Avatar Web-Based Self-Report Survey System Technology for Public Health Research: Technical Outcome Results and Lessons Learned

Craig Savel[1]; Stan Mierzwa[1]; Pamina M. Gorbach (Dr.P.H.)[2]; Samir Souidi[1]; Michelle Lally (MD)[3]; Gregory Zimet (Ph.D.)[4]; Adolescent Medicine Trials Network for HIV/AIDS Interventions

1. Information Technology, Population Council, New York, NY
2. Department of Epidemiology, University of California, Los Angeles (UCLA), CA
3. Alpert Medical School of Brown University, Lifespan Hospital System, and VA Medical Center, Providence, RI
4. Indiana University School of Medicine, Indianapolis, IN

Abstract

This paper reports on a specific Web-based self-report data collection system that was developed for a public health research study in the United States. Our focus is on technical outcome results and lessons learned that may be useful to other projects requiring such a solution. The system was accessible from any device that had a browser that supported HTML5. Report findings include: which hardware devices, Web browsers, and operating systems were used; the rate of survey completion; and key considerations for employing Web-based surveys in a clinical trial setting.

Keywords: Self-Report Data Collection; Electronic Data Collection; CASI; Avatars; HTML5; Smartphones; Web Browsers; Web-Based Survey; Clinical Trials

Correspondence: smierzwa@popcouncil.org

DOI: 10.5210/ojphi.v8i2.6719

Introduction

Given the challenges associated with collecting accurate self-reported data in research studies, new approaches using customizable avatars and online questionnaires are being developed in an attempt to improve the frequency and accuracy of self-reports. In looking for ways to better collect survey data, we developed a technology solution consisting of a Web-based self-report data

collection system that used customizable avatars to collect data. Participants were instructed to take two surveys at specific time periods. Self-created avatars "traveled" with participants such that they would appear during the first and second surveys and also appear if a participant restarted a survey that was not completed on the first attempt. The survey website was HTML5 compatible, but we elected not to use HTML5 local storage because of confidentiality concerns and requirements for data security in this public health research study. The survey was designed to work on any HTML5-compatible browser and on any tablet, smartphone, or computer that had a browser that supported HTML5. Although we recognized that participants who had older browsers (Internet Explorer 8 or earlier, old versions of Firefox and Chrome) might not be able to access and complete the self-report survey, it was felt that most of the target audience would have little trouble doing so. To the best of our knowledge, no other study has specifically examined the devices, Internet browsers, and operating systems used to complete a Web-based self-report survey for a public health research project.

Methods

Many self-report electronic data collection systems in HIV and/or other public health research studies use technology that exists in a controlled environment. The study protocol generally dictates the type of computer or device to be used and the method for presenting the study's survey. This study was an ancillary study to a large clinical trial of pre-exposure prophylaxis use by HIV-negative adolescent males 15–17 years of age conducted at 12 sites in the United States through the NICHD-funded Adolescent Trials Network. Most of these sites were adolescent HIV clinics (ATN 110/113). After completing procedures for the clinical trial, adolescent males enrolled in the trial were offered participation in this ancillary study. If they agreed to participate, they were given choices on how to complete the study questionnaire. Participants were able to access the Web-based survey either from inside a study clinic (using clinic computers) or from a device of the participant's choice (either inside or outside the clinic setting). This meant that the self-report survey system needed to be built such that participants could access the survey from computers, tablets, or smartphones on a variety of operating systems using many different browsers. The system needed to allow participants to create their own customized avatars that would follow them through the questionnaire. It also had to allow them to edit or use the same avatar in a follow-up survey. The customized avatars would appear on each question screen, and they would move to different locations on the screen in order to present the survey questions within a text bubble [1].

During the Web-based self-report survey data collection, information was collected on several technical measures such as which Web browser and operating system was used. The method of data collection was made available via log files that are common in Web servers. Our study used the Microsoft Internet Information Server to capture this information. Several elements will be reported in the Findings section, including the preference of using the interactive questionnaire Web-based survey system, recording the amount of time to complete the electronic survey, and the percentage of participants that completed the survey. Many of the qualities of the very simple end-user screen design, as well as the elements of start and end time, and computer name were adopted from the Population Council ACASI technology solution [2].

Findings (Results)

The study enrolled its first participants in July 2013 and completed enrollment in July 2015. Since each of the 167 study participants should have made at least two visits to the assigned clinic during the study, and may have accessed the survey additional times as needed to complete it, we were able to collect sufficient data. Because of potential confidentiality issues, we were not able to use cookies to track visitors, even anonymously. We were therefore also not able to use Google Analytics. The preferred strategy was to parse the USER AGENT, glean as much as possible, and write that information into a secure database. Browser data were sent to the server via a USER AGENT text string with every request; this occurred automatically when the browser on the computer communicated with the device. A USER AGENT string indicates which browser was used, its version number, and details about the user's system, such as the operating system and version. For various reasons, including incomplete or corrupted USER AGENT strings, some visits to the survey may not have been logged/recorded (meaning communication between the browser and the server).

Table 1. Basic result data on utilized Internet browsers and versions: June 2013–July 2015

BROWSER and VERSION	# VISITS	% OF VISITS
Internet Explorer	**196**	**34%**
7.0	28	
…..8.0	19	
…..9.0	126	
…..10.0	23	
Safari	**149**	**26%**
0.0 (see note)	110	
4.0	1	
5.0	3	
5.1	3	
6.0	1	
6.1	31	
Chrome	**128**	**21%**
0.0 (see note)	118	
18.0	8	
27.0	2	
Firefox	**4**	**< 1%**

16.0	1	
21.0	3	
Android 4.0	**13**	**2%**
Other or unknown	**89**	**16%**

NOTE: Version 0.0 for both Safari and Chrome browsers occurs when the browser version is not part of the string sent to the server.

These data show some notable differences from general statistics for US users. The website Statcounter (http://gs.statcounter.com/#all-browser-US-monthly-201306-201504-bar) collects statistics on Web usage. Among participants taking the survey, the most popular browser was Internet Explorer by a margin of 8%. The second most popular was Safari, and the third most popular was Chrome. (See Table 1.) In contrast, Statcounter shows that for the time period the survey was running, Chrome was the most popular browser at 31%, Internet Explorer was second at 24%, and Safari was third at 23%. Firefox had 11% usage but less than 1% usage for the survey.

What can account for this difference? It is impossible to know, but we hypothesize that more users opted to fill out the surveys at the clinic than we had expected. Businesses, governments, and social service organizations are often "late adopters" of technology, and if users filled out the surveys at the time of the visit to the clinic that could account for the difference. Since survey participants were young people, we expected mobile browsers, especially iPhones or iPads, to be factored in. Apple mobile products use Safari as the default and this can account for the relative greater use of Safari in our survey as opposed to general statistics.

Table 2. Basic result data on operating systems usage: ATN 123* June 2013–July 2015

OPERATING SYSTEM	# VISITS	% OF VISITS
Windows	302	**53%**
Windows 7	264	
Windows 8	4	
Windows XP	31	
Windows Vista	3	
Mac OSX	99	**17%**
Unknown	89	**16%**
Linux	39	**7%**
iOS	24	**4%**
Android	20	**3%**

*ATN – Adolescent Medicine Trials Network

Striking differences were also found between survey respondents' operating system usage (see Table 2) and the general statistics for US users during that time period. In reviewing Statcounter data on operating systems (http://gs.statcounter.com/#all-os-US-monthly-201306-201504-bar), we notice that more than 50% of responses were from a Windows-based computer, whereas 17% were from a Mac OSX-based system.

Although it is more difficult to determine the device type by webserver log files, one can infer a likely device from a log file. For instance, if a log file entry shows a Windows operating system, the device is obviously a Windows PC or laptop. If a log file shows iOS as the operating system, then it is an Apple mobile device. It is much more difficult to determine, however, whether that device is an iPhone or an iPad. Another problem is that for many versions of mobile operating systems, there is no accurate information in the header file. This is especially true of Android devices.

How successful was the survey? To gauge effectiveness, some of the questions we might ask are: What percentage of respondents finished the self-report avatar survey? How many finished on the first attempt? How long did it take to finish the survey?

The number of respondents who started Survey 1 was 154; 96 completed it. The number of respondents who started Survey 2 was 106; 89 completed it. There was only one user who attempted both surveys and completed neither.

Table 3. Totals of users who started and completed surveys 1 and 2

Started Survey 1	Started Survey 2	Completed Survey 1	Completed Survey 2	Completed 1; did not complete 2	Completed 2; did not complete 1	Completed both surveys
154	106	96	89	9	2	87

Among those who started the first survey, 62% finished it. Almost 84% of those who started the second survey completed it. (See Table 3.)

Participants were allowed multiple attempts to complete the survey. Most who completed the survey did so on one try, although a few required multiple tries.

We cannot know, of course, why a participant needed more than one attempt to complete a survey, but it is instructive to compare survey attempts. For example, what were the browsers and operating systems used for each attempt? Were they the same or different? Can we observe patterns?

A quick and preliminary look at survey statistics shows that most of those who logged in more than one time for a given survey logged in using the same browser and operating system as they had used previously. There were a few trends though. Most users who switched went from a Windows machine using Internet Explorer to either Windows using Chrome or a Mac. The second

largest group of switchers went from Safari on iOS to Safari on Mac OS. Those who started on Android tended to stay on Android.

Some participants did not complete the surveys. This includes those who did not complete the surveys at all and those who did not complete the surveys in the allotted time but returned to restart from the beginning.

For participants who did not complete either survey on the first visit, including those who returned and completed the survey later and those who did not, certain questions were "stoppers." In other words, many users who stopped the survey stopped at the same questions. The one that caused the most users to stop was a complicated type of question that was presented in a calendar. Users were required to answer two questions for each calendar date, accessible via a pop-up. Users were not able to advance to the next question until both questions were answered for all dates. Eighteen participants left the survey without completing the question that was presented in that format. Six users logged in, did not complete any questions, and did not return to complete the survey. Other than that, no more than two users were stopped at any other particular question.

Discussion

The Web-Based Self-Report Survey was made available to participants with the option of taking it in a controlled clinic environment or on their own outside the clinic using whatever device they had, wherever they were. Because of confidentiality concerns it is not possible to verify that the surveys were more often completed in a clinic. In a future similarly designed project, it would be beneficial to consider adding logic to the survey to record whether it was actually taken in one of the original clinic sites on a computer belonging to the study.

The Web-Based Self-Report Survey included many of the assumed benefits of electronic survey-taking via the Web: consistency in survey presentation, minimization of errors in data collected because of edit and range checks, and the ability to know when surveys are completed and consequently prevent users from taking a survey they had already completed. The number of studies collecting self-reported data via the Web continues to increase rapidly [3]. The quality of anthropometric data collected using a Web-based questionnaire, with regard to missing and plausible answers, has been shown to be equal to, or better than, that of data collected using a paper version of the questionnaire [4].

To complete the Web-based survey, participants were provided with the secure link to the site as well as a user ID and password to use when logging in. When surveys were taken in the clinic it was much easier to ensure or validate that the actual participant was taking the survey; when the survey was taken away from the clinic, there is a possibility that the participant is being aided or having the survey done by someone other than themselves.

Limitations

For 16% of the surveys, it was not possible to get information on the browser or operating system. These data come from text strings sent from a user's browser to a server. A large percentage of

data were unclassifiable, and this could have had an impact on the outcome of top browser and operating system used, given that the separation between the top three browsers was small: 34%, 26%, and 21%. It is important to remember that there was no absolute requirement to send the survey and no correction mechanism if the data was incomplete or missing. Karl Groves writes in online design journal *boxesandarrows*:

"Server log files are inappropriate for gathering usability data. They are meant to provide server administrators with data about the behavior of the server, not the behavior of the user. The log file is a flat file containing technical information about requests for files on the server." [5]

Since the Avatar-Based Self-Report Survey was administered in the United States, it was assumed that Internet access would be widely available; it was therefore anticipated that, in most cases, participants would perform the survey at home. However, we cannot be sure that the participants had Internet access via their mobile devices and/or home computers. The requirement of Internet access may lead to a higher rate of completion in clinics if studies such as this are conducted in the developing world where many such public health research projects are likely to take place. In addition, for those who did not have Internet access, it was not assessed if this was associated with any sociodemographic factors. If future similar projects are to include self-report Web-based surveys in the context of a clinical trial, we would recommend reviewing data on Internet broadband access availability. The National Broadband Map (NBM) is an available resource that is created and maintained by the National Telecommunications & Information Administration in collaboration with the FCC, 50 US states, 5 territories, and the District of Columbia (www.broadband.gov). By reviewing the Internet access data available to households in the United States, one could do a scan to ensure that adequate coverage is available. Household broadband adoption rates have increased dramatically over the past decade, from about 4% in 2000 to nearly 70% in 2011 (6). Although this is quite an increase, the latest US Census Population Surveys do suggest there is still a gap in home Internet access. Current Population Survey data from 2003 to 2011 demonstrate a persistent 12–13 percentage point gap in broadband adoption rates between metropolitan areas and nonmetropolitan households (6). Depending on the demographic characteristics of the survey participants in particular studies, it could also be useful to focus more specifically on the Internet access that is available to particular age groups as well as the race and/or ethnic background of the householder. Such data is available in the US Census American Community Survey Reports. As of 2014, it was reported that for individuals in the age range of 15–34 years, Internet access was available to 77.4% of households. In addition, Internet access was available to 76.2% of White-only households, 60.6% of Black-only households, 86% of Asian-only households, and 65.9% of Hispanic of any race households [6]. These data could be helpful in scanning a prospective survey participant population to determine the probability of using a Web-based self-report survey outside the clinic. In a study that compared adolescent survey completion via telephone versus web-based, overall 41.5% completed the survey online as compared to 59.8% via the telephone interview [7]. This finding also indicates that considering the method for administering an adolescent survey may be valuable, rather than assuming that a web-based approach is optimal. Finally, the surveys were conducted at 12 locations across the United States. The sites may have implemented the study differently and it may have been easier for them to manage reimbursement and retention if participants completed the survey in the clinic. This may have had an effect on the survey data discussed above.

Conclusion

Public health researchers, particularly in the social science and epidemiology arenas, continue to consider technologies that would aid in obtaining more accurate response options when doing self-report surveys. There is ample research on conducting Web-based surveys, but more knowledge and research into self-report public health surveys is needed.

We did find that the majority of our surveys were completed on Windows-based computers with a corresponding Internet Explorer browser. For participants who needed more than one visit to complete a survey, the largest percentage went from a Windows-based computer running Internet Explorer to an Apple OS (either Mac OS or iOS) running Safari.

For future projects requiring the use of a self-report survey that allows respondents to use their own personal devices (BYOD), we would consider adding several new elements to the solution. These changes or additions would include: better logging of the exact browser, version, and operating system used; the time it took to respond to each question; access to the Internet and device access away from the clinic for participants; and individual participant demographics.

Acknowledgments

We thank Sarah Thornton and the ATN Data and Operations Center at Westat for their collaboration in setting up the operational data collection process at the many sites involved in this research study. We acknowledge the contribution of the investigators and staff at the following sites that participated in this study: University of South Florida, Tampa (Emmanuel, Straub, Enriquez-Bruce), Children's Hospital of Los Angeles (Belzer, Tucker), Children's National Medical Center (D'Angelo, Trexler), Children's Hospital of Philadelphia (Douglas, Tanney), John H. Stroger Jr. Hospital of Cook County and the Ruth M. Rothstein CORE Center (Martinez, Henry-Reid, Bojan), Tulane University Health Sciences Center (Abdalian, Kozina), University of Miami School of Medicine (Friedman, Maturo), St. Jude's Children's Research Hospital (Flynn, Dillard), Baylor College of Medicine, Texas Children’s Hospital (Paul, Head); Wayne State University (Secord, Outlaw, Cromer); Johns Hopkins University School of Medicine (Agwu, Sanders, Anderson); The Fenway Institute (Mayer, Dormitzer); and University of Colorado (Reirden, Chambers). We would like to acknowledge Irene Friedland, at the Population Council, for the thorough edit of the paper she provided. The comments and views of the authors do not necessarily represent the views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The study was scientifically reviewed by the ATN’s Community Prevention Leadership Group. Network, scientific and logistical support was provided by the ATN Coordinating Center (Wilson, Partlow) at The University of Alabama at Birmingham. The investigators are grateful to the members of the local youth Community Advisory Boards for their insight and counsel and are indebted to the youth who participated in this study. This work was supported by the Adolescent Medicine Trials Network for HIV/AIDS Interventions (ATN) and NIH support, Bill Kapogiannis and with supplemental funding from NIDA and NIMH Grant NICHD 5 U01 HD 40533 and 5 U01 HD 40474.

References

1. Savel, C., Mierzwa, S., Gorbach, P., et al. 2014. Web-based, mobile-device friendly, self-report survey system incorporating avatars and gaming console techniques. *Online J Public Health Inform*. 6(2), •••. PubMed

2. Mierzwa, S., Souidi, S., Friedland, I., et al. 2013. "Approaches that will yield greater success when implementing self-administered electronic data capture ICT systems in the developing world with an illiterate or semi-literate population." New York: Population Council.

3. Bonn, S.E., Trolle Lagerros, Y., and Bälter, K. 2013. How valid are web-based self-reports of weight? *J Med Internet Res*. 15(4), e52. doi:http://dx.doi.org/10.2196/jmir.2393. PubMed

4. Touvier M., Méjean, C., Kesse-Guyot, E., et al. 2010. Comparison between web-based and paper versions of a self-administered anthropometric questionnaire. *Eur J Epidemiol*. 25(5), 287-96. doi:http://dx.doi.org/10.1007/s10654-010-9433-9. PubMed

5. Groves, Karl. 2007. "The limitations of server log files for usability analysis," boxesandarrows. http://boxesandarrows.com/the-limitations-of-server-log-files-for-usability-analysis/

6. File, Thom and Ryan, Camille. 2014. "Computer and Internet use in the United States: 2013," American Community Survey Reports. United States Census Bureau, ACS-28.

7. Rivara, Frederick P., Koepsell, Thomas D., Wang, Jin, et al. 2011. Comparison of telephone with world wide web-based responses by parents and teens to a follow-up survey after injury. *Health Serv Res*. doi:10.1111/j.1475-6773.2010.01236.x.

8. Whitacre, Brian, Strover, Sharon, and Gallardo, Roberto. 2015. How much does broadband infrastructure matter? Decomposing the metro-non-metro adoption gap with the help of the National Broadband Map. *Gov Inf Q*. 32, 261-69. http://dx.doi.org/10.1016/j.giq.2015.03.002