

# Treasure Explorers - A Game as a Diagnostic Assessment Tool

Bernardo Pereira Nunes<sup>\*†</sup>, Terhi Nurmikko-Fuller<sup>§</sup>, Giseli Rabello Lopes<sup>‡</sup>,  
Sean W. M. Siqueira<sup>†</sup>, Gilda H. B. de Campos<sup>¶</sup> and Marco A. Casanova<sup>\*</sup>

<sup>\*</sup>Department of Informatics - PUC-Rio, Rio de Janeiro, RJ - Brazil  
{bnunes, casanova}@inf.puc-rio.br

<sup>†</sup>Department of Applied Informatics - UNIRIO, Rio de Janeiro, RJ - Brazil  
{bernardo.nunes, sean}@uniriotec.br

<sup>‡</sup>Federal University of Rio de Janeiro, Rio de Janeiro, RJ - Brazil  
giseli@dcc.ufrj.br

<sup>§</sup>Oxford e-Research Centre - University of Oxford, Oxford, United Kingdom  
terhi.nurmikko-fuller@oerc.ox.ac.uk

<sup>¶</sup>Department of Education - PUC-Rio, Rio de Janeiro, RJ - Brazil  
gilda@ccead.puc-rio.br

**Abstract**—Understanding students’ strengths and weaknesses can help in the design of teaching materials to successfully bridge identified gaps. Formal exams are useful to gauge the extent of learners’ memorised information, but have been critiqued for not reflecting the true extent of learners knowledge. Diagnostic assessment assists teachers in setting task-specific plans and goals, for both individual students, and the learner-group as a whole. In this paper, we describe *Treasure Explorers*, a Game With a Purpose (GWAP) with a multilayered structure that facilitates the learning process, promotes user retention, and is designed to reward contributing players. A comprehensive evaluation based on game log and TAM model was conducted. The log of over 5,500 records reveals that aspects of engagement and learning had occurred, and qualitative evaluations show the applicability and usability of games as diagnostic assessment tools. *Treasure Explorers* assists teachers in identifying student difficulties, and has significant potential to help educators plan for lesson content.

**Keywords**—Diagnostic assessment; serious games; gamification

## I. INTRODUCTION

Learning and teaching are complex endeavours [1], [2]. The need to evidence that desired learning outcomes have been reached necessitates the evaluation of education as an observable, quantifiable, and measurable phenomenon. This gives rise to standardised testing, and influences both the pedagogical process and the design of teaching materials.

Knowledge of the extent and type of students’ prior learning and abilities allows educators to set task-specific plans and goals in terms of the needs of individuals and the group as a whole. Awareness of the students’ strengths and weaknesses can aid this process and support the design of teaching materials to successfully bridge identified gaps.

Learners are assessed through frequent formalised testing, ranging from nationally acknowledged examinations to short classroom quizzes. But does exclusive reliance on formalised testing provide an accurate measure of all learners’ abilities? Alternative methods such as diagnostic assessment

(traditionally used to identify of students who might benefit from additional support e.g those with dyslexia), are known, but may be beyond the remit, scope, and access of some teachers. The restraints of time may dissuade others from additional, potentially extra-curricular assessments. Learners may be reluctant to participate (perceived) additional testing.

Standardised testing is designed to ensure all aspects of the examination process from the questions to the test environment and the scoring procedures are identical across the board, but some students do not perform to their full potential when formally assessed [3]. Existing informal approaches are usually conducted in a casual manner with no assignment of individual scores [4]. Mostly qualitative, they are based on observations, discussions, interviews, and portfolios, and can be demanding on the educator’s time.

To help circumvent the problem, we propose the use of a Game with a Purpose (GWAP) as an alternative, additional, complementary method for informal assessment. *Treasure Explorers* is an interactive learning environment that can be used to perform diagnostic assessments playfully and informally. Although gamification in education is not a new concept, *Treasure Explorers* is unique in combining diagnostic assessment and games to assist teachers to identify patterns in learners’ prior knowledge, skills and abilities. It supports the development and delivery of appropriate learning plans to reach specific, desired teaching goals.

*Treasure Explorers* brings diverse research agendas together, and benefits from the synergy of interdisciplinary perspectives and approaches. It consists of layers of riddles and sub-games that players solve and create to gain points. The initial design and implementation were inspired by the successes of citizen science projects such as Duolingo<sup>1</sup> and Zooniverse<sup>2</sup>. *Treasure Explorers* incorporates reward

<sup>1</sup><https://www.duolingo.com>

<sup>2</sup><https://www.zooniverse.org>

schemes and links to social media to encourage initial take up and increase participant retention rates. Our aim was to establish the diagnostic potential of games, which diversity (and complement) formalised education and assessment. We hypothesise that GWAPs help in the assessment of students' levels of knowledge, and remove barriers to learning.

In this paper, we describe *Treasure Explorers*. We summarise findings from diagnostic assessments, evaluate obtained results, and conclude with a view to future work.

## II. THE GAME

*Treasure Explorers* consists of layers of riddles and sub-games. When a player successfully solves one, (s)he gains points and advances to the next level. If unable to do so, the player can acquire clues released through various associated sub-games, of which there are three types: the *Quiz*, the *Image Tagger* and the *Multiple Connector*. If the correct answer is entered within the allocated time and number of available tries, a clue to the original riddle is released. The length of time and level of difficulty are assigned by the creator of the sub-game. Top players can create new riddles.

The *Quiz* has a multiple-choice layout, where the creator specifies the number of available attempts a player has for identifying the correct answer. This sub-game requires prior knowledge of specialist subjects or general trivia.

In the *Image Tagger*, a player labels a single or group of images. These may be prompted by questions, or simply be the most basic of visual descriptions (items, colours), depending on the design. It can be set up to require very little prior knowledge of any given subject, and with limited time available, it is the most fast-paced of the sub-games.

The *Multiple Connector* is an opportunity to link diverse digital resources, such as images to text. This is the most demanding of the creator in the game design phase. Requiring a degree of interlinking between different datasets, it has the most potential for creative approaches to information assessment and discovery. Of the three sub-games, it is most suitable not only for the repetition of learnt facts, but also for the analysis and interpretation of those facts to correctly match and link between disparate pieces of information.

*Treasure Explorers* has a simple model for user-engagement. The aim was to facilitate and encourage training, with a conscious effort to cater for different learning styles, and to lower barriers to engagement. There is a simple point-and-click user-interface, and at most, players need to type in answers. The required levels of computer literacy are low, and the game is accessible to anyone on the Web.

Players log in to *Treasure Explorers* with their Facebook profiles. This approach allows participants to discuss and promote the game within their social network, encourages other players to join, and the playfully competitive ranking element enables friendly competition between players. It provides us with the opportunity to collate demographic

Table I  
PLAYERS OF TREASURE EXPLORERS

Age Group	Players								
	Students			Young Teachers			Senior Teachers		
	under 19 y.o.			19 < x ≤ 30			above 30 y.o.		
Gender	M	F	Both	M	F	Both	M	F	Both
Total	10	7	17	2	9	11	7	15	22

data, with the caveat that this is based on users' self-description, and on such a social and informal platform, it might not be (deliberately or accidentally) entirely accurate.

## III. EVALUATION SETUP

Player engagement was used to evaluate the suitability of *Treasure Explorers* as a diagnostic assessment tool. There are three categories of players: students, young teachers, and senior teachers (Table I). Both sets of teachers were encouraged to use the game in the classroom in the upcoming semester: this was found to increase their interest in playing it. Engagement (as with the students) was determined by examination of the activity log. A complementary evaluation, based on the Technology Acceptance Model (TAM) proposed by Davis [5] was conducted to verify acceptance of the game amongst teachers, their plans to for future use, and the role of games in diagnostic assessment.

Our evaluation focused on games created by the authors in accordance to the Brazilian National Educational Plan: mostly logic and language-based questions. At the time of the evaluation, the players ranged from 15 to 64 years-old. Students were not informed they were under assessment, nor was playing the game a compulsory activity, ensuring genuine engagement (based solely on the log<sup>3</sup>) was measured.

### A. Log-based Evaluation

A total of 50 players were observed for the duration of one calendar month. The majority belonged to the approached focus groups, but additional players were attracted through links shared by existing one via their social media networks.

Post-login actions (e.g. page clicks, etc.) are stored in a log. In this period, 5,556 actions were recorded: 2,999 of them were related to playing. These activities were divided into five categories: *Game Access*, *Leaderboard*, *Sharing Behaviour*, *Attempts Analysis* and *User Engagement*.

### B. Technology Acceptance Model Evaluation

*Treasure Explorers* was evaluated qualitatively. The main objectives were verifying the role it could play in academic activity, and assessing its adoption-rate as a tool by teachers. A questionnaire to capture *perceived usefulness* (PU) and *perceived ease-of-use* (PEOU) was created based on the TAM [5]. Opinion-mining questions were also included.

<sup>3</sup>The log records all player interactions with the game.

Table II  
PAGE VISITS BY GROUP AND GENDER

Players	Page Hits		
	Male	Female	Total
Students	1505	1818	3323
Young Teachers	231	927	1158
Senior Teachers	332	743	1075
<b>Total</b>	2068	3488	5556

PU maps “the degree to which a person believes that using a particular system would enhance his or her job performance”, whereas PEOU assesses “the degree to which a person believes that using a particular system would be free of effort” [5]. The questionnaire, consisting of seven PU questions and five PEOU questions (each associated with a 5-point Likert scale of agreement) was sent to the teachers. In total, 11 teachers answered the questionnaire anonymously.

#### IV. RESULTS

##### A. Log-based Evaluation Results

**Game Access.** Numbers of page hits per group and gender are shown in Table II. Students were more active than the other two groups combined, and the game was more popular among females (63%). On average, players navigated to 106 pages (there was large standard deviation of  $\sigma=155.4$ ).

**Leaderboard.** A leaderboard was used as the basis for analyses of access rates, and the identification of the top player demographics. The top five players were found to include one player from each of the young and old teachers groups, and three players from the students group.

The leaderboard was quite popular amongst the players who were responsible for 9% (235) of the total number of hits to pages (excluding sub-games and riddles). Students were the most active group, scoring fewer points per game on average, but playing a greater number of games and for longer periods of time, achieving high scores. This illustrates the success of the game in engaging and retaining players.

**Sharing Behaviour.** Players can publish their results on social media. Such updates are possible only for the riddles, not each individual sub-game. When a player shares their score, a message is displayed: “%Player name% has just solved the riddle ‘What goes around the world but stays in one corner?’. (S)He has just got 50 points”. A total of 50 riddles were shared via Facebook. Again, students were most likely to publish their results (68% of shared results).

**Attempts Analysis.** From an educational perspective, the number of attempts needed for a player to solve a game is not as relevant as the learning outcome. According to Piaget [6], learning occurs through two cognitive processes:

assimilation and accommodation. Initial attempts at a new task involving previously unfamiliar material often result in error, but as learning occurs, success becomes more likely. From an educational perspective, repetition is key. From the perspective of the game, players are more likely to persevere if rewarded for their success. The point of intersection between the two is that a task must be complex enough to be challenging, but not too difficult so as to be demoralising.

*Treasure Explorers* has an option to limit the number of attempts a player can have. These are generally related to the number of possible answers, i.e., a quiz with four options should allow no more than four attempts. For each wrong answer given, the number of received points decreases.

For the *Quiz*, no significant gender-based differences were observed in the number of attempts needed. Players with more experience were quicker to solve the quizzes (1.16 and 1.19 times for senior and young teachers respectively, c.f. the students who on average took 1.27 attempts).

The *Tagger* and *Multiple Connector* were more challenging. Female student players of the *Multiple Connector* outperformed their male counterparts (2.00 attempts, c.f. 3.72 ). Students needed a greater number of attempts than either the young or senior teachers (3.23, 2.09 and 1.95 respectively) for the *Multiple Connector*. 3.13, 2.08, and 2.19 attempts were needed respectively for the *Image Tagger*: female players on average used 2.88 attempts, male players 2.08.

The analysis was repeated on the riddles. Students, young and senior teachers used 573, 2.52 and 1.51 (with a standard deviation of 4.17, 3.6 and 0.8) attempts respectively.

Table III summarises the results of the attempts analysis per game, group and gender.

Table III  
AVERAGE ATTEMPTS PER GAME

Players*	Attempts								
	Quiz			Multiple			Tagger		
	M	F	Both	M	F	Both	M	F	Both
Students	1.28	1.25	1.27	3.72	2.00	3.23	2.15	4.83	3.13
Young Teachers	1.08	1.23	1.19	1.80	2.17	2.09	2.00	2.11	2.08
Senior Teachers	1.24	1.08	1.16	2.70	1.27	1.95	2.60	1.90	2.19

**User Engagement.** Target audience engagement was evaluated as an indicator of the viability of future development to support other domains. There was little behavioural difference between the genders: male players had 2,984 page views while female players accessed 2,576 pages. Students (3,323 page views) were three times more engaged than the young and senior teachers (1,158 and 1,079 respectively).

##### B. Questionnaire Results

As described in section 5.2, a questionnaire divided into three categories: PU, PEOU, and opinion mining questions was circulated. The average agreement of the PU was 3.98 ( $\sigma=0.29$ ) out of 5 points on a Likert scale with a Cronbach’s

alpha of 0.92; the PEOU had an average of 4.06 ( $\sigma=0.36$ ) with a Cronbach's alpha of 0.73. The coefficient of internal consistency Cronbach's alpha greater than 0.70 indicated a high reliability of the results. The results for the TAM point to the usability of *Treasure Explorers* as a tool for diagnostic assessment and its acceptance by teachers.

The opinion mining questions contained both open and multiple-choice questions following a 5-point Likert scale of agreement. The first captured the enjoyment felt by the player, and for the exception of two respondents, all teachers awarded the game 4 or 5. This resulted in an average of 4 ( $\sigma=1.34$ ). Another question assessed whether players felt under assessment. An average of 1.72 ( $\sigma=0.78$ ) shows that the game was not perceived to be a formal assessment.

The question to receive the highest grade was "This game can be used as a complementary educational resource to the traditional methods of education" with an average of agreement of 4.5 ( $\sigma=0.68$ ). Another suggesting the replacement of traditional methods with playful approaches had the lowest average (3.72), but the highest deviation ( $\sigma=1.34$ ).

Questions related to the identification of student difficulties obtained positive feedback (4.09 and  $\sigma=0.94$ ). Teachers could discover the challenges and problems faced by learners by examining the answers recorded in the logs, and reported that the game could be useful in assisting in the preparation for their classes (4.18 and  $\sigma=0.87$ ). A strong belief (4.27  $\sigma=0.9$ ) that this approach would make students engage more with relevant information was reported. Many planned to use the game in the future (average 3.45 and  $\sigma=1.57$ ).

Five respondents gave feedback using the open question. Comments referred to the score system, particularly in that teachers were expecting the score to be used later, for either student assessment or as a record of the experience. One teacher stated that this was the first experience her students had with this kind of game, and that they enjoyed it, having played it without knowing that it would be used as a tool for evaluating performance or prior knowledge.

## V. RELATED WORK

A memorandum from the Canadian Ministry of Education [7] discusses the importance of diagnostic assessment in support of student learning, and their role in assisting teachers and school staff to identify pedagogical needs, monitor progress and set individual or group learning goals. Suitable assessment tools help teachers reduce the gap in student achievement and meet the criteria outlined in the curriculum.

Shute *et al.* [8] examine the benefits and challenges of summative and formative assessment, proposing stealth assessment<sup>4</sup> as a method for determining learners' skills during active learning. It is similar to our approach as both methodologies aim to find evidence of learning and achievement during highly interactive and immersive tasks.

<sup>4</sup>This is a type of formative assessment.

Buzzetto-More *et al.* [9] present several assessment methods and case studies that aim to satisfy learning and teaching goals. The research agenda is one of identifying methods that can be used to measure myriad different skills. Our primary focus has been to assess students' strengths and weaknesses in solving written riddles (requiring existing knowledge, as well as reading comprehension and writing).

GWAPs have been proposed as a way to solve hidden tasks or those where entertainment is not the main focus of the activity. Von Ahn [10] suggests different strategies to solve large-scale computational problems through games, an example being the ESP game, where players label images based on a guess as to the word the other player used to describe the same image. *Treasure Explorers* shares some similarity with the ESP game with the aim of identifying students' strengths and weaknesses to further set learning plans and learning goals individually and collectively.

Games have been used as a motivational tool for learning [11]. This branch of games, called Serious Games or GWAPs, aims at entertaining while educating. They are often used as complementary learning material to stimulate interest and engage students to achieve specific learning goals, knowledge and skills. Derbali and Frasson [12] discuss the importance of motivation and student engagement as a natural part of any learning process. They claim motivated students are more likely to perform challenging activities with better performance. Bellotti *et al.* [13] discuss the educational effectiveness of GWAPs, drawing attention to the assessment of player's performance as a prerequisite for a successful learning experience. They argue that GWAP must have well-defined goals and elaborated techniques to collect learners' performance while playing, a perspective endorsed by Guillén-Nieto and Aleson-Carbonell [14]. These approaches show the importance of games in motivating students to learn specific concepts, in achieving a set of learning goals, and in retaining users [15]: following these guidelines, *Treasure Explorers* logs all information that can be used *a posteriori* to assist teachers on setting specific learning plans.

## VI. DISCUSSION AND OUTLOOK

*Treasure Explorers* is a multilayered GWAP, designed to support the processes of education and learning evaluation in an informal learning environment. It includes a number of choices and feedback mechanics designed to motivate players, and the riddle and sub-game structure is purposely designed to reward players' curiosity and encourage them to explore the game further. As players begin to understand what is required and the how the pattern of the game repeats itself, their self-esteem and desire to complete all the clues grows. Points, bonus scores, and levels all encourage the feeling of *fiero*, "the personal feeling of triumph over adversity" [16]. Finally the opportunity to share achievements stimulates emotions responses varying from amusement and

inspiration to envy. *Treasure Explorers* can be shown to contain elements that address each of the “Four Fun Keys” which Lazarro [16] says every game should strive for: *Hard Fun, Easy Fun, Serious Fun* and *People Fun*.

The leaderboard and user evaluations illustrate a high level of student engagement, and support the notion that students are very motivated to learn through enjoyable experiences such as games or other informal learning environments.

Player behaviour demonstrates the desire to and practice of promoting and sharing their success with others.

The results indicate a gender division between the players, with female players being more involved than their male counterparts. Different genders were also seen to excel in solving different types of questions. This interesting discovery could help support teachers in preparing for classes and in the selection of educational resources more closely tailored to the idiosyncratic needs of individuals.

The attempts analysis illustrated the games’ relative complexities. Since the aim is to facilitate learning, rather than scoring, observed increases and decreases in the attempts rate become an evaluation tool. Further testing and analyses across a greater dataset are needed to establish whether there are patterns to question or sub-game type, player demographics, or other elements.

Analysis using the TAM model has shown that *Treasure Explorers* was accepted and approved of by teachers as a tool to support diagnostic assessment. There was extensive agreement on the suitability of the game to assist in the identification of student difficulties, and its potential to help plan lesson content. Evaluation by the players recorded an ease of playing, and a low barrier to engagement.

Future development of *Treasure Explorers* will see wider dissemination, and the generation of new content. A greater number of observable patterns in the relationship between learning styles and riddle types is likely to emerge.

*Treasure Explorers* can be accessed at: <http://research.ccead.puc-rio.br/treasureExplorers/>.

#### ACKNOWLEDGMENT

Partly funded by CNPq and FAPERJ under grant 444976/2014-0 and E-26-102.256/2013, respectively.

#### REFERENCES

- [1] D. V. DAY, “The difficulties of learning from experience and the need for deliberate practice,” *Industrial and Organizational Psychology*, vol. 3, no. 1, pp. 41–44, 2010. [Online]. Available: <http://dx.doi.org/10.1111/j.1754-9434.2009.01195.x>
- [2] J. S. Brown, A. Collins, and P. Duguid, “Situated learning and the culture of learning,” *Education Researcher*, vol. 18, no. 1, pp. 32–42, 1989.
- [3] D. Boud, R. Cohen, and J. Sampson, “Peer learning and assessment,” *Assessment & Evaluation in Higher Education*, vol. 24, no. 4, pp. 413–426, 1999. [Online]. Available: <http://dx.doi.org/10.1080/0260293990240405>
- [4] C. A. Christie and M. Rose, “Learning about evaluation through dialogue: Lessons from an informal discussion group,” *American Journal of Evaluation*, vol. 24, no. 2, pp. 235–243, 2003. [Online]. Available: <http://aje.sagepub.com/content/24/2/235.abstract>
- [5] F. D. Davis, “Perceived usefulness, perceived ease of use, and user acceptance of information technology,” *MIS quarterly*, pp. 319–340, 1989.
- [6] J. Piaget, “Piaget’s theory,” in *Piaget and His School*, ser. Springer Study Edition, B. Inhelder, H. Chipman, and C. Zwingmann, Eds. Springer, 1976, pp. 11–23. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-46323-5\\_2](http://dx.doi.org/10.1007/978-3-642-46323-5_2)
- [7] Ontario Ministry of Education, “Diagnostic assessment in support of student learning,” in *Policy/Program Memorandum No. 155*. Ontario Ministry of Education, January 2013, pp. 1–5.
- [8] V. Shute and Y. Kim, “Formative and stealth assessment,” in *Handbook of Research on Educational Communications and Technology*, J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop, Eds. Springer, 2014, pp. 311–321. [Online]. Available: [http://dx.doi.org/10.1007/978-1-4614-3185-5\\_25](http://dx.doi.org/10.1007/978-1-4614-3185-5_25)
- [9] N. A. Buzzetto-More and A. J. Alade, “Best practices in e-assessment,” *Journal of Information Technology Education*, vol. 5, pp. 251–269, 2006.
- [10] L. von Ahn and L. Dabbish, “Designing games with a purpose,” *Commun. ACM*, vol. 51, no. 8, pp. 58–67, Aug. 2008. [Online]. Available: <http://doi.acm.org/10.1145/1378704.1378719>
- [11] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, “A systematic literature review of empirical evidence on computer games and serious games,” *Computers & Education*, vol. 59, no. 2, pp. 661–686, Sep. 2012. [Online]. Available: <http://dx.doi.org/10.1016/j.compedu.2012.03.004>
- [12] L. Derbali and C. Frasson, “Assessment of learners’ motivation during interactions with serious games: A study of some motivational strategies in food-force,” *Adv. in Hum.-Comp. Int.*, vol. 2012, pp. 5:5–5:5, Jan. 2012. [Online]. Available: <http://dx.doi.org/10.1155/2012/624538>
- [13] 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, 2013.
- [14] V. Guillén-Nieto and M. Aleson-Carbonell, “Serious games and learning effectiveness: The case of it’s a deal!” *Computers & Education*, vol. 58, no. 1, pp. 435–448, Jan. 2012. [Online]. Available: <http://dx.doi.org/10.1016/j.compedu.2011.07.015>
- [15] B. B. Morrison and J. A. Preston, “Engagement: Gaming throughout the curriculum,” in *Proceedings of the 40th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE ’09. New York, NY, USA: ACM, 2009, pp. 342–346. [Online]. Available: <http://doi.acm.org/10.1145/1508865.1508990>
- [16] N. Lazarro, *Understand Emotions*. Charles River Media, 2009.