VALVE’S APPROACH TO PLAYTESTING: THE APPLICATION OF EMPIRICISM

Mike Ambinder, PhD
Experimental Psychologist

VALVE
Goal

- Review pros/cons of various playtest methodologies
- Discuss which data is best derived from which methodology
- Focus more research on user research
Overview

- Valve’s (external) playtest philosophy
- Traditional playtest methodologies
  - Qualitative
- Technical playtest methodologies
  - Measured
Overview

• Traditional Playtest Methodologies
  – Direct Observation
  – Verbal Reports
  – Q&As

• Technical Playtest Methodologies
  – Stat Collection/Data Analysis
  – Design Experiments
  – Surveys
  – Physiological Measurements
Valve’s Game Design Process

- Goal is a fun game →
  - Game designs are hypotheses →
  - Playtests are experiments →
  - Evaluate designs off playtest results →
- Repeat
Playtesting Goal

- Fun
- Not bug testing
- Not game balancing
- DEFINITELY not focus testing
Ancillary Benefits

- Idea generation
- Identify problem areas
- Solve design arguments
- Aid other production aspects
Valve’s Philosophy

- We want to make informed decisions
  - Get data early, get data often
  - Iterate constantly
- We don’t know what’s best (players do)
- Create a feedback loop between design and playtest
Valve’s Philosophy

- Playtesting continues after we ship
  - Gameplay stats
  - Forum responses
  - Fan feedback
- Always gathering data for the future
  - Patches/updates
  - Upcoming games
Traditional Methods

- Direct Observation
- Verbal Reports
- Q&As
Direct Observation
Direct Observation

- “Typical” playtest
  - Watch people play the game
  - Observe their gameplay/behavior
  - Simulate at-home experience

- Have a design goal
Direct Observation

+ Get a feel for player interaction with game
+ Importance of what people do—not what they say
  – Presence of observers can bias results
  – Salient event can slant interpretation
  – Behavior requires interpretation
Verbal Reports

Zombies are scary… the assault rifle is my favorite…
Verbal Reports

- Think-aloud protocol:
  - People describe their actions as they play
  - Unprompted and uncorrected
- In conjunction with direct observation
Verbal Reports

+ Enables realtime glimpse into player thoughts, feelings, and motivations
+ Bring up unnoticed details
+ Effective for ‘why’ questions
  – Interfere with gameplay/create an artificial experience/distracting
  – Inaccurate and biased
Q&A
Q&A

- Structured (usually) querying of playtesters
- Validate playtest goals
- Source of supplemental information
Don't shoot teammates!
Q&A

- Answer specific design questions
- Determine specific player intent
  - Group biases (anchoring, social pressure, saliency, etc.)
  - People don’t know why they do what they do
  - Potential for biased questions
Our Q&A Procedure

- Survey
- Individual Q&A
- Group Q&A
- Be cautious
Benefits of Traditional Methods

+ Nothing beats direct gameplay observation
+ Determine major gameplay, navigation, and content issues
+ Get an idea of player thoughts/mental models
+ Get feedback on design choices
Issues with Traditional Methods

- Artificial gameplay sessions
- Many potential biases
- Distorted data (interpreted behavior)
- Lack of empiricism
- Missing elements of objectivity
- Sometimes difficult to establish emotions, baselines, and independence
Technical Approaches

- Stat collection/analysis
- Design experiments
- Surveys
- Physiological measurements
Stat Collection/Analysis
Stat Collection/Analysis

- Record of gameplay behaviors
  - Deaths, level times, friendly fire, ...
- Objective measurements
- Aggregate perspective
- Quantify behavior
- Opportunity for analyses
  - T-tests
  - Regressions
  - ...
<table>
<thead>
<tr>
<th>Achievement name</th>
<th>% of players</th>
</tr>
</thead>
<tbody>
<tr>
<td>DROE AND DROP: Recover a Survivor from a Smoker's tongue before he takes damage</td>
<td>52.4%</td>
</tr>
<tr>
<td>BEAN VALAD: Make 100 headbutts in 1000 seconds</td>
<td>90%</td>
</tr>
<tr>
<td>TONGUE TWISTER: Kill a Smoker who licks grabbed you with his tongue.</td>
<td>68.6%</td>
</tr>
<tr>
<td>BLOODreetings: You or another Survivor take no damage after being rammed on by a Smoker.</td>
<td>88.6%</td>
</tr>
<tr>
<td>HIT BODYGUARD: Prevent a Survivor from an attacking Infected 50 times.</td>
<td>81.4%</td>
</tr>
<tr>
<td>TANK-BUSTERS: Kill a Tank without it doing any damage to a Survivor.</td>
<td>84%</td>
</tr>
<tr>
<td>PREDATOR: Protect 20 infected in a single explosion.</td>
<td>81.7%</td>
</tr>
<tr>
<td>NO SMOKING SECTION: Kill 100 Smokers or they are pulled behind Survivors.</td>
<td>78.5%</td>
</tr>
<tr>
<td>OUTBREAK: Take a crew of infection, then pass it on to someone else.</td>
<td>70.4%</td>
</tr>
<tr>
<td>HUNTER HUNTER: Drive a hunter off of a penned and helpless Survivor.</td>
<td>76.4%</td>
</tr>
<tr>
<td>101 COMPLAINTS: Get 101 infected on file.</td>
<td>75.6%</td>
</tr>
<tr>
<td>HEAD CLOSET: Release a survivor trapped in a closet.</td>
<td>74.6%</td>
</tr>
<tr>
<td>TOWING INFECTED: Light a Tank with a Molotov.</td>
<td>72.8%</td>
</tr>
<tr>
<td>WITCH HUNTER: Kill a Witch without any Survivor taking damage from her.</td>
<td>70%</td>
</tr>
<tr>
<td>NO-ONE LEFT BEHIND: Beat a campaign with one Survivor.</td>
<td>64.3%</td>
</tr>
<tr>
<td>FINAL TAP: Kill an infected with a single blow from behind.</td>
<td>66.6%</td>
</tr>
<tr>
<td>GROUND COVER: Save a locked Survivor from a Special Infected while on the ground.</td>
<td>64.1%</td>
</tr>
<tr>
<td>DEAD STOP: Finish a marathon up to 100 kilometers.</td>
<td>63.3%</td>
</tr>
<tr>
<td>BURN THE WITCH: Light a Witch with a Molotov.</td>
<td>67.6%</td>
</tr>
<tr>
<td>MERCY KILLER: Survive the No Mercy campaign.</td>
<td>97.8%</td>
</tr>
<tr>
<td>JUMP SHOT: Meet a monster while he's leaping.</td>
<td>59.9%</td>
</tr>
<tr>
<td>TOLL COLLECTOR: Survive the Death Toll campaign.</td>
<td>54%</td>
</tr>
<tr>
<td>DEAD BARON: Survive the Dread 24th campaign.</td>
<td>53.6%</td>
</tr>
<tr>
<td>Challenge</td>
<td>Percentage</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>GROUND COVER</td>
<td>64.1%</td>
</tr>
<tr>
<td>Save another Survivor from a Special Infected while on the ground.</td>
<td></td>
</tr>
<tr>
<td>DEAD STOP</td>
<td>63.3%</td>
</tr>
<tr>
<td>Punch a Hunter as he is pouncing.</td>
<td></td>
</tr>
<tr>
<td>BURN THE WITCH</td>
<td>61.6%</td>
</tr>
<tr>
<td>Light a Witch with a Molotov.</td>
<td></td>
</tr>
<tr>
<td>MERCY KILLER</td>
<td>57.6%</td>
</tr>
<tr>
<td>Survive the No Mercy campaign.</td>
<td></td>
</tr>
<tr>
<td>JUMP SHOT</td>
<td>55.8%</td>
</tr>
<tr>
<td>Headshot a Hunter while he's leaping.</td>
<td></td>
</tr>
<tr>
<td>TOLL COLLECTOR</td>
<td>54%</td>
</tr>
<tr>
<td>Survive the Death Toll campaign.</td>
<td></td>
</tr>
<tr>
<td>DEAD BARON</td>
<td>53.9%</td>
</tr>
<tr>
<td>Survive the Dead Air campaign.</td>
<td></td>
</tr>
</tbody>
</table>
Stat Collection/Analysis

+ Objective notions of player behavior
+ See global trends
+ Readily enables comparisons, baseline establishment, and metric creation
+ Track changes over time
  – Averages hide extreme examples
  – Miss nuance (lacking context)
  – Requires rigor
  – Can see ‘illusory’ patterns
Design Experiments
Design Experiments

- Hypothesis testing
  - Compare two or more conditions
  - Collect data
  - Verify hypothesis

- Predict player behavior
  - Define set of variables
  - Investigate resulting relationships
# TEAM FORTRESS 2

## THE SCOUT UPDATE

The results are in, the update's out! Now it's time to...

**PLAY BALL!**

### Community Voted Unlockables Order

<table>
<thead>
<tr>
<th>Rank</th>
<th>Unlockable</th>
<th>Votes</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Force-a-Nature</td>
<td>17,219</td>
<td>42.53%</td>
</tr>
<tr>
<td>2</td>
<td>The Sandman</td>
<td>13,906</td>
<td>34.10%</td>
</tr>
<tr>
<td>3</td>
<td>'Bonk' Energy Drink</td>
<td>9,462</td>
<td>22.37%</td>
</tr>
</tbody>
</table>

© 2008 Valve Corporation; all rights reserved. Valve, the Valve logo, Half-Life, the Half-Life logo, the Lambda logo, Steam, the Steam logo, Team Fortress, the Team Fortress logo, Opposing Force, Day of Defeat, the Day of Defeat logo, Counter-Strike, the Counter-Strike logo, Source, the Source logo, Valve Source and Counter-Strike: Condition Zero are trademarks and/or registered trademarks of Valve Corporation.
Design Experiments

+ Enables more informed decision-making
+ Objective answer
+ Saves time in the long run
  - Costs time (in the short run) and money
  - Right questions aren’t always clear
  - Proper experimental design is a process
Surveys
Surveys

- Set of standardized questions
- Forced choice responses
- Quantify feedback/opinions
- Player categorization
How challenging were the following enemies (1 = very easy; 7 = very hard)?

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boomer:</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Common Infected:</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Hunter:</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Smoker:</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Tank:</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Witch:</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

Please rank order your preference for the following weapons from 1 (most liked) to 12 (least liked)

Assault Rifle _____
Auto Shotgun _____
Dual Pistols _____
Gas Can _____
Hunting Rifle _____
Molotov Cocktail _____
Mounted Turret _____
Pipe Bomb _____
Pistol _____
Propane Tank _____
Pump Shotgun _____
SMG _____
Surveys

+ Get less biased responses
+ Validate responses (repetitive questions)
+ Forced choice helpful for revealing preference
+ Ratings enable time-based comparisons
  - Eliminate nuance
  - Difficulty in converting ratings to meaningful decisions
  - Limited solution space
Physiological Measurements
Physiological Measurements

- Measurements of biological response
- Create proxies of player state
- Involuntary
- Objective—can’t be faked
- Quantify emotion
Valence and Arousal

- Valence = positive or negative emotion
- Arousal = magnitude of emotion
Adapted from Lang (1995)
Heartrate

- Beat to beat interval
- Measure baseline rate and changes
- Most basic measure of arousal
- Fourier transforms to distinguish emotion

Heartrate

+ Simple to collect
+ Accurate correlate of arousal
+ Good metric for comparison
  - Intrusive
  - (Sometimes) delayed response to stimuli
  - Variable
Skin Conductance Level

- Electrical resistance of the skin
  - Correlate with arousal
  - Maybe other emotions as well
- Can look for spikes (both responsive and anticipatory)
Skin Conductance Level

- Excellent correlate with arousal
- Good metric for comparison
- Adept at detecting transient responses
  - Intrusive
  - Susceptible to other factors
- Direct I/O relationship doesn’t exist
Eyetracking

- Camera focused on the eyes
- Determine where the eyes are looking
- Real-time insight into player thought processes
- Blink rate/pupil dilation
DANS, KÖN OCH JAGPROJEKT

På jakt efter ungdomars kroppsspråk och den "synkretiska dansen", en sammansmältning av olika kulturers dans, har jag i mitt fältarbete under hösten rört mig på olika arenor inom skolans värld. Nordiska, afrikanska, syd- och östeuropeiska ungdomar gör sina röster hörda genom sång, musik, skrik, skratt och gestalter känslor och uttryck med hjälp av kroppsspråk och dans.

Den individuella estetiken framträdde i kläder, frisyrer och symboliska tecken som förstärker ungdomarnas "jagprojekt" där också den egna stilen i kroppsrörelserna spelar en betydande roll i identitetsprövningen. Upphållsrummet fungerar som offentlig arena där ungdomarna spelar upp sina performance liknande kroppsspråk.\[\text{Power}][\text{Point}][\text{Text}][\text{Diagram}][\text{Image}][\text{Table}][\text{Figure}]
Eyetracking

+ Effective metric of player attention/gaze
+ Excellent tool for interface design
+ Provides understanding of scene interpretation
  - Expensive
  - Can be intrusive
  - Time consuming
  - Can lead to costly over-analysis
Face Recording

- Observation of facial expression
- Determination of player emotion
- Tied into gameplay
Face Recording

+ Provides emotional context
+ Excellent metric of player emotion
  - Intrusive
  - Requires experienced coders
  - Not always reliable
  - Biased reactions
EEG

- Measurement of electrical potentials in the brain
- Various frequencies are correlated with emotional state
  - Alpha (relaxation)
  - Beta (thinking, engagement)
  - Delta (fatigue)
EEG

+ Good at measuring arousal, engagement, etc.
+ Potential for fairly sophisticated determinations down the road
  - Expensive
  - Very intrusive
  - Noisy
  - Hard to control/validate
EMG

- Sensors placed at varying points on the face
- Measurement of facial muscle contraction/relaxation
- Determinant of emotion based on ‘action units’
EMG

+ Most accurate measure of emotion
+ Real-time determination
- Expensive
- Very intrusive
Other Techniques

- Body temperature
- Gesture recognition
- Muscle tension
- ...
Physiological Measurements

+ More objective measurements of player state
+ Quantifiable emotional response
+ Analysis/comparison metrics
  – Expensive
  – Intrusive
  – Artificial experience
  – Requires experimental control
Benefits of Technical Approaches

+ Application of empirical data to game design
+ Objective (for the most part)
+ Quantify behavior
+ Enable testable hypotheses about player emotional state
Issues with Technical Approaches

- Expensive
- Resource intensive
- Impractical
- Lacking nuance
Summary

- Do your QA early
- Understand pros/cons of existing methods
- Correctly frame design questions
- Be aware of emerging technologies
Acknowledgments

- Charlie Burgin
- Lars Jensvold
- Marc Nagel
- Steve Bond
- John Morello