PERFORMANCE METRICS FOR SERIOUS GAMES:

WILL THE (REAL) EXPERT PLEASE STEP FORWARD?

Loh, C. S. & Sheng, Y. (2013)

Abstract

- *Human Performance literature shows behavioral differences between experts and novices*
- *Experts make decisions differently from novices (many years of practice to achieve mastery)*
- *Competency is a demonstrable attribute based on a person's course of action in problem solving*
- *Telemetry: tracing people's actions and behaviors (as user-generated data) remotely for performance assessment (web navigation, animal movement)*

Experts vs Novices

- Very well-studied phenomenon in T&L & psychology
- Behavioral indicators vary widely
  - Ranging from 'time-to-task completion' rate,
    to mental representations of knowledge,
    to gaze patterns in scanning for information
- Observable & Measurable competency changes
  - Novices ← Competent Users → Experts
  - Novices follow rules (often blindly)
  - Experts (appear to) break/ignore rules at will (because they detect subtle cues that are not obvious to novices)

Serious Games

- *Serious games: designed to support knowledge acquisition and/or skill development*
- *Entertainment ← Digital Games → Serious*
  No ← Performance Assessment → Required
- *ROI: Stakeholders (T&L industries) need "measurable evidence of training or learning"*
- *Gap in Literature: few know what to do*
  - *Thus far, sell games but not assessment reports*
  - *Industry have different criteria for assessment (really complicated if you are an educator)*

Performance Metrics & Analytics

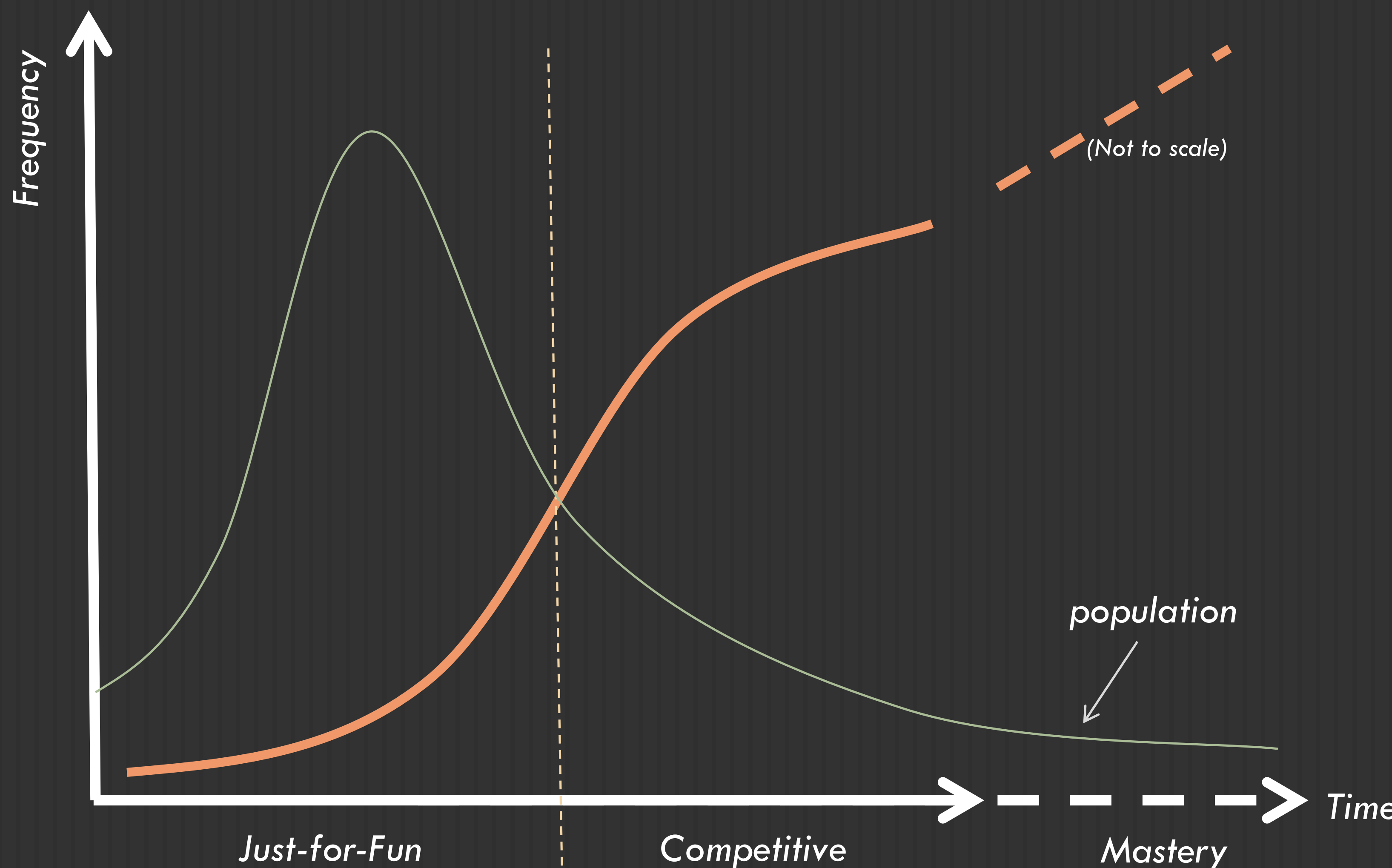
- *Serious Games (for T&L) can provide training so that novices → competent users → experts*
- *To satisfy the needs of stakeholders (for ROI)*
  - *Need STANDARDIZED measurable Performance Metrics to quantify observable changes in competency*
  - *Identify potential metrics*
  - *Test for viability*
  - *Incorporate as SErious Games Analytics (SEGA)*
  - *A set of established performance metrics and industrial standards for measuring competency with SG*

Considering Entertainment Games

- 'Just-for-Fun' mode
  - Why would you want to 'performance assess' me?

- Just for fun?
  - Burger eating competition, Drinking, Car race, etc.
  - Fun → Competition (still fun?)

Different Kind of Games/Players



Considering Competitive Games

- 'Competition' mode: BEST players (in….)
- Best against someone (PvP) → glory and fame, Hall of Fame, Leader board
- Best against self (ghost car) → self improvement
  - Best Time (of completion)
  - Best Route (of navigation) → Trajectory-based
  - Best Utility (of 'limited' resources)
  - Best Collector (of badges)
  - Best Strategy → Objective-based (combination of time, route, resources, etc.)

Best Strategy (Objective-Based)

- Combinations of Time, Route, Resources…
- Many combination
- To start examining the problem, we limit our scope to just the order of completion
  - If you need eggs, shower gel, and video game (how would you shop at Wal-Mart?)
  - Can include Time and Route (but not a must)
- Future: compared ORDER with TIME and/or ROUTE

Similarity in Degree of Competency

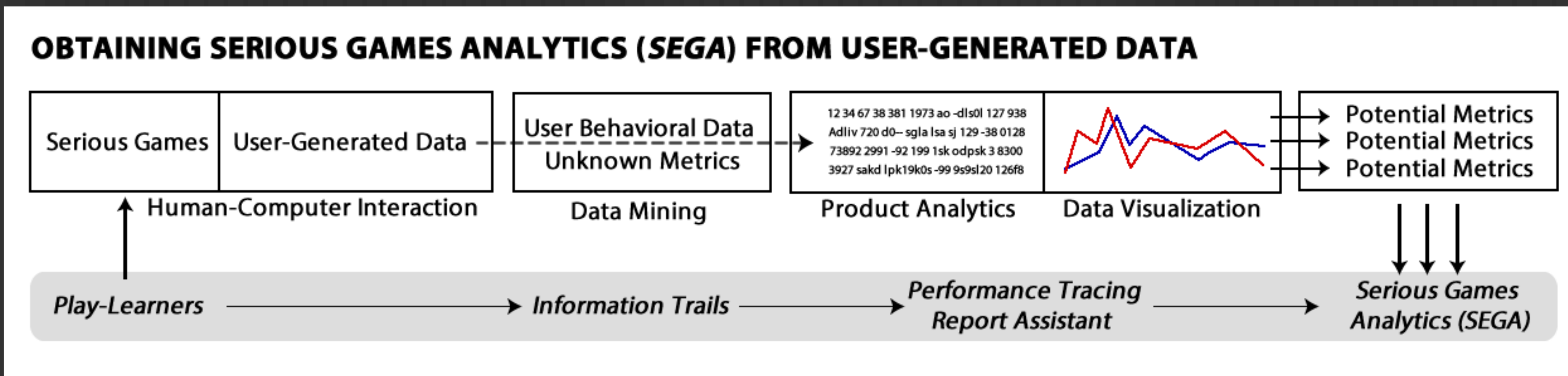
- Since competency is characterized by an observable course of actions taken during problem solving
- Are there differences between course of actions of experts vs novices?
- We compared how closely match the two sets of traces are against one another.
- We calculated the Similarity Index for each player and identified individuals whose performances approach/match that of the experts.

Novice (0) ← Similarity Index → (1) Experts

Logs, Trigger Events

- User-generated data can be collected using a variety of methods
- Information Trails (Loh, 2007), Game Telemetry (Zoeller, 2010)
  - Remote Locale where interaction occurs (online)
  - Event 'Listener'
  - Transmitter/Receiver
  - Home base for database storage and analysis
    - Multiple data points (snowballing effect → massive)
- Analytics → add visualization (reporting purpose)

Information Trails



- *Loh, C. S. (2013). Improving the Impact and Return of Investment of Game-Based Learning. International Journal of Virtual and Personal Learning Environments. 4(1): 1-15.*
- *Loh, C. S. (2012). Information Trails: In-process assessment for game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds). Assessment in game-based learning: Foundations, innovations, and perspectives. (pp.123-144) New York, NY: Springer. [Chapter 8]*
- *Loh, C. S. (2009). Researching and Developing Serious Games as Interactive Learning Instructions. International Journal of Gaming and Computer Mediated Simulations. 1(4): 1-19.*

Route-based Performance Metrics

String Similarity

- *Statistical method devised to determine if two strings/records are similar enough to be duplicates in Record Linkage analysis*
- *Advance uses include facial recognition, DNA sequence similarity, fingerprinting, etc.*
- *Have been used in the analysis of sequences in poker and computer strategy games*
- *But NOT in the differentiation and ranking of human performance (assessment)*
  - *Many types: wikipedia.org/wiki/String_Metric*

String Similarity for Assessment

- Jaccard Similarity Coefficient (or Jaccard Index, JAC)
  - Measure the similarity between two sample sets by dividing the size of their intersection by the size of their union

$$JAC(A, B) = |A \cap B| / |A \cup B|$$

- JAC value ranges from 0 (two completely different strings) to 1 (two identical strings)
  - Easily understood by nonprofessionals

(0% Similarity) 0 ← JAC → 1 (100% Similarity)

Converting String to Bigrams

Example:

- *String A {12345} → Bigrams {12, 23, 34, 45}*
- *String B {13452} → Bigrams {13, 34, 45, 52}*
- $|A \cap B| = |\{34, 45\}| = 2$
- $|A \cup B| = |\{12, 23, 34, 45, 13, 52\}| = 6$
- $JAC(A, B) = |A \cap B| / |A \cup B|$

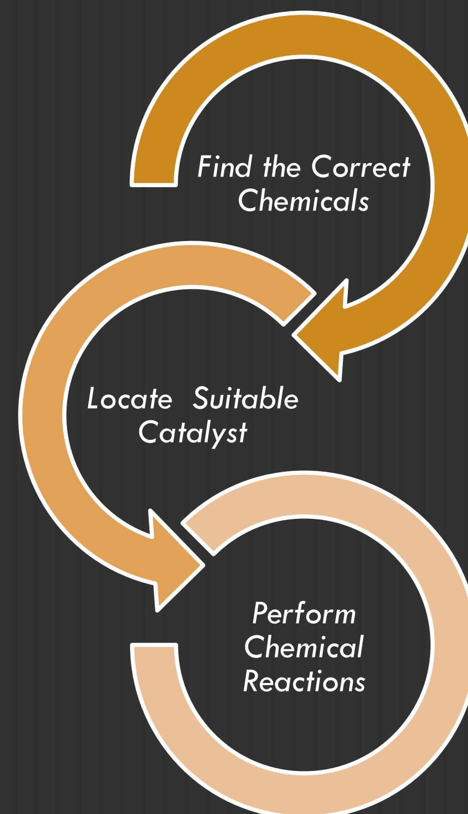
$$= 2 / 6$$

$$= 0.333$$

Story-based Serious Games

Military-style objectives (Search and Rescue)

STEM-based Objectives (Chemical Reaction)



Obtaining 'Action Sequence'

- *Competency may be measured using "observable course of actions" within serious game environments*
- *Depending on player's course of actions (i.e., order of checkpoints visited), an action-sequence can be obtained for each player*
- *In our case,*
  - *Action-sequences happen to start and end with 1 (due to mission giver)*
  - *E.g., 12345671, 13456271, etc.*
- *Consider cases such as 134, 1567, etc. ??*

Findings

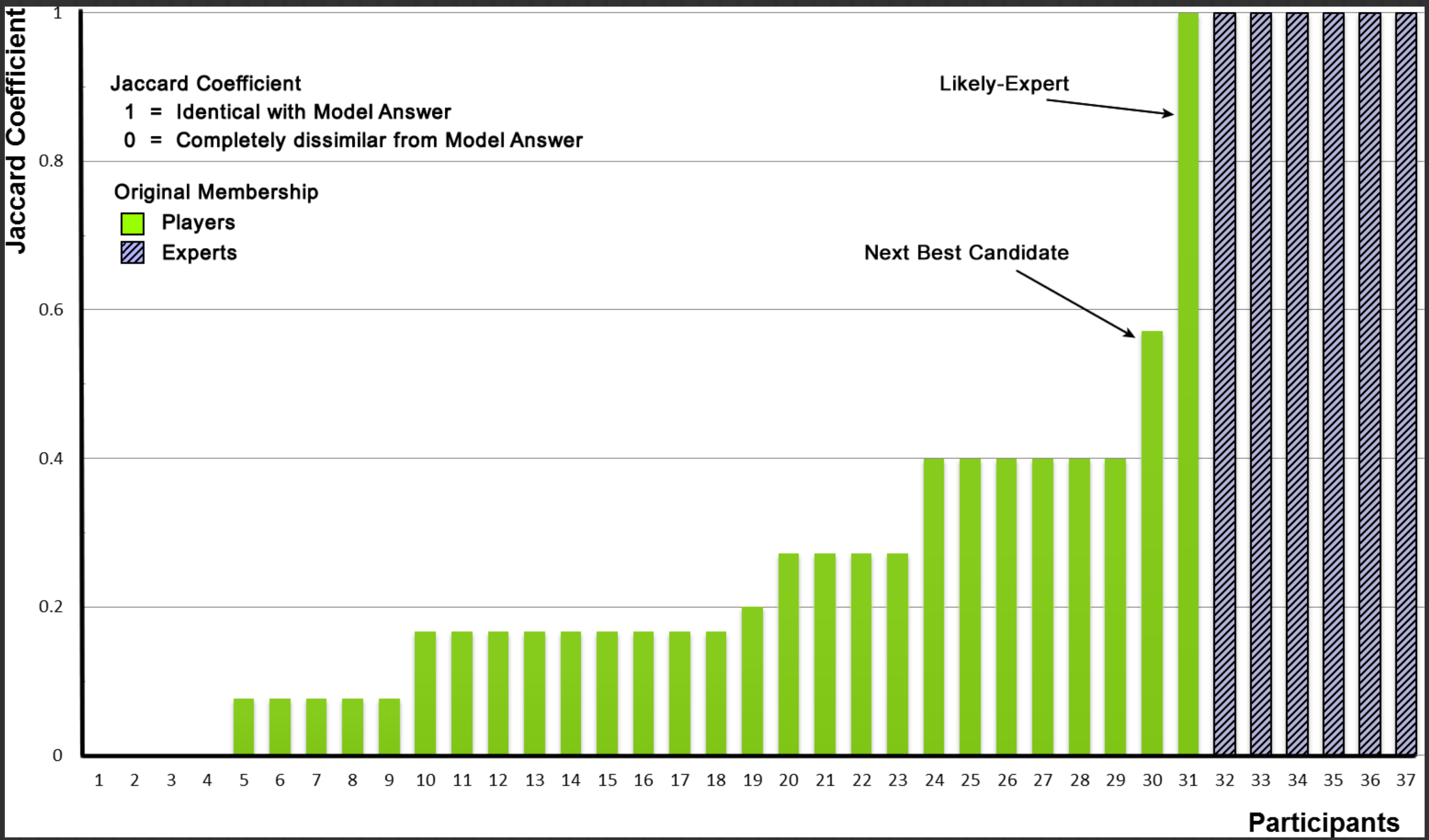
Player Ranking By JAC Values

ID	Number/Identity	JAC Values	Level	Ranking
1 - 6	Design/Testing Team	1	--	Real Expert
7	1 Player	1	1	Expert-rank
8	1 Player	0.57	2	Likely-Expert
9-14	6 Players	0.40	3	Average
15-18	4 Players	0.27	4	Below Average
19	1 Player	0.20	5	Below Average
20-28	9 Players	0.17	6	Below Average
29-33	5 Players	0.08	7	Below Average
34-37	4 Players	0	8	Non-Gamer

Findings

- *Participants who self-identified as avid game players did not automatically score high on JAC.*
- *Only one player achieved Expert rank (JAC =1)*
  - *Never played this game before but had prior game design experience – might explain competency in problem solving using serious game.*
- *Next best player (JAC = 0.57)*
- *The rest falls quickly below 0.5 towards 0*
  - *Performed poorly (low competency, expected)*

Next Best Player



Classification Accuracy

- *We use discriminant analysis with jackknife reclassification to further evaluate the classification accuracy using JAC*
  - *also known as leave-one-out cross-validation*
  - *Particularly useful for small samples where it is difficult to divide the entire data into training and validation datasets.*
- *JAC did a nearly perfect job (97.3%) in reclassification, misclassifying only 2.7% (1 player) out of the total 37 observations.*
- *The success rate was significantly better than the 50% expected by chance ($p < 0.001$).*

By Chance?

- *Simulated sample of 60 experts and 310 players achieve similar result.*
- *Jackknife success rate for simulated sample is 97.48% (with SD = .98%)*
  - *Recall Jackknife for actual data is 97.3%*
- *Better than expected by chance*

Interesting Side Notes

Example: String C = {13}

- Drop out of network (did not complete game)
- Performance by "Time of completion" alone would therefore be erroneous
- JAC = 0 (not always)
- Hence, incomplete data need not be thrown away (conserve economy: little wastage)

Future Research

- *Scenario in this paper depicts 1 model answer*
  - *All experts agree that there is only 1 solution*
- *What if the experts do not agree? Or if there are multiple model answer?*
- *How does String Similarity hold up to Time-of-Completion? (Which one is a better metric?)*

Conclusion

- *Researchers* have suggested that a data-driven approach and an evidence-centered design are much better assessment methods that will foster real adoption of serious games.*
- *Findings in this study suggest string similarity to be a viable performance assessment metric for serious games.*
- *Hope this will encourage others to look into finding appropriate performance metrics for SEGA in the future.*

** [3, 33, 34, 36, 37] referenced in paper*

Publication

- *Loh, C. S., & Sheng, Y. Y. (online first, 2013).*
- *Measuring the (Dis-)Similarity between Expert and Novice Behaviors as Serious Games Analytics.*
- *Education and Information Technologies.*
- *DOI: 10.1007/s10639-013-9263-y*

LinkedIn & Research Gate

- Christian S Loh, Ph.D.

Virtual Environment Lab (V-LAB)
Southern Illinois University
Carbondale, IL, USA
csloh@siu.edu

- Yanyan Sheng, Ph.D.

Dept of Educational Measurement & Statistics
Southern Illinois University
Carbondale, IL, USA