Assessing Verbal Creativity with Digital Educational Games

Stephen Howell and Tony Veale
stephen.howell, tony.veale {@ucd.ie}
School of Computer Science and Informatics
University College Dublin
Dublin, Ireland
Abstract

There is growing interest in utilising digital educational games for assessment of traditional classroom learning, but the challenge of adapting the game difficulty to the individual ability of the player remains. A critical aspect of successful game design is to ensure that the player is placed in a state of flow, such that they are neither frustrated by the difficulty of the game nor bored by its simplicity. In digital educational games, this means that the difficulty of the game must be tailored to the individual learner’s ability. This paper presents an Intelligent Tutoring System which features a digital educational game that assesses the player’s verbal creativity, specifically in the area of word association skills. By designing the game as a casual game (a casual game employs common game modalities such as simple control systems and immediate feedback), there is a reduced learning curve to the game play. The Intelligent Tutoring System implements dynamic difficulty adjustment and allows a wide range of players with different verbal abilities enjoy the game. To examine the effectiveness of this assessment strategy, a pilot experiment was performed where third level computing students undertook a traditional verbal creativity test (Remote Associates Test) and then played the digital learning game. This paper examines the pertinent game design issues and presents the findings from the pilot experiment.

Keywords

Digital Educational Games, Intelligent Tutoring Systems, Dynamic Difficulty Adjustment, Verbal Creativity
1. Introduction and Motivation

Digital games are a popular pastime with many students and a growing activity in many market segments. Commercial games are marketed to male and female, young and old alike, some claiming to improve health and fitness or mental acuity while entertaining the player (Buttussi & Chittaro, 2010, Dixon et al., 2010). Increasingly, educators are exploring commercial games in the classroom to encourage student interest and participation (Rosas et al., 2003). Games which are designed and marketed as having a positive health or educational effect are dubbed serious games (Zyda, 2005). There is growing interest in academic research to measure the effects of such games.

There are several challenges to developing games for students. Technical challenges include the large and costly development and design teams required to produce them (Torrente et al., 2010). Some games are more successful than others, and some researchers theorise that this is because the more successful games place the player in a ‘state of flow’ (Csikszentmihalyi, 1991). Unfortunately, it is difficult to design a game to place a player in a state of flow. Furthermore, having developed the game, the educator is faced with the problem of ensuring the pedagogical material is complete and pertinent. Mandated syllabus changes may render this difficult work outdated, making the game less effective or useless. This leads to the idea that a means to algorithmically generate the data for the educational game would save considerable development time and effort. However, a single assessment will not suit all students, and while a summative assessment for a class could be a ‘single use’ assessment that all students must take, a formative assessment that could analyse a student’s incremental progress would require a test that is dynamically adjusted so that the difficulty suits the student currently taking the test.

This paper describes the process of play testing a serious game, BuzzWords, to assess verbal-creativity. The issues outlined above, flow states, algorithmic data generation and dynamic difficulty adjustment are all addressed and implemented for BuzzWords. BuzzWords employs a significant assessment database which was largely algorithmically generated. BuzzWords uses game play mechanisms designed to place the player in a state of flow. In order to ensure that BuzzWords was academically sound, it was implemented as an Intelligent Tutoring System with dynamic difficulty
adjustment. This paper focuses on the play testing and assessment aspects of BuzzWords.

2. Serious Casual Games

Serious games are educational games that are still fun to play. Traditionally, they can take many forms, from simulation and strategy games to adventure games. The attendant costs of creating a game mean that serious games are just as costly as entertainment market games to produce. Recently there has been growth in casual games, games that have relatively no plot or characterisation but are simple and fun to play (Fulton & Fulton, 2010). Creating a casual game is comparatively cheap yet they can be significant revenue generators. Possibly the most well known casual game is Tetris. Simple game play with only 3 possible moves, minimalist user interface and infinite game play combinations make Tetris a highly playable and popular game (Jordan, 2009). Players enjoy a short learning curve, yet the game proves engaging as players replay continually to improve their previous score. Casual games are possibly unique in that publishers claim that the majority of players are female. PopCap Games, publisher of Bejewelled and other games also claims greater numbers of age fifty-plus players (Schiesel, 2007).

Causal games feature a very simple learning curve, repeated ‘try-again’ game play and simple, easy to use interfaces. These features make them ideal for educational games, as the student player does not need lengthy instruction on how to play, and the more times they play, the more detailed pedagogical data can be generated. In essence, because they are enjoying the experience, the student is volunteering to be assessed repeatedly. Many educators would welcome the scenario where students are asking to be tested.

BuzzWords is a casual game; as game play required word matching to assess verbal creativity, a narrative game was unnecessary. A typical game of BuzzWords involves a short introduction that informs the player as to the goals and basic game play techniques of the game. Once the game begins, they are presented with a selection of eight words in blocks at the bottom of the game window. A falling word must be positioned by the player so that it will connect with one of the existing words. Somewhere amongst these words on the bottom row will be some that match the falling word in a particular way. In BuzzWords, they match to form compound terms.
Compound terms are made from two words which may be combined to form a new term such as ‘dining room’ or ‘race car’. There are many thousand compound terms in the English language. BuzzWords uses a scoring mechanism that requires the player to make the best compound match that they can from the falling word and all the resting words. There can be several matches, but only one will be the best or rarest match. Over time the difficulty (or rarity) of the words increases or decreases based on the player’s performance. There is an element of strategy in the game, as the player is required to vertically match words but should they land on a word which does not make a valid match, the in-play word will simply rest where it lands. In future rounds these words become a hindrance as they will hide a word underneath it that the player may later need. Furthermore, if the player makes a horizontal match accidentally, the words lock together and this may also cause a hindrance. Should the player make any valid vertical match, both words in the match will flash and then explode, with a score floating in their place. The score is directly related to the rarity of the matched words with rarer compounds scoring higher than common compounds.

Figure 1 BuzzWords in play
2.1 Engaging Games & Flow

It would be inadvisable to develop a serious game and simply hope it is engaging. Design considerations are needed to ensure that the game has every chance of being engaging and fun. The game’s educational content is still of utmost importance, but if the game is not fun to play, it will fail to assess or entertain the student. This section outlines the research into game and psychological engagement which formed the basis for BuzzWords (Howell & Veale, 2006a).

Early research in this area focused on engaging game heuristics (Malone, 1982). This research describes aspects to educational games such as fantasy, challenge and curiosity. Malone hoped that the spread of cheap computers would eventually mean computers in the home and more in schools. Engaging games were a means to encourage this trend. More recent analysis of engaging games includes models for research and development of serious games (Garris, Ahlers, & Driskell, 2002). The authors cite studies that improved learning from serious games and formulate a model of these characteristics. Adding to Malone’s engaging game aspects, they add rules/goals, sensory stimuli and control. They include several learning outcomes, including skill-based, affective and cognitive.

Motivational theories behind engaging games include intrinsic motivation theory, self-determination theory and interest theory (Medina, 2005). Medina concludes that an engaging game must have challenges, exploration and social interaction.

A psychological theory that describes the sense of immersion and engagement in any activity is flow (Csikszentmihalyi, 1991). Csikszentmihalyi described flow as the state of immersion that a professional experienced when they were performing a complex task such as a sports player in a game or a surgeon operating on a patient. To induce a state of flow, an activity should have clear goals, focus the player’s attention, create a feeling of being ‘lost’ in the activity, create a distorted sense of time, give immediate feedback to the user, be balanced so the challenge matches their ability and be rewarding. Researchers have examined flow as a critical aspect to games (Fullerton et al., 2006). Some games have attempted to implement flow concepts, such as The Cloud Game, Flowers and Flow. These were originally designed by academic researchers, but were also released as commercial games (Chen, 2007). Further research has examined how to create a state of flow to ensure an engaging game is
also educational (Bizzocchi & Paras, 2005), (Guynup & Demmers, 2005), (Dickey, 2006). In order to make BuzzWords engaging and fun to play, the game play design emphasised activities that would create a state of flow in the player, such as clear goals, immediate feedback and balance to the game. However, a sound educational framework was also necessary, so BuzzWords was implemented as an Intelligent Tutoring System.

### 2.2 Intelligent Tutoring Systems

Intelligent Tutoring Systems evolved from early educational applications as a need for a formalised structure to electronic teaching tools was identified. An ITS is an expert system which attempts to supplement traditional teaching methods by adaptively instructing and assessing a student. The main distinction between the passive e-learning systems and an ITS is that the ITS builds a model of the student’s knowledge of the domain being instructed. This allows the system to incrementally or episodically develop a student’s understanding in various directions, usually in a student directed fashion (Freedman, 2000).

The basic structure of an ITS contains Domain, Tutor, Student and Communication modules. These have been described in detail by ITS researchers (Polson, Richardson, & Soloway, 1998), (VanLehn, 2001). An ITS can be adapted to suit a variety of subjects and educational environments, from early multimedia to game environments. The domain module is the complete set of subject matter relating to the subject that the ITS aims to present and assess. These systems are easier to produce for mathematical subjects, as firm logic rules can be set for the domain module to analyse the data (Virvou & Moundridou, 2000). If a class of students consistently fail to answer a particular form of question correctly, the domain module may flag this as a problem area and update its rules accordingly (Abbott, 2006).

The tutor module uses aggregate usage data from the student module to determine the next lesson or section of the subject matter to present to the student. The student module acts a model of the student’s knowledge, storing the composite data set of the student’s performance with the ITS. The communication module accepts input from the user and displays syllabus data or some form of assessment. The communication module is in effect the user interface. Van Eck reviews a number of ITS
implementations including algebra, geometry, physics and computer programming languages (Van Eck, 2006).

2.3 Domain Module Data Generation

BuzzWords required a domain module with data to allow it to assess verbal creativity. This is a difficult task, as there is no standard format of domain module data, and it required a data collection, refinement and compilation of the relevant data. Usually this requires an education expert to compile, organise and categorise the data. BuzzWords uses a defined subject area (compound terms in the English language) that were algorithmically extracted from a computational linguistics database known as WordNet (Miller, 1995). A broad search of WordNet was performed, which produced a list of 36,821 compound terms. Unfortunately, many of these terms were inappropriate for the game. A filtering process to produce a valid set of compound terms from the complete list was required (Howell & Veale, 2009). Oddly formatted terms, such as dates or times or any term that used a numeral in either word of the compound were filtered along with Latinate terms. Slang or colloquial terms were also filtered. The final, filtered list contained almost 30000 compound terms.

The domain module data would require meta-data about the compound nouns to discern the challenge or difficulty of a particular term. This was achieved by using a measurement of how common a compound term phrase was using English text on the internet. Each compound term was cross-referenced with its web frequency. The final data could then be incorporated into the domain module of BuzzWords.

The frequency metric allows a dynamic difficulty adjustment mechanism be built which analyses previous moves and determines in real time the player’s progress. This enables BuzzWords to change the difficulty during the game. Note that a poorly performing player will have the difficulty eased automatically. This is a significant departure from traditional games, where the difficulty level only increases, or games where the user must specify the difficulty level manually (Howell & Veale, 2006b).
3. Testing & Results

BuzzWords is a word matching game that tests player’s familiarity with certain words. This verbal creativity measurement goal is similar to traditional analogy standard aptitude tests or the Remote Associates Test (Mednick, 1968). The RAT requires a word be supplied which matches a triad of given words (e.g. library, phone, chapter could be matched by book). This test is based on Associative Theory and describes the creative processes used when connecting unrelated objects or words to form new expressions. As the original RAT is aging and tends to have particularly American-centric language and some archaic expressions, the word triads and solution words used in the testing were drawn from a more recent set (Bowden & Beenman, 2003).

The hypothesis of the pilot test was that there was a positive correlation between performance in the RAT and performance in BuzzWords. An initial experiment to test this hypothesis deployed the game to volunteers (n = 31). Participants were all 3rd level college students studying computing or science. All students were required to be proficient in reading English (as the game was only available in English) but some participants were not native English speakers (n = 4). All participants were between 18 and 40 years old. The testing process involved a 30 minute RAT followed by an open-ended BuzzWords game. The first 30 minutes of play for each player were recorded. Analysis of these logs produced the number of moves (a move being a word placement in the game, not a positional adjustment), total score, maximum difficulty level achieved and average difficulty level throughout the game. Some experiment participants (n = 4) were unable to complete the second 30 minute BuzzWords play test for various reasons, this was noted when they withdrew from the experiment.

The final analysis of native English speaking participants who completed both the RAT and the full playtest of BuzzWords was performed on a subset of participants due to the language and time factors described (n = 26). Analysis of results included the RAT score (out of 100), max difficulty level achieved in BuzzWords, average difficulty level achieved and score. The participants performed poorly overall on the RAT (M = 36.66, SD = 15.25). The number of moves (word matches) made by players varied (M = 285.88, SD = 77.61) with the minimum number of moves by any player being 156, the maximum 468. The dynamic difficulty level adjusted from the initial starting level of 20 for all players. The maximum difficulty level achieved
represented the most difficult level the player played, even if it was for only one move (M = 31.19, SD = 4.34). There was a positive correlation between the RAT score and maximum level achieved (r(26) = 0.471, p = 0.015). However, the average level achieved by the player is a more accurate reflection of their verbal creativity as measured by BuzzWords (M = 26.59, SD = 4.00). The correlation between RAT score and average level was stronger (r(26) = 0.653, p = 0.000). A scatter plot of RAT score vs. average level can be seen below in figure 2.

![Scatterplot of RAT Grade vs Average Level](image)

Figure 2: Scatter plot of RAT Vs Average Level

4. Conclusions and Future Work

A number of observations are possible based on the test results. The primary result of interest is that there is a correlation between average level attained in BuzzWords over time and performance in the RAT. Whereas the RAT is generally a single use only test, BuzzWords has many thousands of compounds to keep the player engaged, regardless of their level. It can also be inferred that as the average difficulty level is indicative of player’s ability, so the dynamic difficulty adjustment mechanism is functioning well. Furthermore, although the experiment did not test flow states in players, the game design did not require detailed play instructions or complex
strategies to be mastered, so it could be argued that the flow state factors were facilitated by the game design. Once the test was over, participants were informed that they could leave if they wished. Most participants (over 80%) chose to continue playing the game after the 30 minute game session was complete. This may be indicative that they were engaged in the game and did not want to stop playing, even though other games were available to them. If measuring the state of flow was the objective of the experiment, the players would have been measured with the Flow State Scale. Students were observed comparing their scores after the experiment, even though the game expressly avoided a high-score table as it was deemed educationally inappropriate. Although the game records and displays the details of each individual game, these results can only be shared verbally.

Finally, the algorithmically generated data for the ITS domain module provided a wide variety of compound terms, which meant that each player could have a tailored game play experience.

This experiment was performed to test the hypothesis of correlation between performance on the RAT and average difficulty level achieved in BuzzWords, and a correlation is evident from the results shown above. There were some issues affecting the outcomes that became evident after the experiment was performed and results analysed. The most obvious issue was the poor performance on the RAT by the students (M = 36.66, SD = 15.25). The RAT score was calculated by assigning one mark for every correct answer over the 30 questions, and no partial marks were awarded for answers that matched some but not all of the words in the question triple. It would be expected that some students would perform better than others, but no student scored higher than 66.7%. It is likely that the random selection of RAT questions, which were sampled on increasing difficulty scores from the Bowden research, happened to choose questions that were either too difficult or inappropriate for some cultural or other reason. The RAT questions which proved most difficult included the triple fight, control, machine (answer: gun) and the triple grass, king, meat (answer: crab). An improved selection process for RAT triples will be used in future experiments. Other issues included the fact that some participants were not native English speakers (n = 4). Although it would not be fair to exclude volunteers who wished to participate in the experiment on the basis that they were not native speakers, it is important to note participants who are not native speakers so that their results can be incorporated into the analysis. Fortunately in this experiment the
number of non-native speakers is small and not likely to distort the results. Finally, some students were forced for various reasons to leave before playing BuzzWords for a full 30 minutes (n = 4). In ideal circumstances, it would be preferable to have all play testers complete the game. Overall, the experiment facilitated the testing of the game based assessment being a valid method to assess verbal creativity. Future experiments include a test of whether players are entering into flow states, and whether the use and non-use of the dynamic difficulty adjustment mechanism has a measurable effect on the player’s state of flow.

In conclusion, serious games in the classroom can engage the student and turn a traditionally unpopular task, test taking, into a fun activity. There are however hurdles to developing such games. These can be overcome by developing causal games that focus on simple game play. The educational aspects must be catered for, and some subjects can be algorithmically data mined to produce the game data. In order to ensure the game is suitable for all students, the game should dynamically adjust the game difficulty, which will give balance to the game and encourage a state of flow. Although these hurdles are by no means insignificant, it is possible to produce serious casual games with simple but engaging game play and sound pedagogical aspects. It remains to be seen whether games will become a staple tool of the educator, but modern tools and techniques for developing games are simplifying the process and within the foreseeable future, it will be possible to rapidly prototype, test and deploy games into the classroom. Whether this will be a boon or a curse remains to be seen.

5. References


