SHORT COMMUNICATION

A Method for Longitudinal Behavioral Data Collection in *Second Life*

1 Introduction

Persistent online virtual environments, whether game worlds like *World of Warcraft* or social worlds like *Second Life,* provide social scientists with the opportunity to collect longitudinal behavioral profiles from users. These environments allow behavioral measures of interesting variables at the individual and group level to be collected and analyzed. For example, studies have examined mutual gaze and personal space in *Second Life* (Friedman, Steed, & Slater, 2007; Yee, Bailenson, Urbanek, Chang, & Merget, 2007). Behavioral data also allows researchers to avoid self-report questionnaires, which have been shown to produce unreliable measures (Slater, 2004).

On the other hand, even though these virtual environments are treasure troves of data for social scientists, typical social science curriculums do not provide researchers with the necessary background skills (e.g., programming, databases, integration across platforms) to collect data from these emerging environments. Our goal here is not to provide, by any means, a cutting-edge solution from a technical perspective, but rather, to provide a foundational framework that others can easily modify for a wide variety of purposes. The solution we describe allows researchers to capture avatar-related data from *Second Life* (SL) at a resolution of one minute or less over a period of weeks.

1.1 Assumed Background Knowledge

While we will describe the solution at a fair level of detail and an online addendum (http://www.nickyee.com/pubs/secondlife.html) will provide all the scripts to replicate our framework, it is not our intent here to teach *Second Life*'s scripting language (LSL), web-based programming languages, or database languages. As such, we will assume that the reader has some background knowledge of LSL, PHP, and MySQL, enough to modify the provided scripts.

It must be pointed out that using PHP and MySQL implies that one has access to a web host that has these packages installed. Both PHP and MySQL are open-source packages and most commercial web hosts provide both in their basic services. Both are popular and standardized web-based tools that many computer science graduate students or IT staff will be able to assist with.

2 Technical Architecture

In the hypothetical study for which our solution is designed, imagine that we wanted to track behavioral data from 10 participants, each using their own SL avatar, over 2 weeks at a resolution of 30 s, capturing every avatar-related variable that SL gives us access to (e.g., Cartesian coordinates of locomotion, whether the avatar is typing, every character the user types, changes to body size, etc.). Also assume that each participant is active in SL for at least 5 h each week.

Due to the constraints in LSL, it is not possible to store large amounts of data within SL directly. Thus, any data

Nick Yee
Jeremy N. Bailenson
nyee@stanford.edu
Department of Communication
Stanford University
Stanford, CA 94305

Presence, *Vol. 17, No. 6, December 2008,* 594–596

collection in SL must be piped to an external database. However, since LSL does not provide any direct methods for connecting to databases, a web-based script must act as the bridge between LSL and the database.

2.1 The LSL Script

The LSL script uses a timer function to collect and send avatar-related variables every 30 s. In the scripts provided, the variables for setting the timer interval are clearly marked and can be changed according to the desired collection resolution. The provided script should be attached to an object and thus will serve as a tracking device. The script will start when users attach the object to themselves. To avoid attaching the object to the avatar visually, we recommend attaching the object to the HUD element (Heads Up Display) instead. In this way, only the user can see the object. And while the object can be anything, including being invisible, we recommend coloring the object with a bright or noticeable color to serve as an indicator of whether the tracking device is being worn. A further note is that changing outfits in SL can detach all previously attached objects. Thus, participants should be reminded to check whether the tracking device is still attached after changing outfits.

2.2 The PHP Scripts

The LSL script will send the collected data to a set of PHP scripts that provide the bridge to the MySQL database. The SL variables are sent via the URL of the PHP page. One additional consideration is that we want to minimize opening connections to the database as this is a costly server-side operation. To buffer the incoming stream of SL data, it is first appended to a text file. Once the text file reaches a certain size, it is piped to the database in one session. There are three PHP scripts in the solution. One script contains the shared connection function to the database. Another script handles the incoming stream of numeric variables. And the final script handles the intermittent chat data (which occurs whenever the user chats, rather than every 10 s). As a precautionary measure, all incoming data is also stored in text files indexed by avatar name in addition to being piped to the database.

2.3 The MySQL Database

The MySQL database is composed of three tables. The table schemas are also provided in the online addendum. One table lists the users. The PHP scripts automatically update this table when a new avatar name is spotted and this table also stores overall metrics for easy administrative tracking, such as total time seen online, last date and time seen online, and total lines of chat. Another table stores the numeric variables associated with the avatar. The final table stores the chat data.

2.4 Data Analysis in MySQL

An additional advantage of storing data in a database directly is that it makes it possible to parse the data easily. For example, the database structure makes it easy to find the total number of lines of chat each user typed, the average number of characters in each chat exchange, or to find the most popular zones visited by the participant sample.

3 Additional Technical Considerations

3.1 Sampling Resolution

Expected web access counts should be extrapolated from the desired sampling resolution and sample size. For example, at a 10 s resolution with 50 participants, the website might be accessed 3,000 times per hour during peak usage. The web host should be queried as to the feasibility of this access rate. The best way to reduce web access counts is to lower the sampling resolution.

3.2 Database Access Deterioration

Database access and storage speed deteriorate linearly with the current number of rows stored in the ta-

ble. One simple solution to prevent the database from stalling is to periodically pipe the incoming data to a new table and to merge the tables as necessary when data collection ends. If it is expected that data will be analyzed on a week-by-week basis, then the final table merge may not be necessary.

3.3 LSL Idiosyncrasies

It must also be remembered that SL was created as a public sandbox, and many features exist to prevent players from antagonizing each other using scripts. Thus, certain avatar-related variables that one might assume that LSL provides are not available or return a dummy value. For example, while there are *x* and *z* coordinates associated with avatars, this does not represent the avatar's size. However, the *y* coordinate does represent the avatar's height. Researchers should consult LSL's list of methods and attributes as different behavioral measures are considered in a research design to ensure that LSL allows one to collect data on that variable.

4 Online Addendum

All the scripts necessary to replicate this solution are provided in an online addendum located at http://www.nickyee.com/pubs/secondlife.html. The scripts provided are also well-commented to allow for modifications.

References

Friedman, D., Steed, A., & Slater, M. (2007). Spatial social behavior in Second Life. In C. Pelachaud (Ed.), *Intelligent virtual agents 2007* (pp. 252–263). Berlin: Springer-Verlag.

Slater, M. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators and Virtual Environments, 13,* 484–493.

Yee, N., Bailenson, J. N., Urbanek, M., Chang, F., & Merget, D. (2007). The unbearable likeness of being digital: The persistence of nonverbal social norms in online virtual environments. *Journal of Cyber Psychology and Behavior, 10,* 115–121.