Designing a Serious Game for Eliciting and Measuring Simulated Taxpayer Behavior

TOBIN BERGEN-HILL, ROB CREEKMORE, JOHN BORNMAN

The MITRE Corporation, McLean, VA 22012

Abstract. Serious games are designed for an ultimate purpose other than pure entertainment. In the case of a business simulation game, the design typically is centered on a "table top exercise" environment where the player is learning successful business practices. The virtual reality based technology and the design methodology described in this paper goes well beyond such typical business game design. It introduces more immersive three dimensional taxpayer behavior elements to the game – an element not usually found in commercially available business simulation games – and also triggers tax-related behaviors and records the resulting behavioral response data in order to test behavioral hypotheses. The virtual reality based serious game design methodology is discussed, including the development team's adaptation of the game to test a specific experimental hypothesis regarding taxpayer behavior. Various techniques and technologies for measuring taxpayer behavior are described, and their effectiveness is evaluated in the context of an experimental test session involving human subjects. Follow-on work and applications of this approach are proposed in the conclusion.

Keywords*: agents, data collection, virtual reality environment, serious games, taxpayer behavior, modeling and simulation*

Introduction

MITRE has recently concluded its research into adapting virtual-reality-based serious game technology to allow organizations such as the Internal Revenue Service (IRS) to simulate taxpayer compliance behavior well enough to test hypotheses around taxpayer outreach. The goal is to understand reasons for noncompliance and to encourage voluntary compliance through education and other outreach activities. While a player is immersed in a virtual environment there are potentially many opportunities to extract meaningful data related to their actions. To adequately test experimental hypotheses around proposed methods for influencing behavior, the challenge is twofold: to elicit the kind of behavior desired, and to select the data that best indicates the decision-making process of the player. This paper describes the serious game technology aspects of MITRE's research into eliciting and measuring player behaviors within a

virtual reality based business simulation game that includes tax-related activities of a sole proprietor. This paper explores how technology and design methodology challenges were addressed to ensure a robust method for organizations to test their hypotheses about the best outreach methods for understanding and influencing taxpayer compliance. Additional papers are being submitted for publication that more specifically describe from a social science point of view the methodology and results for the data collection and analysis aspects and the implications of this research for studying taxpayer behavior.

A key assumption of this experiment is that the data on the decisions made by the player can be correlated with that individual's personal experience and attitudes towards running a small business. An additional assumption is that this system can effectively simulate how they would behave in real life, which therefore allows organizations such as the IRS to test their hypotheses about the effectiveness of particular outreach approaches in influencing the tax-related decision-making process of small business owners.

Serious Game Design Method

To design a serious game capable of eliciting these desired behaviors, the MITRE team followed the Simulation Experience Design Method, a four-phase approach that focuses on the interactions between the player and the game system (Raybourn, 2007). By following these four phases – "interaction," "narrative," "place," and "emergent culture" – throughout the design and implementation of the game, the MITRE team sought to create an environment capable of fostering meaningful behaviors that lead to understanding taxpayer compliance.

During the initial phase, which involves defining the interactive roles, the MITRE team interviewed several volunteers that had previous experience in running small businesses. The goal was to determine the major stakeholders involved in a business and to enumerate their day-to-day interactions plus identify any issues they face during a typical year of business. After some discussion – after assurances that the collected information would remain anonymous – the volunteers divulged various ways that tax laws might be transgressed. The information gathered from these sessions led the team to decide to structure the game around the ownership of an independent ice cream store. The independent nature of the store allows the player (in the role of the sole proprietor) more freedom in business-related decisions. In particular, the choice of ice

cream as the product forces a regular complete turnover of inventory of small-priced items, thereby allowing more flexibility in transaction policies (e.g. accepting only cash transactions) and in inventory management (e.g. claiming loss of inventory when in actuality it was sold for cash). As a whole, this arrangement provided the basis for a challenging set of scenarios that a sole proprietor may face.

The second phase – telling the story – involved transforming the valuable information provided during the first phase into an extensive narrative that explicitly listed every possible interaction and decision facing the player. The narrative made it easy to review the decision points in the game and identify those which would elicit tax-specific behaviors, which was crucial to supporting the selected hypothesis for this experiment: that taxpayer compliance could be affected by a change in the tax form instructions. The level of compliance of a sole proprietor would be assessed based on the results of those tax-specific decisions.

This narrative guided the development of the game environment and the creation of non-player characters (NPCs) in the third phase. The MITRE team selected OpenSimulator (or "OpenSim"), an open source multiplayer virtual environment system based on the Second Life environment, as the development platform for the game (OpenSim, 2012). The inherent NPC generation capability of the OpenSim server allowed the team to populate the business simulation with representations of the various stakeholders: bookkeeper, salesperson, supplier, previous owner, and tax preparer. To direct the NPCs and to trigger various events in the game, the team developed a script-driven engine that sent commands to the OpenSim server. The script driven NPCs provided consistent, objective interactions with all test subjects, something that human-controlled characters – due to their subjective nature – could not guarantee. Another interactive element of the environment involved in-world simulated computer monitors that could display web pages, using the inherent 'Media on a Prim' capability (Linden & Linden, 2011). Some of them permitted player interaction (e.g. an expenses table displayed on a virtual laptop that allowed the user to tally expenses by category when filing their taxes).

The MITRE team deviated slightly from the Simulation Experience Design Method during the fourth phase, which is usually about creating a lasting experience outside of the game itself through feedback and motivation. Since our interest was more about understanding the game play behaviors and testing our hypotheses, the players were asked to complete two on-line

surveys, revealing their demographics and tax compliance preferences. The second online survey was based on the Tax Compliance Inventory (Kirchler & Wahl, 2010), the results of which could be correlated to the player's in-game tax law compliance behaviors. Additionally, during a post-game interview session, the test instructor asked the participant a series of questions to (1) generate feedback as to whether the game provided a realistic environment and (2) discern between intentional *deception* and *confusion* for non-compliant answers on their tax forms. To motivate the players in the full experiment, the MITRE team offered a cash award based on the player's game score, which is based on the amount of net income earned during the simulated year, minus any tax penalties and including any tax refunds. Such a reward was based on the need for *saliency*, one of the sufficient conditions for a valid controlled microeconomic experiment (Smith, 1982).

The team's main focus during this last phase, however, was the development of automatic event logging techniques to support robust data analysis for determining player intent (Nacke, Lindley, & Stellmach, 2008). During the game, the 'chat logs' (or transcripts of the text-only interchanges between the player and the NPCs) were recorded into an external database using a system called "SL>>DB" (Aubret, 2009). As the player carried out various tax-related decisions (e.g. whether to make quarterly tax payments, whether to record cash income), the resulting choices made by the player were also recorded into a database. These decisions were displayed to the test instructor at the end of the game, for reference during the post-game interview session mentioned earlier. For example, in the later stages of the game, a tax preparer NPC asked the player for the values to insert into a Schedule C and Form 1040; the player's responses were then stored by line number and form in the database, and the response was evaluated as to whether it was compliant so it can be addressed during the interview. Another crucial recorded data point was the timing of the player's responses after a character posed a question that relates to these tax-related decisions. Long reaction times can be an indicator of intentional deceit (Sheridan & Flowers, 2010) or a lack of comprehension (Kelly, Maris, & Özyüre, 2009). However, this must be coupled with other evidence to determine the reason behind the player's apparent non-compliance with tax laws.

Finally, the entire sessionwas captured using two video recording technologies: a webcam to record player facial/bodily reactions to the game and any verbal feedback they provide; and

frame-grabbing software to record the player's game session as they saw it. The webcam was positioned so that both the player and the screen are visible. The purpose of this arrangement was to capture the visual in-game context for any "talk-aloud" comments or facial expressions made by the test subject while they played the game e.g., a puzzled look on a test subject's face could be correlated to the tax preparer NPC asking a confusing question in the game.

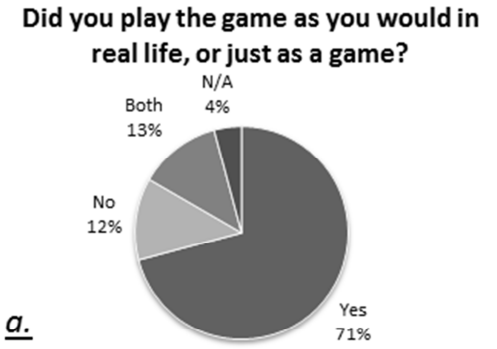
Findings

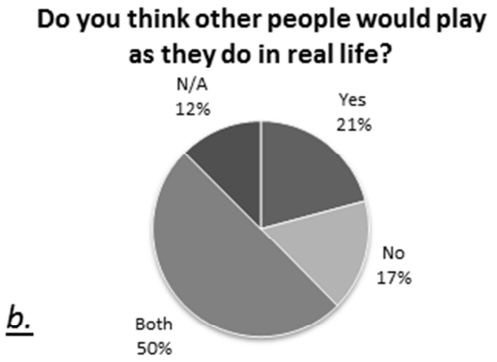
The MITRE team conducted experimental sessions in January and February of 2013 involving 24 test subjects that consisted of business and economics students recruited from George Mason University. During each session, the test subject played a version of the business simulation game that simulated managing the ice cream store business in two phases: (1) a tutorial where the player is learning the business from the previous NPC owner and (2) a full game that simulates the first quarter of a full year of running the business on their own. Both the tutorial and the full game presented the test subjects with opportunities to employ questionable business practices. To test the hypothesis mentioned earlier (that taxpayer compliance could be affected by changing the tax form instructions), half the test subjects received a regular set of instructions for filling out the Schedule C, and the other half received a modified version that used more punitive language in the introduction regarding potential imprisonment for violating tax laws. A test instructor was on hand to provide instructions initially and conduct the interview afterwards. All of the recording technologies described earlier were used during these sessions. The following paragraphs provide an assessment of the effectiveness of the various technologies in eliciting and measuring behaviors.

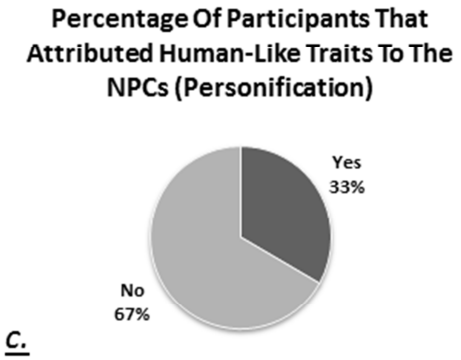
The webcam was an excellent resource for recording the player's emotional responses to the game and to the questions posed afterwards. The team manually generated transcripts of the post-game interview from these videos, which were analyzed for trends and clues to the nature of non-compliant responses (i.e. whether the participant was intentionally 'cheating' on their taxes or whether they were confused).

As mentioned previously, some of these interview questions were designed to determine whether the participant was fully engaged in the virtual environment. Figure 1 displays the results in graph form of these engagement questions and unsolicited player comments during

game play. Graph 1a shows that the majority of participants responded that they played the game as if they owned the business in real life, as opposed to playing it simply as a game. Graph 1b shows how the test subject thought about whether other game players were similarly engaged; the results show that they were uncertain about this. Graph 1c is based on sets of key words found in the participant's transcripts that attributed human-like traits to the various characters they encountered in the game; a third of the participants exhibited this personification behavior. Graph 1d is based on participants' recorded chat interactions that went beyond simply answering the questions posed by the NPCs. Of all the participants, 21% had posed unsolicited questions to the NPCs (e.g. "Mr. Frostman [the previous business owner NPC], have you ever been audited by the IRS?"); a third of them had issued a greeting upon meeting NPCs for the first time; 21% had stated their thoughts to the NPCs without solicitation (e.g. "Sir, I do not believe that this is the correct purchase price per the invoice"); and 29% provided commentary on their reaction to certain events in the game (e.g. "Yay!").







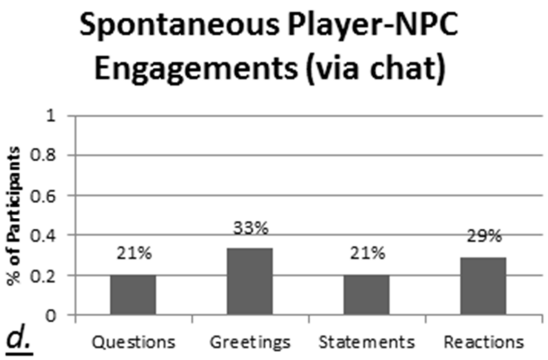


Figure 1: Participant engagement results. a) Responses to a question about the participant's engagement in the virtual environment. b) Responses to a question about their thoughts on other participants' engagement. c) Percentage of participants whose verbal responses exhibited personification of the NPCs. d) Instances in the chat recordings where the participants engaged with the NPCs beyond answering questions

Unfortunately, for one of the test sessions, one of the test instructors had not enabled the web recording; consequently, the post-game interview for that session was lost. Perhaps the recording could be triggered automatically in future experiments. Additionally, the task of searching the video for relevant "talk aloud" comments was prohibitively time consuming. The use of commercially available speech recognition software was briefly explored to automate this task, but the software required several minutes of 'training' by the speaker before it could accurately transcribe text from speech. Consequently, no transcripts of these in-game comments were generated for this experiment.

The frame-grabbing software successfully captured the non-verbal player avatar behaviors in the game, along with the contextual information surrounding the decisions made by the player. It went a step beyond the chat recording mechanism in that it captured the text that the player would enter, but later delete, as they changed their mind about their response; the chat recorder only captured the final text that the player submitted. This provided insight into the thought process behind the player's decisions.

The online surveys were effective in capturing the demographics of the test subjects and characterizing their tax compliance index. As noted above, a more detailed exploration of this aspect of the research is being published in other papers.

The script-driven engine has proven to be particularly useful. A custom scripting language gave the team the flexibility to add any commands needed to support the game scenario expressed in the narrative. However, each command needed to be coupled with a capability provided by the other software components – some of which varied in their reliability – that would perform the command, e.g., the NPC controls, dynamic objects in the environment that would appear or disappear on command, wall-sized displays of web pages, sensor objects that would report the presence of the player, player responses and actions to store in the database, etc. Consequently, the effectiveness of the script relied on the functionality and responsiveness of these components of varying reliability. Most of these components were matured in time for the

experiment, with the exception of the NPC functionality. For this reason, the script for the full game only covered the first quarter of the simulated business year.

The bulk of development time was spent on the NPC functionality. The initial version of the OpenSim server had no native NPC support; so, a custom version was developed, which consumed many staff hours to perfect. This capability required a dedicated client machine for each NPC. Since the full game could involve up to twelve possible characters, the number of supporting machines needed was prohibitively large. Fortunately, late in the development cycle, a new version of OpenSim was introduced with native NPC support, meaning that the characters could be generated on the server without requiring dedicated client machines. But its capability for moving the characters along a string of waypoints was problematic, resulting in the NPCs occasionally drifting between locations. Also, the new server did not support inventory items for the NPCs. The script needed the ability to detect when the player had given the NPCs certain items (e.g. payment for services), which was beyond the ability of the OpenSim server. As a workaround, the test instructor needed to log in to the game as an 'Inventory Valet' character using a traditional client; the player would give the items to the valet, which the script could detect. This arrangement was clumsy, at best, but it allowed the narrative to move forward.

While the web pages displayed on various in-game computer monitors added to the realism of the business environment, the toughest challenge was in keeping the displayed information up to date. The web pages can query the database for relevant information, e.g. cash sales for the current quarter or emails from the past three months. But the web page needed to know the current quarter and year. This was solved by adding commands to the script-driven engine to store the current quarter and year in the database; the web page would query these variables, then structure the other queries accordingly. Populating the database with enough detailed historical financial information (e.g. sales, expenses, taxes paid) to support the tax-related aspects of the narrative became the next challenge. Nominal figures were used as placeholders until a simplistic microeconomic model was developed to determine sales figures based on pricing, supply, demand and salesperson skill. It was extremely difficult to keep all this information in sync with the game timeline. These issues were corrected for the final experiment, providing the participants with a business simulation game containing financially consistent information.

Conclusions

The February 2013 experiment has proven the viability of designing a virtual reality based serious game for capturing taxpayer behaviors to support testing of behavioral hypotheses. The virtual environment and its supporting technologies created by the MITRE team have successfully recorded the transcripts, actions and thoughts of the experimental test subjects as they carried out the role of a sole business proprietor of an ice cream store. The data capturing technologies that proved themselves the most useful – namely the surveys, the webcam, the frame grabbing software, the chat logs, and the script-driven engine combined with the back-end database – were crucial to generating the data needed to make conclusions regarding the experimental hypothesis. Not only are the player-generated responses recorded in the data but also the actual financial records from the year of simulated business which, combined with the interview and survey data regarding the player's tax compliance attitudes and reasoning during game play, provides a basis of comparison to easily identify areas of tax law compliance versus deviance. Additionally, the results demonstrated the ability of this virtual reality technology to engage test subjects sufficiently enough to generate responses similar to managing an actual business.

The next step for the MITRE team will be to align the experimental design with a hypothesis suggested by the IRS. Several presentations and discussions with IRS staff have resulted in a number of promising potential hypotheses. Once a hypothesis is selected, some adjustments to the scripts will be required to match the narrative to the hypothesis. Then the team will solicit an appropriate experimental group according to the target demographics relevant to the hypothesis. The group should be of significant size – far more than the 24 participants from this first experiment – in order to generate enough data to make meaningful conclusions. By providing convincing evidence as to the validity of the hypothesis, the team will demonstrate more thoroughly to the IRS this new means for validating proposed taxpayer outreach methods well beyond current means based on surveys or focus groups.

Acknowledgements

The authors would like to thank Kathryn Laskey and Walter Andrew Powell from GMU's C4I Center for their assistance in recruiting the test subjects for the experiment. Additionally, the authors are thankful for the participation of the students of George Mason University.

References

OpenSim. (2012). Retrieved March 2012, from OpenSimulator Web Site: http://opensimulator.org/wiki/MainPage

Aubret, L. (2009). *SL >> DB: Simple Database Storage for Second Life*. Retrieved March 2012, from Second Life Marketplace: https://marketplace.secondlife.com/p/SLDB-Simple-Database-Storage-For-Second-Life/74318?id=74318&slug=SLDB-Simple-Database-Storage-For-Second-Life

Kelly, S., Maris, E., & Özyüre, A. (2009). Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21, 260-267.

Kirchler, E., & Wahl, I. (2010). Tax compliance inventory TAX-I: Designing an inventory for surveys of tax compliance. *Journal of Economic Psychology, 31*, 331-346.

Linden, J., & Linden, B. (2011). *Shared Media - Second Life*. Retrieved March 2012, from Second Life Knowledge Base: http://community.secondlife.com/t5/English-Knowledge-Base/Shared-Media/ta-p/700145

Nacke, L., Lindley, C., & Stellmach, S. (2008). Log Who's Playing: Psychophysiological Game Analysis Made Easy through Event Logging. *Second International Conference on Fun and Games*, (pp. 150-157). Eindhoven.

Raybourn, E. (2007). Applying simulation experience design methods to creating serious game-based adaptive training systems. *Interacting with Computers, 19*, 206-214.

Sheridan, M., & Flowers, K. (2010, December). Reaction Times and Deception - the Lying Constant. *International Journal of Psychological Studies, 2*(2), 41-51.

Smith, V. (1982). Microeconomic Systems as an Experimental Science. *The American Economic Review, 72*(5), 923-955.

Weber, E. U., Blais, A. R., & Betz, N. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making, 15*, 263-290.