Performance Metrics for Serious Games

Will The (Real) Expert Please Step Forward?

http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6632633&tag=1

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

Yanyan Sheng, Ph.D.
Department of Educational Measurement & Statistics
Southern Illinois University
Carbondale, IL, USA
ysheng@siu.edu

***Abstract*—The literature on human training performance has long attested to the behavioral differences between experts and novices, in which 'competency' is a demonstrable attribute based on a person's *course of action* in problem solving. The advances in technology have made it possible to trace players' actions and behaviors (as user-generated data) within an online serious gaming environment for performance assessment purposes. In this study, we introduce string similarity as a performance metric to identify likely-experts among a group of unknown performers (mixture of novices and experts) according to their in-game course of action in problem solving. Our findings indicate that string similarity is both viable and potentially useful as the first performance metric for Serious Games Analytics (SEGA).**

Keywords—Information Trails, string similarity, performance metrics, serious games analytics

I. INTRODUCTION

A learning activity without assessment is informal at best and comparable to the endeavor of hobbyists. As serious games are "designed to support knowledge acquisition and/or skill development," not only is player-performance assessment an important, if not the most important, aspect of serious game evaluation [1–3], it is also a necessary component for these games to set themselves apart from other entertainment games [4]. Since the advances in technology have made it increasingly easy to trace players' in-game actions and behaviors for performance assessment [5–7], stakeholders in the training and learning industries have begun asking for "measurable evidence of training or learning" to justify their investment and to ensure a good rate of return [8].

Nevertheless, it remains difficult to determine what serious game assessment really looks like because different industries have different performance metrics and assessment criteria. Lacking an established set of standardized performance metrics for serious games, the evidences available are frequently limited to player logs. As 'big data' and game analytics [9] become even more important in informing business decisions in the future, we proposed that a logical step forward is to test, verify, and establish a set of standardized performance metrics, both to satisfy the needs of these stakeholders and to create a baseline for *SErious Games Analytics* (SEGA).

It is known that 'best completion time' is probably the most recognized performance metric used by many first-person shooters, mazes, and puzzle games. A common strategy is to pit players against one another (or themselves) to compete for a spot on the high-score leaderboards depending on how fast players can beat the game or game levels. While the concept of 'best time' (fastest winner) is very intuitive and makes an effective performance metric for entertainment games, the appropriateness of the approach relies heavily on the learning situations. Specifically, in scenarios where play-learners must think critically before applying their skills or knowledge in problem solving, 'best time' can be detrimental to learning. Because those who work/play under time pressure are often tempted into making hasty decisions or taking chances [10–12], such risky behaviors can lead to poor decision habits and even workplace disasters when left unchecked [13].

Other than 'best time', what metric can one use to measure the play-learner's performance in serious games?

II. MOTIVATION

A. Behavioral Differences Between Novices and Experts

The differences between experts' and novices' behaviors in problem solving and decision making has been a very well-studied phenomenon in the training and psychology literature [14, 15]. The indicators of expert-novice behaviors vary widely and can range from time-to-task-completion rate, to mental representations of knowledge, to specific gaze patterns in scanning for information [16]. Novices have a tendency to follow the rules *blindly* when solving problems because they have yet to acquire the context in which those rules operate. As they grow in competency to solve problems, they will gradually learn to apply the right rules with the right conditions. Thus, it is possible to observe competency because it is demonstrable based on a person's *course of action* in problem solving. Experts, who are so in tune with the tasks at-hand, are able to detect cues that are not obvious to non-experts. As a result, experts can be seen as solving problem based on intuition while breaking or ignoring rules, at will.

Proceedings of the 18th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games (CGAMES 2013). (July 30-Aug 1, 2013). Louisville, KY.

B. Ranking of Play-Learners by Performance

The evidences of expert-novice behavioral differences have been reported among airline pilots, teachers, surgeons, nurses, programmers, sportsmen [17–19], and digital game players [20]. The expressed ability to differentiate and rank play-learners by their performance can be a desirable and valuable feature in these fields and professions. Professional schools (such as music performance, business, and medical schools) regularly make use of auditions and selection examinations to rank and select candidates who are more likely to succeed in their programs for admission purposes [21–23]. In cases where a limited number of scholarships, promotions, and job positions are available, stakeholders would require the ability to not only differentiate but also rank candidates by performance. For example, a data mining company recently commissioned a Portal 2 mod to 'pre-select' job applicants: only those who solved an in-game puzzle were invited to apply (see http://jobs.wibidata.com/puzzles/).

As serious games continue evolving into an instrument for training, trainers and stakeholders will need performance metrics that are able to rank learners according to their mastery of the subject – to recognize the top performers for certain scholarships, leadership positions, promotions, etc. Although it is currently not possible for computers to judge effectively the nuances of human performances, this could happen in the future if the right combination of performance metrics can be determined to allow stakeholders to quickly identify individuals who are 'likely experts' from a crowd of masses – hence, the need for SEGA. In summary, identifying performance metrics (or efficient means to measure competency and mastery) is a worthwhile and valuable undertaking for serious games.

C. Metrics to Identify the Best Gamer or Performer

Besides 'best time', there are several other game metrics that game developers have used to help identify best gamers for ranking on the leaderboards. For example, role-playing games and first-person shooters tend to use the number/rate of missions achieved as the preferred performance metrics to identify best gamers. Other measures include the number of kills (i.e., enemy killed), the amount of gold collected, and the amount of experience points gathered. Game developers may also combine several game metrics to yield one composite score for ranking.

Since competency can be characterized by an *observable* course of actions taken during problem solving, we traced the course of actions of a group of experts and a group of novices within a serious game environment and compared the two sets of traces to determine how closely match their actions are. We then calculated the similarity index for each player to identified individuals whose performances approach/match that of the experts.

III. String Similarity Metrics

A. Comparing Unknown Performance to Known Values

String similarity metrics is a statistical method devised to determine if two strings/records are similar enough to be duplicates [24] in Record Linkage analysis [25]. The impetus was to clean out large databases of name-record to remove extraneous data and duplication for data-mining [26]. Although computer scientists have used string similarity to analyze a variety of sequences in poker and computer strategy games, such as [27] and [28], as far as we know, the similarity metrics has yet to be used in the differentiation and ranking of human performance.

B. Jaccard Index (JAC)

After reviewing several available string similarity metrics (http://wikipedia.org/wiki/String_Metric), we determined the best metric to be the Jaccard Similarity Coefficient [29], which we will report in this study. The Jaccard Similarity Coefficient (or Jaccard Index, JAC) can be used to measure the similarity between two sample sets by dividing the size of their intersection by the size of their union:

$$JAC(A, B) = |A \cap B| / |A \cup B| \quad (1)$$

Because the JAC value for two identical strings is 1 and for two completely different strings is 0, the values are easily understood by nonprofessionals:

$$0 \leq \text{Jaccard Index} \leq 1$$

(0% ⟷ 100% similarity)

To calculate the similarity between two strings, *A* and *B*, the strings must first be converted into bigrams. The bigrams for string *A* = 12345 are {12, 23, 34, 45}, and for string *B* = 13452 are {13, 34, 45, 52}. The intersection of the bigrams of *A* and *B* is {34, 45} and the union is {12, 23, 34, 45, 13, 52}. Therefore, $|A \cap B| = 2$ and $|A \cup B| = 6$. Thus, JAC is calculated to be 2 / 6 = 0.333.

IV. Methodology

A. Materials and Participants

The serious game used in this study is a military-oriented game with a *search and retrieve* mission comprised of seven checkpoints. From a designer's viewpoint, this is a nonlinear game because players may visit the checkpoints in any order they choose. From a player's viewpoint, upon entering the game, a mission-giver located at checkpoint 1 will brief the player with the mission to *search and retrieve* five villagers (checkpoints 2-6) who are hiding from potential enemies and a blacksmith (checkpoint 7) living at the far end of a forest. Once checkpoints 2-6 have been visited, the villagers will return to their 'home' and the checkpoints will disappear from the game environment.

Depending on the player's *course of actions* (i.e., order of checkpoints visited) in accomplishing the mission, an 8-digit string (action-sequence) can be obtained for each individual: e.g., 12345671, 13456271, etc. After completing the mission, the player must report to the mission-giver (at checkpoint 1) to be debriefed and 'end' the game. An action-sequence of 13456721 would indicate that a player visited checkpoints 1, 3, 4, 5, 6, 7, 2, and then back to 1 (in that order). All action-

sequences began and ended with 1 because checkpoint 1 was the mission-giver, who briefed and debriefed the players about the game mission at the beginning and the end of the game.

B. Data Collection and Visualization Tool

Because no one else knows the game better than the designer and the testing team, they served as Experts in this study and produced six sets of Expert data. Thirty-one students from a university contributed to the project as participants (with unknown performance). None of them had any prior experience with the game. User-generated data, which include coordinates of movement, time-stamps, number of villagers retrieved, and specific game events, were traced at regular intervals using an *in situ* data collection method called, *Information Trails* [2]. The collected user data were viewed in real-time using the data visualization tool: Performance Tracing Report Assistant (*PeTRA*) [7]. A total of 104233 data points were collected in this study.

V. Data Analysis

According to the Expert team, the ideal course of actions for the mission is 12345671. We calculated the JAC values for all participants and ranked them by JAC to determine their (dis-)similarity with the expert's (model) answer. Only one player (out of 31) achieved the ideal of JAC = 1, followed by the next best player at JAC = 0.57. This means the player with the highest possible JAC value (1) has attained the Expert rank, while the one with JAC = 0.57 can be regarded as a Likely-Expert (see Table 1). Beyond this, the JAC values fell quickly below 0.5 towards 0 – a clear indication that the players were not familiar with the game at all and performed poorly (i.e., low competency, which is to be expected).

Interestingly, participants who identified themselves as avid game players did not automatically score high on JAC. Although the player who achieved Expert rank has never played this game before, she did have a prior game design experience – which might explain her competency in problem solving using serious game.

TABLE I. Player –Ranking by Jaccard (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

A. Jackknife Reclassification Success Rate

The point-biserial correlation coefficient between JAC and the Expert/Player dummy variable (where Experts were coded with 1s and Players as 0s) indicated a strong, positive and significant linear relationship ($r_{pb} = 0.839$, $p < 0.01$). To evaluate the classification accuracy using JAC, a discriminant analysis with jackknife reclassification [30] (also known as leave-one-out cross-validation) success rate was carried out, which is 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 in reclassification, misclassifying only 2.7% (1 player) out of the total 37 observations.

The success rate of 97.3% was significantly better than the 50% expected by chance ($p < 0.001$). This is further investigated using simulated data following [31]. In particular, we used the sample means and standard deviations in the original data as estimates of the true parameters for experts and players, then simulated their Jaccard values as random draws from normal distributions with these parameters. Given this, two new data sets were generated for each of the 500 replications: a training data set with sample sizes being the same as those in the original data, and a validation data set with 60 experts and 310 players. With uniform priors, the discriminant functions developed based on the training data set correctly assigned an average of 97.48% (with a SD of 0.98%) of the subjects in the validation data set. This matches the jackknife reclassification success rate of 97.3% almost exactly, and further supports that the reclassification using JAC is better than expected by chance.

B. New Performance Metric for SEGA

This study represents the first step towards defining new performance metrics that are suitable for use as SEGA. We have shown that string similarity is a viable means to empirically quantify the degree of (dis)similarity between experts' and novices' course of actions (or competency) in problem solving within a serious game environment. We collected the in-game actions of a group of participants with unknown performance and successfully compared their action-data to the action-data of the experts.

We were able to obtain an 'Expert Similarity Index' and express empirically the similarity distance between a player and the expert. Converting the (dis)similarity of novices' and experts' actions into an index is a necessary step because it facilitates the ranking of the players according to how similarly their performances are to the experts.

C. Contribution

We have shown that it is possible to use JAC to measure the similarity between player action-sequence and that of the expert. Specifically, we were able to rank successfully the players' performance (to facilitate the selection of best/better players) without the need to collect more data (e.g., demographics). The ability to identify expert-like players out of a crowd of players with unknown performance without prior knowledge about their game-playing history or achievements is no small feat; it is also of possible value to the online gamer profiling industry.

We submit that JAC is much more robust than the 'best time' metric. In one case, a player was 'booted' from the online game due to technical issue and was forced to end the game prematurely. Based on the criteria for 'best time,' this player

would rank top on the leaderboard because s/he would appeared to have completed the game in record time! However, since the action-sequence of the players was {13} (JAC = 0), we knew immediately that s/he had failed to complete the game. The low JAC value also indicated his/her dis-similarity to the Experts, although this is not guaranteed. An additional advantage of JAC is that it is rather tolerant towards incomplete data, and thus, incurs little wastage from a data-mining point of view.

A player's in-game actions can be inferred to as a function of the user's understanding of the learning problems (what they know), problem-solving skills (what they are able to do), and decision-making strategy (how they go about doing it). Although time pressure can be useful in injecting fun, competitiveness, and motivation into the learners, in many learning situations, it can be detrimental to learning and performance assessment. A learner's competency should first and foremost reflect their metacognitions, decision-making processes, and strategies in dealing with the problems at hand; this is represented by their *course of actions* in problem solving.

Even though this study made use of a military-oriented mission, the string similarity metric approach detailed here is very user-friendly for performance assessment in non-military and business training settings. The similarity index introduced in this study is an important breakthrough because it converts players' competency in problem solving into an easily-comprehensible index.

D. Future Research

In a follow-up study, we will compare the effectiveness of string similarity index against commonly used game metrics to determine its usefulness as a performance metric for SEGA in differentiating (likely-)experts from novices, or play-learners who are highly competent from the less competent.

VI. Conclusion

Game designers and researchers are turning to capturing in-game actions of gamers as game analytics for monetization [9], and game-play experience improvement (usability) [32, 33]. However, no one has taken advantage of these user-generated data for performance assessment of play-learners in serious games, until now.

Although there are already a number of common metrics for comparing (entertainment) gamer performance such as best time, number of kills, number of game objectives met, etc. it is unclear if these metrics are suitable for use as performance metrics in serious games. Despite their direct application and possible transfer into military training games, very few of them are appropriate for serious games that deal with business training and traditional classroom learning. It is doubtful that any manager would agree to the view of 'clients as enemies' in business training, or for parents to approve of the tallying of kills as high-scores for classroom learning!

It has been suggested that a data-driven approach [34] and an evidence-centered design [3, 33] are much better assessment methods that will foster real adoption of serious games [36], [37]. Findings in this study suggest string similarity to be a viable performance assessment metric for serious games. We hope this will further lead to the development and establishment of newer and better metrics for SEGA in the future.

Acknowledgment

The authors wished to extend our thanks to Professor Emerita Sharon Shrock for providing valuable feedback to an earlier draft of this paper, and our doctoral students, T. Zhou and I. H. Li, for their assistance in data collection.

References

[1] F. Bellotti, B. Kapralos, K. Lee, P. Moreno-Ger, and R. Berta, "Assessment in and of Serious Games: An overview," *Advances in Human-Computer Interaction*, vol. 2013, p. 11, 2013.

[2] C. S. Loh, A. Anantachai, J. H. Byun, and J. Lenox, "Assessing what players learned in serious games: In situ data collection, Information Trails, and quantitative analysis," in *Proceedings of the Computer Games: AI, Animation, Mobile, Educational & Serious Games Conference (CGAMES 2007)*, 2007.

[3] V. J. Shute, I. Masduki, O. Donmez, V. P. Dennen, Y.-J. Kim, A. C. Jeong, and C.-Y. Wang, "Modeling, Assessing, and Supporting Key Competencies Within Game Environments," in *Computer-Based Diagnostics and Systematic Analysis of Knowledge (Part 4)*, D. Ifenthaler, P. Pirnay-Dummer, and N. M. Seel, Eds. Boston, MA: Springer US, 2010, pp. 281–309.

[4] D. Michael and S. Chen, *Serious games: Games that educate, train, and inform.* Boston, MA: Thomson Course Technology PTR., 2006.

[5] R. Thawonmas and K. Iizuka, "Visualization of online-game players based on their action behaviors," *International Journal of Computer Games Technology*, vol. 2008, pp. 1–9, 2008.

[6] G. Wallner, "Play-Graph: A methodology and visualization approach for the analysis of gameplay data," in *8th International Conference on the Foundations of Digital Games (FDG 2013)*, 2013, pp. 253–260.

[7] C. S. Loh, "Information Trails: In-process assessment of game-based learning," in *Assessment in Game-based Learning: Foundations, Innovations, and Perspectives*, D. Ifenthaler, D. Eseryel, and X. Ge, Eds. New York, NY: Springer, 2012, pp. 123–144.

[8] C. S. Loh, "Using in situ data collection to improve the impact and return of investment of game-based learning," in *Old Meets New: Media in Education – Proceedings of the 61st International Council for Educational Media and the XIII International Symposium on Computers in Education (ICEM&SIIE'2011) Joint Conference*, 2011, pp. 801–811.

[9] A. Canossa, M. Seif El-Nasr, and A. Drachen, "Benefits of game analytics: Stakeholders, contexts and domains," in *Game analytics: Maximizing the value of player data*, M. Seif El-Nasr, A. Drachen, and A. Canossa, Eds. London, UK: Springer-Verlag, 2013, pp. 41–52.

[10] H. Ben Zur and S. J. Breznitz, "The effect of time pressure on risky choice behavior," *Acta Psychologica*, vol. 47, no. 2, pp. 89–104, Feb. 1981.

[11] R. Pieters and L. Warlop, "Visual attention during brand choice: The impact of time pressure and task motivation," *International Journal of Research in Marketing*, vol. 16, no. 1, pp. 1–16, Feb. 1999.

[12] M. E. Young, S. C. Sutherland, and J. J. Cole, "Individual differences in causal judgment under time pressure: Sex and prior video game experience as predictors," *International Journal of Comparative Psychology*, vol. 24, no. 1, pp. 76–98, 2011.

[13] C. D. Wickens, A. Stokes, B. Barnett, and F. Hyman, "The effects of stress on pilot judgment in a MIDIS Simulator," in *Time Pressure and Stress in Human Judgment and Decision Making*, O. Svenson and A. J. Maule, Eds. Boston, MA: Springer US, 1993, pp. 271–292.

[14] S. E. Dreyfus and H. L. Dreyfus, "A five-stage model of the mental activities involved in directed skill acquisition," Berkeley, CA, 1980.

[15] S. E. Dreyfus, "The five-stage model of adult skill acquisition," *Bulletin of Science, Technology and Society*, vol. 24, no. 3, pp. 177–181, Jun. 2004.

[16] J. Underwood, "Novice and expert performance with a dynamic control task: Scanpaths during a computer game," in *Cognitive processes in eye guidance*, G. Underwood, Ed. Oxford: Oxford University Press, 2005, pp. 303–323.

[17] A. Hofer, "Exploratory comparison of expert and novice pair programmers," in *Lecture Notes in Computer Science including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, 2011, vol. 4980 LNCS, pp. 218–231.

[18] B. Law, M. S. Atkins, A. E. Kirkpatrick, and A. J. Lomax, "Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment," in *Proceedings of the Eye Tracking Research & Applications Symposium - ETRA 2004*, 2004, pp. 41–48.

[19] A. M. Williams and P. R. Ford, "Expertise and expert performance in sport," *International Review of Sport and Exercise Psychology*, vol. 1, no. 1, pp. 4–18, 2008.

[20] W. Boot, A. F. Kramer, M. Fabiani, G. Gratton, D. J. Simons, X. I. Wan, M. S. Ambinder, L. E. Thomas, S. J. Colcombe, J. Agran, K. Low, and Y. Lee, "The effects of video game playing on perceptual and cognitive abilities," *Journal of Vision*, vol. 6, no. 6, p. 75, 2006.

[21] T. C. Chamberlain, V. M. Catano, and D. P. Cunningham, "Personality as a predictor of professional behavior in dental school: Comparisons with dental practitioners," *J Dent Educ.*, vol. 69, no. 11, pp. 1222–1237, Nov. 2005.

[22] P. Hughes, "Can we improve on how we select medical students?," *Journal of the Royal Society of Medicine*, vol. 95, no. 1, pp. 18–22, Jan. 2002.

[23] J. E. Rockoff, B. A. Jacob, T. J. Kane, and D. O. Staiger, "Can you recognize an effective teacher when you recruit one?" Cambridge, MA, p. 56, 13-Nov-2008.

[24] M. Bilenko and R. J. Mooney, "Adaptive duplicate detection using learnable string similarity measures," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, 2003, pp. 39–48.

[25] W. E. Winkler, "The state of record linkage and current research problems," Washington, DC, 1999.

[26] A. E. Monge and C. P. Elkan, "An efficient domain-independent algorithm for detecting approximately duplicate database records," in *Proceedings of the SIGMOD 1997 Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1997, pp. 23–29.

[27] J. Rubin and I. Watson, "Similarity-based retrieval and solution re-use policies in the game of Texas Hold'em," in *Case-Based Reasoning, Research and Development*, vol. 6176, I. Bichindaritz and S. Montani, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 465–479.

[28] S. J. Louis and C. Miles, "Playing to learn: case-injected genetic algorithms for learning to play computer games ," *Evolutionary Computation, IEEE Transactions on*, vol. 9, no. 6, pp. 669–681, 2005.

[29] C. S. Loh and Y. Sheng, "Measuring the (dis-)similarity between expert and novice behaviors as serious games analytics," *Education and Information Technologies.* (in press).

[30] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, no. 1, pp. 1–11, 1968.

[31] J. W. White and B. I. Ruttenberg, "Discriminant function analysis in marine ecology: Some oversights and their solutions.," *Marine Ecology Progress Series*, vol. 329, pp. 301–305, 2007.

[32] L. Nacke and C. A. Lindley, "Flow and Immersion in First-Person Shooters: Measuring the player's gameplay experience," in *Proceedings of the 2008 Conference on Future Play*, 2008, pp. 81–88.

[33] L. Nacke, C. Lindley, and S. Stellmach, "Log Who's Playing: Psychophysiological Game Analysis Made Easy through Event Logging," in *Fun and Games*, vol. 5294/2008, P. M. E. AL, Ed. Springer Berlin / Heidelberg, 2008, pp. 150–157.

[34] J. M. Thomas and M. E. DeRosier, "Toward effective game-based social skills tutoring for children: An evaluation of a social adventure game," in *Proceedings of the 5th International Conference on the Foundations of Digital Games (FDG '10)*, 2010, pp. 217–223.

[35] V. J. Shute and M. Ventura, *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA: MIT Press, 2013, p. 102.

[36] D. Ifenthaler, D. Eseryel, and X. Ge, *Assessment in Game-based Learning: Foundations, Innovations, and Perspectives*. New York, NY: Springer, 2012, p. 465.

[37] C. S. Loh, "Researching and developing serious games as interactive learning instructions," *International Journal of Gaming and Computer-Mediated Simulations*, vol. 1, no. 4, pp. 1–19, Jan. 2009.