

Using Client-Side Event Logging and Path Tracing to Assess and Improve the Quality of Web-Based Surveys

Thomas M. White, MD, MS, MA¹ and Michael J. Hauan, MD, MPH, MTS, MA²,

¹Bureau of Evidence Based Medicine and Practice Guidelines, New York State Office of Mental Health, New York, NY, ²Department of Health Management & Informatics, University of Missouri, Columbia, MO

Web-based data collection has considerable appeal. However, the quality of data collected using such instruments is often questionable. There can be systematic problems with the wording of the surveys, and/or the means with which they are deployed. In unsupervised data collection, there are also concerns about whether subjects understand the questions, and whether they are answering honestly. This paper presents a schema for using client-side timestamps and traces of subjects' paths through instruments to detect problems with the definition of instruments and their deployment. We discuss two large, anonymous, web-based, medical surveys as examples of the utility of this approach.

INTRODUCTION AND BACKGROUND

The quality of data collected is limited by both the reliability and validity of the instruments themselves, and by the process used to collect the data. However, assessing the quality of instruments and their deployment is challenging, and seldom done. Since much of the data used in evidence-based medicine comes from surveys, assessment instruments, and structured interviews, there is a need for faster and easier techniques for assessing and improving their quality.

According to psychometric theory¹, even minor changes in the wording, formatting or order of asking questions can affect the reliability and validity of an instrument. Measurement studies can assess the psychometrics of instruments, and compare them with different instruments or versions. Unfortunately, such studies are seldom done. Moreover, these psychometrics may change when instruments are deployed in novel settings, languages, or populations.

Web-based data collection poses additional threats to reliability, since the user experience is influenced by variations in processor speed, screen size, operating systems, browsers, font size, colors, and network bandwidth and reliability². Some studies have shown that small formatting changes can alter results^{3, 4}, while others are finding that formatting can have little impact⁵. There are still few heuristics for predicting the impact of changes in formatting, and whether those impacts might vary by target population.

Threats to data quality can be divided into two types: those related to the *definition* of the instrument itself, and those related to the *deployment* of the instrument. As

shown in Table 1, the definition of an instrument includes the wording and order of questions, and whether they are understandable. The data quality can also be affected by the deployment of the instrument. Changes in user interface, hardware, screen size, colors, fonts, and other presentation attributes can affect subjects' ability to read questions and navigate through surveys. Moreover, in interviewer-assisted instruments, the interviewers can bias the data collection. Table 1 also shows that data quality questions can be asked at both individual and aggregate levels, thus allowing for the detection of systemic problems, as well as problems in individual data sets.

The quality of individual answers is affected by subject opportunity, ability, and motivation⁶. Deployment problems like slow connections and poor user interfaces can reduce subject opportunity and ability, leading to nonresponses, dropouts, and aggravation. Poorly worded or designed surveys can also reduce subjects' ability to give the desired answer, as well as affect motivation.

People who complete anonymous web-based surveys have a range of motivations, not all of them altruistic. Some people might answer randomly while testing or exploring an instrument. Others complete surveys multiple times to bias the results. Many may want to answer honestly, but are threatened by sensitive questions, or concerns about anonymity. Some might misunderstand questions due to poor wording, or poor choices of colors and fonts. Lastly, some might mark incorrect answers, either because of confusing answer options, overly restricted choices, or poor user interfaces.

Although the rate of random or inaccurate responses may be low, and differ little from what would be seen on a paper-based interview⁷, random responses due to uncertainty or desire to avoid discomfort can be high. A recent study⁸ showed that between 31% and 99% of subjects expressed opinions on fictitious issues, rather than indicate that they were unfamiliar with the topic.

Epidemiological and public health research tends to eschew web-based research due to these limitations. However, given the growing reach of the web, and its convenience, speed, and low cost, increasing amounts of medical research could be conducted via the web if the data quality issues could be improved.

	Aggregate	Individual
Target Question	What is the overall incidence of depression? What is the incidence by sub-populations?	Which subjects are depressed? Is Joe depressed?
Definition (clarity and sensitivity of questions)	Which questions are problematic? Are skipped or left unanswered? Have answers frequently changed? Take longer than expected to answer? Are answered too quickly? Cause people to quit the survey?	Identify individuals who had trouble Who skipped or changed many answers? Who stopped prematurely? Who answered too quickly / slowly? Whose answering speed varied over time? Which questions did Joe find problematic?
Deployment (effects of user interface, order, browser, network, operating system, and hardware)	Was the deployment adequate? Average server and network delays? Which pages took longest to load/draw? Did adequacy vary with time of day? Does user interface affect completion rates, speed, satisfaction, answer distribution? Number of questions per screen? Fonts, colors, layout, question order?	Identify individuals with deployment problems Who had excessive time between pages? Were they more likely to drop out?

Table 1. Questions relating to the quality of instrument definition and deployment that can be addressed using the Dialogix path and timestamp data.

In interview-administered surveys, subjects' uncertainty or confusion is manifested in their comments, vacillating responses, and changed answers⁹. We hypothesized that we could use path tracing to detect similar behaviors. Likewise, we hypothesized that we could detect lack of cooperation on uncomfortable questions by identifying questions that were answered too quickly, or that subjects attempted to skip.

Using two anonymous web surveys as examples, this paper will present generalizable methods for detecting problems with web-based surveys, and for identifying response patterns with questionable data quality.

METHODS

MUSecuritySurvey

The MUSecuritySurvey instrument was used in a study to assess the misuse of and attitudes towards patient-identifiable health information¹⁰. All 35,000 faculty, students, and employees at University Missouri-Columbia were sent an email inviting them to participate in this anonymous survey, which was available for two weeks. Of these, 2558 subjects started the survey, and 1765 completed it.

Since we plan to conduct a multi-center study of these issues, we wanted to assess and improve the quality of this instrument. We hypothesized that subjects might resist answering sensitive questions by quitting the survey at that point, skipping those questions, or vacillating when selecting an answer. Still others might pick random answers. We used event logging to help identify problem questions and potentially spurious answers.

AutoMEQ

The AutoMEQ is a consumer-oriented decision support tool that asks subjects about their sleeping habits and tells those with winter depression when to use light therapy for

optimum benefit. The AutoMEQ is advertised on a not-for-profit web site providing education for people interested in Seasonal Affective Disorder¹¹. Between November 2001 and January 2002, 1090 people started the AutoMEQ, and 359 (33%) people completed it.

Since the AutoMEQ is meant to be a key component of an anonymous, international, multi-lingual, epidemiological study, we need to optimize its quality, standardize it, ensure that all translations are equivalent, and minimize the impact of the use of anonymous subjects. This entailed assessing why people did not complete it, and identifying potential sources of unreliable data.

In addition to monitoring event logs, we embedded a check for inconsistent answers. The AutoMEQ contains 19 questions that comprise a scale¹² that measures the degree to which subjects are morning or evening types (larks or owls). If subjects' get an intermediate score, and their score is a mix of strongly lark and owl attributes, then either they answered dishonestly, questions were confusing, or they represent an under-studied population. Subjects who fit this pattern trigger an alert question that gives them the option to change their responses.

Deployment

Both instruments were authored and deployed using Dialogix¹³, which uses a custom interpreter and Java servlets to read an abstract definition of an instrument and manage the data collection. We previously described the schema for the content and process data collected via Dialogix¹⁴. The client-side event logging makes it possible to determine how long the browser spent drawing the page, how much time subjects spent interacting with each question, and whether they skipped questions or changed their answers before moving on to the next page. These logs also assess network delays, and detect when subjects backtracked to change answers on previous pages.

Analyses

Perl programs packaged with Dialogix aggregate and transform the log and data files into views that can be readily imported into and analyzed by the commercial statistical package SPSS (SPSS Inc, Illinois). These programs generate scripts that create the databases and data dictionaries, and include a set of standard analyses to address many of the questions listed on Table 1.

One table includes path information for each question. These data include the number of times each question was viewed, skipped and answered; how many times the answer was changed; and the history of answers given. They detect whether subjects vacillated between several answer options before making their selection; record the path the subject took after answering the questions (e.g. whether they pressed “next” or “back”); and whether answers became applicable or inapplicable due to backtracking and changes in branching patterns. They also indicate which group of questions were the last viewed and answered, thus identifying questions that correlate with non-completion of the instrument. Thus, these data address problems in instrument definition.

The second table includes timing information for each question. It records the amount of time spent reading and answering each question, and the number of characters in each question and associated answer options, thus allowing the subjects’ response times to be normalized for the amount of content viewed. Associated reports determine the mean and median response times per question, and identify outlier questions that are answered two quickly or slowly. The overall mean normalized response time for all questions is used to identify outlier questions in aggregate. These data can also be used to detect trends in answering speed, such as a fatigue-induced slowing, or aggravation-induced speeding up of answering near the end of long instruments. Thus, these data can answer the time-related questions about definition quality listed Table 1.

The final table indicates the per-screen network, browser, and server delays. The mean values for these can be used to assess the adequacy of the deployment, and identify individuals whose browsers or operating systems are causing excessive delays. The impact of user interface upon deployment can be studied by analyzing the timestamp and path logs of parallel versions of surveys that only differ by a single user interface attribute.

RESULTS

MUSecuritySurvey

Non-completers: The most common places for subjects to stop answering the survey were the last screen, which asked subjects’ for their opinions about patient-identifiable health information (PHI); the first screen, which introduced the survey; and the seventh screen, which gave a verbose definition of PHI. The non-

responses after the seventh screen suggest that more lay-language should be used there. The low answering rate of the opinions mirrors the comments we received from non-medical subjects who had no opinion on these issues.

Our concern that subjects might refuse to answer sensitive questions about misusing passwords was not confirmed. Only 12 of the 345 subjects who were asked questions about misusing passwords chose to terminate the interview rather than select an answer. However, of the 868 asked about password misuse, there were 47 cases where subjects tried to skip the question before being told that they must answer.

Problem Questions: Several subjects commented that the answer choices for some questions were too restrictive. For example, they asked that an “other” category be added to the departmental affiliation question. We hypothesized that subjects who are dissatisfied with the available choices might vacillate when selecting an answer, backtrack to change an answer, or take longer than usual to select an answer. Of 1882 people who were asked this question, 113 (6%) vacillated before selecting a response, and 164 (8.7%) tried to skip it. Changed, vacillated, and skipped answers were also common for the opinion questions; for which several subject asked that we add a “no opinion” option. By contrast, the incidence of this behavior on other questions was ten fold less. These results affirm our strategy of using these analyses to detect potential problem questions.

Deployment Problems: The median server delay between questions was 29 milliseconds. There were 10 outliers with delays of 3 to 13 seconds, four of which were from non-completers. Upon examining the logs, we discovered that these tended to be the subjects who first used the system after the nightly log rotations. Thus, we should pre-load the servlets before the first user tries to access the system.

AutoMEQ

Non-completers: Sixty-seven percent of people who started the AutoMEQ did not complete it. Of those 730 people, 167 (23%) stopped after the introduction, 151 (20%) stopped before answering all 19 questions, with 49 of those stopping after the first question. The remaining 57% stopped after the second-to-last page, which is a tailored report of their results. Examining the wording of that page makes it clear why they stopped. Rather than being told to press “next” to complete the instrument, they are asked to press “next” if they would like a printout of the answers they gave. We hypothesize that we can improve completion rates by re-wording these instructions to make pressing “next” less optional.

Identifying inconsistent answers: Twenty percent (154) of the 747 subjects who answered all of the questions entered inconsistent answers, showing a mix of strongly lark and owl traits. Of these, 60 subjects elected to review their answers, and 29 of them changed their

answers such that they became consistent, making the trigger question inapplicable.

The visualizations in Figures 1 and 2 can be used to further assess the honesty of subjects who asserted that their inconsistent answers truly reflect their circadian rhythms. The typical respondent answers all 19 questions in sequence, does not trigger the question detecting inconsistent answers, and takes between 5 and 15 minutes to complete the instrument and review their results. Figure 1 shows four paths that mock subjects took through the instrument. The Y-axis represents the question within the instrument, and the X-axis shows how many screens-full of questions the subject has seen. The blue, violet, and red lines represent subjects who answered all questions in sequence. Unlike the blue line, the purple and red lines triggered the consistency alert. The purple subject claimed that its answers were accurate, and was allowed to continue. The red subject, however, elected to return to the beginning and change its answers, after which the consistency alert was no longer triggered.

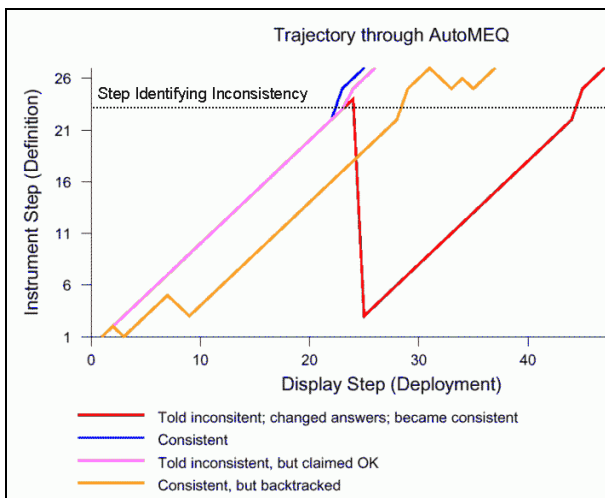


Figure 1: Visualizing the path through an instrument.

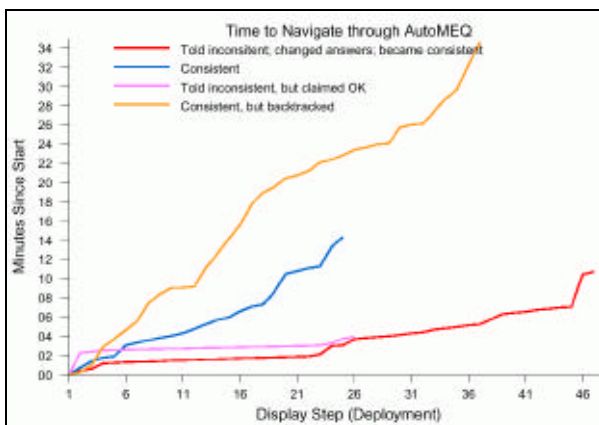


Figure 2: Visualizing slow and fast responses.

We can use timing information to try to distinguish between subjects who are just testing the system, and those who might have unusual circadian rhythms. Figure

2 shows how long it took for each of the subjects from Figure 1 to complete the AutoMEQ. The blue subject took an expected amount of time. The violet and red subjects, however, took less than two minutes to complete the 19 questions (after reading the introduction pages). It is not realistic for a subject to answer each question in 7 seconds. Thus, these answers must have been spurious, as also detected by the consistency trigger. Interestingly the red subject took five minutes when changing his answers, which is within the normal range, so the changed answers are more likely to be honest. Finally, the orange subject took an inordinate amount of time to complete the AutoMEQ. We could return to the raw data to determine whether this was caused by network or browser delays; and identify whether English is her primary language from her browser identification. Subjects who take an appropriate amount of time to answer the questions, trigger the consistency alert, yet claim that this represents their true patterns, warrant further study.

DISCUSSION

These results demonstrate how timing and path information can be used to detect systematic problems with the definition and deployment of instruments, as well as subjects who appear to be entering spurious answers.

There are hundred of published studies that discuss the promise and challenges of web-based research¹⁵. Some mention that timestamps can be used to detect spurious answers¹⁶. However, to the best of our knowledge, ours is the first to use client-side event logging to remove the bias of network and browser delays from these timestamps. Moreover, our approach can normalize for the amount of content displayed, even when that content is tailored. Finally, we believe this is the first example of the use of path information (like vacillation and changed answers) to detect problems in the definition of questions, and in individual answers to those questions.

Path and timing analyses are crucial for one epidemiological study that is using Dialogix¹⁷. It is assessing antisocial behavior in Iland and U.S. Puerto Rican youth, since the phenomenology of disease appears different in the two locations. Whether this is a true difference, or an artifact of problems with the instrument and/or deployment is one of the key questions. The timing and path information collected by Dialogix are being used to help resolve this issue.

This approach for identifying problem questions can also speed and enhance the process of standardizing medical assessment instruments. Moreover, it can aide efforts to detect and code potential differences among related versions of those instruments¹⁸. These tools can also facilitate ongoing psychometric analysis of instruments as they are deployed in new populations, settings, and languages. Thus, these techniques may take us a step further towards web-based epidemiological and public health studies.

This approach is generalizable to a broad range of web-based instruments. For example, our IRB has proposed studying how long subjects spend reading different sections of consent forms so that the wording and organization can be improved. The ability to detect changed answers and backtracking can facilitate web-based testing in which the authors wish to give subjects the correct answers at the end. Web-based testing is difficult without such logging¹⁹. These logs can also be used to assess how subjects navigate through tailored medical education sites, and thus help identify pages needing revision.

These results are limited by the fact that the theoretical relationships between timing, motivation, and validity are still in the formative stages. For example, what is an appropriate threshold for stating that vacillating data entry or changed answers may reflect a true problem with individual data? Such thresholds may differ for questions about demographics, history, and opinion, and would need empirical validation. Likewise, what is the threshold, if any, for determining that a question was answered too quickly; and how crucial is content analysis to specifying that threshold? Modeling is needed to determine whether simple character counts are adequate, or whether sentence analyses is needed. There are also many opportunities for exploring the effect of user interface upon these values, from which heuristics might be developed.

Our efforts would also benefit from more sophisticated and user-friendly visualization tools, like those found in some advanced systems for analyzing user-interface event logs^{20, 21}. Such tools support 3D views of paths through static web sites, but they do not work with dynamically generated web-sites, like those needed for complex surveys²². Once adapted to support dynamic content, such tools could greatly help efforts to improve the quality of web-based surveys.

CONCLUSION

Web-enabled surveys have questionable validity until they have been standardized. Further, web based data collection can be polluted by subjects who answer dishonestly or are confused. This paper presents the first demonstration of how timestamp and path tracing can be used to detect systemic problems with the wording or ordering of questions within instruments, even after they have been deployed. Such an approach can facilitate the longitudinal psychometric analysis of instruments as they are deployed in new settings, languages, and populations. This paper also demonstrates how the same data can be used to detect potentially spurious answers on the part of respondents to anonymous surveys. Together, these approaches can help improve the quality and availability of assessment instruments, and extend efforts to make web-based surveying a viable public health and epidemiological technique.

REFERENCES

1. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York: McGraw-Hill; 1994.
2. Couper M. Web surveys: a review of issues and approaches. *Public Opin Q* 2000;64(4):464-94.
3. Wyatt J. Same information, different decisions: format counts. Format as well as content matters in clinical information. *Bmj* 1999;318(7197):1501-2.
4. Couper MP. Web Survey Design and Administration. *Public Opin Q* 2001;65(2):230-253.
5. Bell DS, Mangione CM, Kahn CE, Jr. Randomized testing of alternative survey formats using anonymous volunteers on the World Wide Web. *J Am Med Inform Assoc* 2001;8(6):616-20.
6. Groves RM, Cialdini RB, Couper MP. Understanding the decision to participate in a survey. *Public Opinion Quarterly* 1992;56:475-495.
7. Pettit FA. A comparison of World-Wide Web and paper-and-pencil personality questionnaires. *Behav Res Methods Instrum Comput* 2002;34(1):50-54.
8. Graeff TR. Uninformed response bias in telephone surveys. *Journal Of Business Research* 2002;55:251-259.
9. Mathiowetz NA. Respondent Expressions of Uncertainty. *Public Opinion Quarterly* 1998;62:47-56.
10. Hauan MJ, Patrick TB, White TM. The individual perspective on patient-identifiable health information in a large midwestern university. *Proc AMIA Symp* submitted.
11. Terman M, White TM, Jacobs J. Automated Morningness-Eveningness Questionnaire Self Assessment Version (AutoMEQ-SA). In: <http://www.cet.org/AutoMEQ.htm>; 2001.
12. Horne JA, Ostberg O. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *Int J Chronobiol* 1976;4(2):97-110.
13. White TM, Hauan MJ. Dialogix: a System for Rapidly Developing, Deploying, and Analyzing Research Studies. In: <http://www.dianexus.org:8080/>; 2001.
14. White TM, Hauan MJ. The capture and use of detailed process information in the dialogix system for structured web-based interactions. *Proc AMIA Symp* 2001:761-5.
15. Kaye BK, Johnson TJ. Research Methodology: Taming the Cyber Frontier. *Social Science Computer Review* 1999;17:323-337.
16. Eysenbach G, Wyatt J. Facilitating Research (Chapter 6.3). In: McKenzie B, editor. *Internet and Medicine* (3rd edition); Oxford University Press; in press. p. 211-225.
17. Bird HR, Canino G. Boricua Youth Study (BYS). In: *Antisocial Behaviors in US and Island Puerto Rican Youth* (5/1/98 - 2/8/03); NIMH Grant MH56401; 2000.
18. White TM. Extending the LOINC Conceptual Schema to Support Standardized Assessment Instruments. *JAMIA* in press.
19. Schleyer TK, Forrest JL. Methods for the design and administration of web-based surveys. *J Am Med Inform Assoc* 2000;7(4):416-25.
20. Laskowski S. WebMetrics. In: NIST: <http://zing.ncs.l.nist.gov/WebTools/>; 2001.
21. Etgen M, Cantor J. What does getting WET (web event-logging tool) mean for web usability. In: *Proceedings of the Fifth Conference on Human Factors and the Web*; 1999; Gaithersburg, MD; 1999.
22. Ivory MY, Hearst MA. The State of the Art in Automating Usability Evaluation of User Interfaces. *ACM Computing Surveys* 2001;33(4):470-516.