Data Collection in the Information Age:

Important considerations while constructing

secure web-based surveys

Matthew J. Ross

University of Connecticut, Neag School of Education
Abstract

The world of research is slowly waking to the dawn of the new technology age and its many benefits as well as drawbacks. One of the most eminent issues in the broad technology arena is security. This “problem” is not confined to corporate America; it extends to the world of research as well, and it is only now becoming a hot topic for many Institutional Review Boards. Security has become an issue in the collection, storage, sharing, and destroying of data. Security of personal identity from the researcher’s viewpoint as well as the participant’s web-based surveys is a major concern of many potential participants. The subject to be focused on is the important considerations when designing secure web-based data collection solutions. Reliability and validity, as these both play important roles in decisions about data collection, are also discussed.
Data Collection in the Information Age:

Important considerations while constructing secure web-based surveys

A survey is a document that provides a researcher with a “detailed critical inspection” of a particular subject or event (WordNet, 2004). Research surveys have traditionally been administered orally or in written format. With the continued growth of availability to the Internet in all levels of society, and the innovation of newer and easier programming solutions, an increasing number of surveys are being administered online.

Web-based surveys contain the same elements as their written/oral predecessors. Questions can be formed using multiple choice, quantitative responses, and qualitative responses. Answers are entered using text boxes, text fields, radio buttons, check boxes, and list boxes/menus. The many methods with which to phrase and answer questions give the researcher a multitude of ways to ask key questions and in some cases multiple times.

Surveys are typically designed with the following components: an opening page containing information about the research being done, a consent form, and the survey questions. In the case of some web-based surveys, an identifier page (establishing the identity of the participant) is also included. Before a participant begins the survey, he/she must consent to participating in the research.

Early forms of web-based surveys used email as the means of data collection. Survey results were returned to the researcher in the body of an email. This presented a very time consuming task for the researcher, as the data was then transferred to an analysis tool. The chance of error in the transfer was great, and the chance for lost data due to transmission errors was greater. Today, many web-based surveys are designed as a front-end to a database.
database is a collection of tables relating to a particular function. The tables contain rows (also known as records or cases) and columns (also known as fields or variables). By entering the data directly into a table within a database file, the researcher can port the data set directly to an analysis tool (SPSS, Excel, etc.) without error.

Obtaining Consent

In conducting any research, the first element is the consent form. Obtaining informed consent is essential to successful research. “[Consent] is more than simply having a potential research participant sign a consent form; it is a process by which necessary information is communicated to the participant by the researcher.” (Sales, Folkman, 2000) Obtaining consent can be difficult when administering web-based surveys, dependant on the participant pool. Obtaining the consent of a minor can be very challenging when done over the web. A parent or guardian must also consent to the child participating. Many commercial websites avoid underage populations by requiring a valid credit card number be entered. Although this assumes that every person over the age of 18 has a valid credit card. It also does not embrace the growing trend for ATM/Debit cards and the growing number of under-18 year olds with these cards. Disclosure statements are also another effective way of dealing with underage issues.

On the other hand, obtaining consent of an adult by means of the Internet is much easier than the inconvenience of arranging a meeting, accommodating travel time, and other interference. A researcher can place a consent form for any study on the web, and ask participants to review and accept/decline far in advance of any research activity. The researcher might also have a brief pre-screening survey that can be administered prior to any visits, for time saving analysis.
A person’s identity is protected by the type of consent that person gives. When designing web-based surveys, one method of obtaining consent is more challenging than the other. The easier of the two methods of obtaining consent is consent with confidentiality. This method allows the programmer to attach a true, unique identifier to the data collected. The agreement between the participants and the researcher bars the researcher from releasing that identifier to any other source. It merely acts as a means for follow-up. This unique identifier might be the participants’ social security number, or piece of that number, a unique id associated with the participants employer, university, or school, or some other pre-arranged identifier. This unique identifier can be used to prevent the participant from returning to participate a second or third time, especially in web-survey situations where data is collected only once. This construct must be programmed carefully, as will be discussed later.

Anonymity is the harder of the two methods of consent to accommodate as a programmer. The identity of the participant must be kept separate from the data. This involves a few additional lines of code in order to ensure that the being entered by the participant is kept together without tying the response to a personal identifier of the participant in any way. A pseudo-identifier must be developed by the programmer to accomplish this. The most common problem with data collection under the constraint of anonymity is the ability to discern that a participant has not previously participated, and/or is not misrepresenting him/herself.

One design, used by the staff at the Neag School of Education, for informed consent can be applied to either confidential or anonymous consent. In this design, a unique pseudo-id is set for the participant. In a confidential and some anonymous situations a unique personal identifier is also collected; however, it is not stored until the data has been completely collected. The method of informed consent determines how that identifier is stored. In a model where
confidentiality is the method of consent, the identifier is stored with the data after the survey has been administered (Diagram 1).

The personal identifier is needed at the beginning to determine whether the participant has already taken the survey. A participant taking a survey more than once can pose a threat to the integrity of the data.

In the case of collecting data anonymously, the data collection can operate in two ways. As mentioned previously, one of the key problems with anonymous data collection is ensuring data integrity. If identity is completely anonymous, the researcher can not check the data to make sure a participant did not complete the survey multiple times.
Diagram 2a shows the personal identifier being collected and checked against a database table, similar to the method used for confidential data collection. However, later in the process, the personal identifier is stored in a separate database table, rather than with the data. This identifier is also never stored with a time/date stamp. This new table is that which the personal identifier is check against earlier in the process. Because the identifier is stored in a separate table, the researcher can not connect the data to the participant. The personal identifier table can also be destroyed after the data collection has occurred, further protecting the participant.
Diagram 2b shows the flow of data collection when no personal identifier is being collected. This method should be reserved only for extreme cases of anonymity, as data integrity can not be determined unless the participant is taking the survey in a controlled situation.

The code for these designs is rather simple. A sample of code used for a Professional Development School Teacher Evaluation is given in Appendix A. This example is for confidential data applications. It shows the personal identifier being collected along with the program of study. In the action, the code checks that personal identifier against a separate table containing all the personal identifiers of completed surveys (table PDS_Teacher_ID). If the identifier already exists, the participant is directed to an error page, and then to an alternate website. If the identifier does not exist in the second table the participant continues with the rest of the survey.
This design is important to prevent participant misrepresentation, intentionally bad data, and possible data manipulation. By eliminating the possibility of a person completing the survey more than once or intentionally taking the survey multiple times to skew the data, the survey instrument is more reliable and valid. It also protects the participant by making the data truly confidential or anonymous, depending on the application. LAN administrators and programmers can create databases that encrypt the personal identifier making it impossible to read the identity even if the database table is shown to the researcher or another person. The personal identifier would show as a series of *******.

Reliability and Validity

The next critical piece to the survey instrument design is the validity of data entered or collected throughout the body of the survey. The continuous development of new client-side programming languages (those incorporated into the user's browser software) such as JavaScript provides for the validation of participant data without a need for additional server transactions. This technology provides a vehicle for the minimization of incomplete or erroneous data sets. Server-side languages (those hosted on the web server) such as Perl and CGI provide for the manipulation of data to such a degree that files can be prepared for analysis with little to no researcher intervention, eliminating time consuming and error-prone pre-analysis processing by hand. (White, Carey, Dailey, 2001)

The common mistake made by many researchers when creating web-based surveys is the use of client-side scripts programmed to force the participant’s completion of all data fields. If all the fields are not completed the form can not be submitted. In the USF study, this resulted in two reflections. “First of all, there is an ethical issue. All participants have the right to withdraw from participation in a research study. On paper surveys, participants often exercise the right to
leave specific data fields blank. These same privileges should exist for web surveys.” The second issue brought to light was also with regard to the completion of data fields. “From a practical perspective, failure to allow the participant to omit a field leaves only the option of total withdrawal from the study. On the chance that partial data may still be useful for some purposes, most researchers will probably prefer a partial data set to none at all, reserving the option to eliminate data sets that are not useful.” (White, Carey, Dailey, 2001)

Instead, researchers and programmers should validate the type of data entered. For example, in a field requiring a participant’s date of birth, the researcher would not want text. Typically the date of birth is given in MMDDYY format. The programmer should validate that the information being entered is that format and is numeric rather than textual.

Cost-Benefit Analysis

The benefits of administering a survey via a web-based forum far outweigh the cost to the participant and the researcher. The benefit to the participant is the ability to participate in web-based research at his/her own leisure. In the case of the researcher, the initial cost of the research development and design, and the survey instrument designs are still apparent. However, the need for additional help to collect, aggregate, and analyze the data once collected is no longer needed. Instead, the data is stored to a concise database, which can then be ported to a statistical analysis program. If configured properly, data is much safer stored on a hardware device than being stored in paper format, portable media, etc. The time savings is also a great factor in reducing the cost.

Overall, the cost of hardware is significantly down compared to costs 10 years ago. In fact, most institutions already have data equipment capable of hosting such web-based surveys for small to medium size research projects. Furthermore, the researcher can reach a much
broader audience by means of the web, than what was typically available in a community, university setting, etc. Research shows that about 51 million households currently have internet access, an increase from approximately 11 million in 1994. (UVA, 2004) A report by the Strategis Group predicts “more than 90 million households to be on the Internet in five years. Internet penetration of businesses reached 6.3 million in 1999 and will rise to 8.3 million by 2004. The Strategis Group foresees a combined total of 171 million Internet users in 2004.” (Pastore, 2000) As the numbers grow, so does the potential participant pool for researchers.

**Conclusion**

As technology advances, the research academy will begin to tackle a new forum for the same critical issues dealt with decades ago. Survey administration is not a new subject; web-based surveys are a different medium for administration. How will participants’ rights be protected? How will research be conducted in a fair, reliable way, while not hindering the overall scholarship of the researcher? Institutional Review Boards must confront the reality of advancing technology and all it has to offer the world of research. Continuing advances in technology are not only the problems, but the answers to the problems. If carefully implemented, the research world can reap the benefits of online survey administration many times over.
References


Appendix A

The following pages follow the structure in Diagram 1 (p. 6). The consent page is given in view format, followed by the HTML/CFML code that interprets the data to determine whether the participant has responded previously.

Professional Development School Teacher Evaluation Form

CONSENT FOR PARTICIPATION IN A RESEARCH PROJECT
University of Connecticut

Contact Name: Jacqueline Kehler
Study Title: "Professional Development School Teacher Evaluation"

Invitation to Participate

You are being asked to participate in this evaluation of the Professional Development School clinic placement. All Neag School of Education, Teacher Preparation Program are being asked to participate in this evaluation.

Purpose

The data collected from this survey will be used to make recommendations to the Teacher Candidate Assessment Committee and to the Neag School of Education, Dean's Office for programmatic improvements.

Description of Procedures

You will complete the attached electronic survey and submit it back to the researcher electronically. It is anticipated that completion of this form will take between 10 and 15 minutes.

Risks and Inconveniences

There are no anticipated risks associated with participation in this study. An inconvenience to you for participation may be the 10 to 15 minutes required to complete the survey.

Please enter your NetID:

Program: First-year IBM (Junior)

I Agree
<cfparam name="FORM.STU_ID" default="1">
<cfquery name="qGetMasterID" datasource="evaluations">
SELECT * 
FROM PDS_Teacher_ID 
WHERE STU_ID = '#FORM.STU_ID#'
</cfquery>
<html>
<head>
<title>Processing...</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
</head>
<CFSET #form.YEAR# = "2004">
<cfquery name="StartRecord" datasource="evaluations">
    INSERT INTO PDS_Teacher
    (STU_ID,PROGRAM,YEAR)
    VALUES
    ('#Form.STU_ID#', #Form.PROGRAM#, '#Form.YEAR#')
</cfquery>
<body>
<CFIF qGetMasterID.RecordCount EQ "0">
    <cflocation url="page1.cfm?STU_ID=#FORM.STU_ID#">
<CFELSE>
    <cflocation url="error.cfm?STU_ID=#FORM.STU_ID#">
</CFIF>
</body>
<html>