) vs. Traditional (4.89 ± 0.79)
Results

Statistical analysis:
• Small sample size: preliminary results

• Series of independent two-sample Student's t-Tests found no significant effect of treatment on:
  • Tactical Plan performance $t(5) = -0.23, p = 0.82$, 95% CI [-1.8, 1.5]
  • Mission Command performance $t(7) = 0.8, p = 0.45$, 95% CI [-0.5, 1.1]
  • Holistic Plan performance $t(4) = -0.53, p = 0.63$, 95% CI [-2.2, 1.5]
  • Analysis performance $t(7) = 0.66, p = 0.53$, 95% CI [-1.2, 2.1]
  • Total performance $t(6) = 0.08, p = 0.94$, 95% CI [-1.2, 1.2]

• Non-parametric Mann-Whitney test also indicated no difference in the average performance for the LG-RAID group ($Mdn = 3.0$) and the control group ($Mdn = 6.5$), $U = 19$, $p = .19$

All statistical analysis conducted using StatCrunch
Results

- Qualitative analysis was collected through a post-hoc guided discussion between the Principal Investigator (PI) and SGL
  - The students' treatments were revealed to the SGL (post-hoc) in order that he may detect and describe general trends, differences and similarities between the two groups' performance
    - Experimental treatment:
      - Did not employ phased operational planning
      - Were better able to articulate their plan's sequence of events
      - Produced less structured orders than their control group peers
      - Plans contained more detail
    - SGL's holistic assessment was congruent with our quantitative analysis; there was no significant difference in performance between both treatments
Conclusion

• Low sample size renders our results as preliminary
• No significant difference in performance between both treatments, despite two important factors
  • LG-RAID treatment incurred an approximate 20% time penalty
    • Student briefing conducted in the control group's format
    • Students received only an abbreviated exposure to the simulation
• LG-RAID simulation primarily employed three gaming elements within the simulation to improve Soldier performance in tactical planning:
  • Challenge: validated simulated enemy
  • Uncertainty: simulated engagements may have taken multiple paths
  • Progress: simulation's ability to depict and playback the individual's performance
Conclusion

Lessons Learned:
• Experiment placed the LG-RAID treatment at a time disadvantage
• Soldiers in experimental group felt they would have performed better on their evaluated plan if they had a better familiarization with the simulation

Recommendations for Future Research:
• Future data collection events to increase sample size and reduce probability of Type II error
• Future research will examine the usability and functionality feedback obtained at this, and future, data collection events
• Incorporate Lessons Learned (neutral briefing format & additional training time) in order to increase internal validity
Part 2

StatCrunch Use
StatCrunch Background

- Pearson statistical software package
- Bundled with MyStatLab
  - Students have purchased
  - No extra cost
  - Can be used separately from MSL
- Web-based, lightweight
  - After initial log-in, no internet access required
  - Can be used on laptop, tablet & mobile phone
- Similar interface to Excel, MiniTab, others
StatCrunch Real-World Use

• Statistical techniques performed in this paper using StatCrunch:
  • Independent two-sample T-Tests
  • Independent two-sample confidence intervals
  • Non-parametric Mann-Whitney test
• Other Papers:
  • Simple linear regression
  • Multiple linear regression
  • Single-factor ANOVA
  • Multiple factor ANOVA
  • Wilcoxon Signed Ranks
  • Kruskal Wallis Test
  • Chi-Square Goodness of Fit
  • Normality Goodness of Fit Test
  • Tukey HSD

Students motivated by seeing real-world application of the tool
• In my STA 2023 sections, students bring their laptops and use StatCrunch in the following ways:
  • Visual analysis of data:
    • DotPlot, Histogram, StemPlot, Bar Chart, Pie Chart, BoxPlot, ScatterPlot
  • Descriptive Statistics
    • Mean, SD, Median, IQR, Q1, Q3, Mode, Five-Point Summary
  • Simple Linear Regression
    • Correlation Coefficient, R-square, model construction
  • Normal probability distribution
  • Random sampling
  • Simulation
  • Statistical Inference
    • One-sample proportion hypothesis testing & confidence intervals
    • Two-sample proportion hypothesis testing
    • One-sample mean hypothesis testing & confidence intervals
    • Independent two-sample mean hypothesis testing & confidence intervals
    • Dependent two-sample mean hypothesis testing

*Students motivated by seeing real-world application of the tool*
StatCrunch Demo/Orientation

This Paper

Microsoft Excel Worksheet

Hypothesis test results:
μ₁ : Mean of LG-RAID
μ₂ : Mean of Traditional
μ₁ - μ₂ : Difference between two means
H₀ : μ₁ - μ₂ = 0
Hₐ : μ₁ - μ₂ ≠ 0
(without pooled variances)

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Diff.</th>
<th>Std. Err.</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>μ₁ - μ₂</td>
<td>-0.1495</td>
<td>0.63948958</td>
<td>4.9806787</td>
<td>-0.2337802</td>
<td>0.8245</td>
</tr>
</tbody>
</table>

95% confidence interval results:
μ₁ : Mean of LG-RAID
μ₂ : Mean of Traditional
μ₁ - μ₂ : Difference between two means
(without pooled variances)

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Diff.</th>
<th>Std. Err.</th>
<th>DF</th>
<th>L. Limit</th>
<th>U. Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>μ₁ - μ₂</td>
<td>-0.1495</td>
<td>0.63948958</td>
<td>4.9806787</td>
<td>-1.7952802</td>
<td>1.4962802</td>
</tr>
</tbody>
</table>

Hypothesis test results:
m₁ = median of LG RAID
m₂ = median of Control
m₁-m₂ : m₁ - m₂
H₀ : m₁-m₂ = 0
Hₐ : m₁-m₂ ≠ 0

<table>
<thead>
<tr>
<th>Difference</th>
<th>n₁</th>
<th>n₂</th>
<th>Diff. Est.</th>
<th>Test Stat.</th>
<th>P-value</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>m₁ - m₂</td>
<td>5</td>
<td>4</td>
<td>-2.5</td>
<td>19</td>
<td>0.1905</td>
<td>Exact</td>
</tr>
</tbody>
</table>
Graphs

Variable: Stemplot
Decimal point is 1 digit(s) to the right of the colon.
Leaf unit = 1
0 : 11111222333334
0 : 5556668
1 : 00
1 : 57
2 : 0
2 : 5
3 : 00
High: 40
# Descriptive Stats

![Microsoft Excel Worksheet](image)

<table>
<thead>
<tr>
<th></th>
<th>Column</th>
<th>n</th>
<th>Mean</th>
<th>Variance</th>
<th>Std. dev.</th>
<th>Std. err.</th>
<th>Median</th>
<th>Range</th>
<th>Min</th>
<th>Max</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>9</td>
<td>59.333333</td>
<td>6.3125</td>
<td>2.5124689</td>
<td>0.83748964</td>
<td>59.5</td>
<td>8</td>
<td>55</td>
<td>63</td>
<td>58</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>15</td>
<td>55.266667</td>
<td>7.6380952</td>
<td>2.7637104</td>
<td>0.71358696</td>
<td>55</td>
<td>12</td>
<td>51</td>
<td>63</td>
<td>53.5</td>
<td>56</td>
<td></td>
</tr>
</tbody>
</table>
StatCrunch Demo/Orientation

Simple Linear Regression

**Simple linear regression results:**
Dependent Variable: Weight
Independent Variable: Height
Weight = -442.88235 + 9.0294118 Height
Sample size: 6
R (correlation coefficient) = 0.88093363
R-sq = 0.77604407
Estimate of error standard deviation: 14.141876

**Parameter estimates:**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>Alternative</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-442.882</td>
<td>155.327</td>
<td>≠ 0</td>
<td>4</td>
<td>-2.851</td>
<td>0.0463</td>
</tr>
<tr>
<td>Slope</td>
<td>9.0294</td>
<td>2.4253</td>
<td>≠ 0</td>
<td>4</td>
<td>3.723</td>
<td>0.0204</td>
</tr>
</tbody>
</table>

**Analysis of variance table for regression model:**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F-stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>2772</td>
<td>2772</td>
<td>13.86</td>
<td>0.0204</td>
</tr>
<tr>
<td>Error</td>
<td>4</td>
<td>799.97</td>
<td>199.992</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>3572</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>