# Interview Timing Data:
# Simple yet Powerful Survey Instrument Development Tools

Ruth Heuer
John Doherty
Eric Zwieg

RTI International
Research Triangle Park, North Carolina

Presented at the 62nd Annual Conference
of the American Association for Public Opinion Research (AAPOR)

Anaheim, California

May 17-20, 2007

**Abstract:** Timing data in a computer-assisted telephone interview (CATI)/Web survey is not difficult to collect and is quite useful for subsequent questionnaire development. Time stamp information for predictable interview events (such as login/logout, form submission, interview completion) can be collected in several ways. Each provides different insights into the interview experience, and when used together, they present instrument designers with tools to fine-tune future instruments. Based on strategies used in numerous large-scale studies for the U.S. Department of Education, National Center for Education Statistics, this paper explores all aspects of implementing timers in a Web-based instrument as well as methodologies for analyzing and using those data to shape future questionnaire development. The different types of timers and strategies for determining when to implement each will be discussed. Methods of aggregating and analyzing timing information will be presented. When implemented and analyzed correctly, timing data contains substantially more information than simply the total interview completion times. For example, when designing future instruments that utilize the same or similar items, individual screen timers can be used to assist in predicting administration time and to provide good estimates of overall interview length. In addition, individual onscreen timing data coupled with data on help text usage or item-level missing data can be used to evaluate understandability of question or response-option wording and assist in determining response-option types (checkbox, radio button). Results from the National Postsecondary Student Aid Study (NPSAS), Beginning Postsecondary Students Longitudinal Study (BPS), Baccalaureate and Beyond Longitudinal Study (B&B), and National Study of Postsecondary Faculty (NSOPF) will be presented.

*RTI International is a trade name of Research Triangle Institute*

## Introduction

RTI International (RTI) has collected data for numerous large-scale postsecondary education studies for the National Center for Education Statistics (NCES), including the National Postsecondary Student Aid Study (NPSAS), Beginning Postsecondary Students Longitudinal Study (BPS), Baccalaureate and Beyond Longitudinal Study (B&B), and National Study of Postsecondary Faculty (NSOPF). Prior to 2002, the primary mode of data collection for these studies was computer-assisted telephone interview (CATI) with computer-assisted personal interview (CAPI) follow-up. With expanded capabilities of the Internet, Web-based self-administered data collection has become our primary mode with CATI and, if necessary, CAPI follow-up. This paper reports on our experiences with time stamps and the resulting timing data for NPSAS:04, BPS:04/06, B&B:93/03, and NSOPF:04.

To provide the client with information on respondent burden, our computer-assisted studies for NCES have always collected timing information as part of data collection (typically about 30-minute interviews), but the detailed timing data have not otherwise been used. Recently, however, we realized several uses for timers that allow us to better estimate the length of a questionnaire during the instrument design phase as well as to improve the quality of the data collected.

After a review of the timing data literature, this paper will present the technical details of timers, including definitions of client and server side, considerations for when to use each, and other issues related to implementation of timers. This will be followed by a discussion of the various uses of the timing data in questionnaire development.

## Literature Review

Timers have been used in survey research since the early 1970s. Psychologists have frequently used timers to measure reaction times to questions. For example, Fazio's research on response latencies in opinion research asserted that a subject's accessibility of an attitude could be calculated by measuring the time it took for the subject to answer a question (Fazio et al., 1982). From this he drew several conclusions on cognitive processes that bring about attitude judgments. Bassili (1996) developed a methodology for measuring response latency in CATI using a computer clock to measure the time between the end of the interviewer's question and the respondent's answer with millisecond accuracy. Mulligan et al. (2003) looked at methodologies for implementing and analyzing response latency. As a low-cost, low-maintenance alternative to other response latency timers they advocated the use of *latent* response latency timers (which are invisible both to respondents and interviewers) in interviews.

Bassili and Scott (1996) used response latency to identify bad questions in survey research. Comparing the time it took for respondents to answer three types of bad questions to the time it took for respondents to answer repaired versions of the questions, they found that bad questions took longer for a respondent to answer than the repaired questions.

Other than Bassili and Scott's research, little has been written on the use of timers for subsequent questionnaire design. This paper addresses this gap in the literature by providing the reader with three strategies that have been implemented in major data collections as well as providing background information on how to implement time stamps in a data collection instrument.

## Implementing Time Stamps

Time stamp information for predictable interview events can be collected either as a client-side timer or a server-side timer, depending on the event. Each provides different insights into the interview experience, and when used together, they present instrument designers with tools for future instrument development.

Web survey events that occur at the data collection site are referred to as "server-side," while "client-side" events happen at the respondent's computer. Examples of server events include when a user logs in or out of the survey, when particular data are stored, and interview completion. Client-side events include such things as the time a particular Web page is loaded in the user's Web browser and when a user clicks a particular form control—for example, a button or a hyperlink. The distinction between the two is important to understand because timers of different kinds cannot be combined for analysis.

Server-side time stamps use the time set on the data collection server. As a result, server-side time stamp information is easy to collect and can be relied upon to be consistent, not only for a particular respondent, but also across cases. These timers, therefore, can be reliably related across the entire sample. Client-side timers rely on the time set on the user's PC and as a result should remain consistent within a given case, but cannot be used across cases.

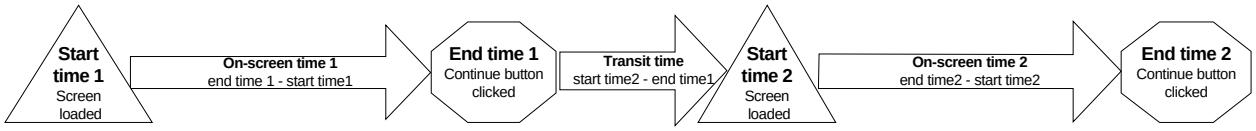
To collect client-side timers, programming code must be sent with the survey form code. Programming languages such as JavaScript are used to interact with the client's computer. Client-side timers present additional challenges, since each Web browser and even different versions of the same browser can implement these programming languages differently. This requires extensive cross-browser testing as well as cross-platform testing to ensure, as much as possible, that the data collected will be reliable. Even with this kind of testing, a user may choose a browser that was not tested, may have made changes to local settings, or may have adjusted the PC time during the survey—any of which can result in timing data that is not useful for analysis. Because of these issues, client-side timing data must be reviewed and cleaned before being analyzed.

In RTI's Web-based data collection, timing data is routinely collected for many events. Server-side event timers are collected at the case level and include timestamps for the first time the case is accessed, the last time the case is accessed, when the case completion status is last updated, and the date the case is completed. In addition, since respondents can complete the surveys over multiple sessions, login and logout timers are collected for each session.

Client-side timers for each case are collected and stored at the form level.[1] These timers are used to compute the onscreen time for each form in the instrument. Two different timestamps are collected with the response data for each form: a page start time and a page end time. The page load event represents the start time and is the time that the form was available for viewing, and the end time is determined when the form submit button is clicked. The difference between the two provides the onscreen time for that viewing of the form. Since respondents are given the ability to back up to previous answers to review or change their responses, total onscreen time is computed for each form as an aggregate of the onscreen timers for each viewing of a particular form.

Transit time is computed by looking at the end timer of the previous form and the start timer of the next. As is the case with the total onscreen times, total transit time for each form is an aggregate value. Total interview length is computed for analysis by adding together the individual total onscreen times with an additional form-level timing that is the computed transit time between forms (see Figure 1).

**Figure 1 Visual representation of onscreen and transit timers: NPSAS:04**




SOURCE: U.S. Department of Education, National Center for Education Statistics, 2004 National Postsecondary Student Aid Study (NPSAS:04).

[1] "Form" is synonymous with "screen"; thus it may correspond with a single questionnaire item or a series of questionnaire items that are displayed on a single screen.

Determining transit time and storing it separately is especially important in Internet-based surveying, since it allows instrument designers to separate the two key pieces that make up the total interview length: 1) respondent burden in understanding and answering a particular question and 2) the time it takes for that information to return to the Web server, be stored, and have the next question sent to the respondent.

As seen in Table 1, transit time generally makes up a small portion of the total interview time. However, although this paper is primarily focused on form-level timing data and its utility in instrument development and refinement, the utility of transit-time information should not be overlooked. RTI has used transit-time information to analyze interview completion by time of day to determine how high-Internet-traffic times of day (e.g., lunch hours or close of business) affect total interview times. Transit time is also often used to pinpoint areas of interest when refining instrument programming.

**Table 1. Average onscreen and transit times in minutes, by response mode: B&B:93/03**

| Instrument section | Average total time | Average total onscreen time | Average total transit time |
|---|---|---|---|
| All web and CATI[1] respondents | 34.6 | 27.4 | 7.2 |
| Web respondents | 34.4 | 24.4 | 10.0 |
| CATI respondents | 34.8 | 29.1 | 5.7 |

[1]CATI = Computer-assisted telephone interview. Computer-assisted personal interview (CAPI) cases were excluded from this analysis.
NOTE: Times are presented separately for time onscreen and time in transit. Interview times are presented only for completed interviews (partial interviews were excluded). This table represents data collected in RTI's first web-based education survey (2003) The transit times for this particular study were affected by a number of factors that have since changed with the technology; for example more respondents were using slower dial-up connections than in later studies, and Internet bandwidth into the survey hosting website was increased after this survey was completed (in part due to timing results from this data collection).
SOURCE: U.S. Department of Education, National Center for Education Statistics, 1993/03 Baccalaureate and Beyond Longitudinal Study (B&B:93/03).

## Utility of Timing Data

The obvious use of timing data is to calculate the average length of an interview to show that the burden on the respondent does not exceed the limit specified by the client and the Office of Management and Budget (OMB). As outlined above, this can be accomplished by simply inserting time stamps at the beginning and end of the instrument and subtracting the start time from the end time to get the length of the interview. Likewise, one can collect timing data for each section of the instrument to determine how much of the interview is spent, for example, obtaining sociodemographic information. But to truly realize the value of timers, time stamps must be inserted on every form (screen) in the instrument.

Our large-scale data collections typically involve a field test prior to the full-scale study. Invariably, our field test instrument is somewhat longer, on average, than the client requirement for respondent burden in the full-scale instrument. Knowing the average length of the interview from the field test and the desired respondent burden, one knows exactly how many minutes and seconds must be eliminated from the instrument to achieve the desired interview length. Armed with that information, together with the individual screen timing data and the percentage of cases each screen was administered to, the instrument designers can begin identifying potential items for elimination. The first step is to multiply the average number of seconds spent on each screen by the percentage of cases that reached that screen (preferably entering the information into a spreadsheet). This yields the number of seconds the interview will be reduced by eliminating a given screen.[2] The next step is to rank screens (in the spreadsheet) for elimination by how critical the data collected are, sometimes taking the administration time into account (i.e., a particular item may be useful, but if it takes a disproportionately long time to administer one must decide whether it is worth losing several other items to keep it). Once items for potential elimination are ranked, then the administration times can be added up, starting with the item that is ranked lowest in terms of importance of data and working down the list until the sum of the administration times is approximately equal to the number of minutes and seconds by which the interview must be reduced. If new items are to be inserted into the full-scale instrument, then additional items will need to be eliminated based on the estimated average administration time of the new items.

[2] Note that this can be done programmatically and directly entered into a spreadsheet package.

Table 2 presents an example of this process based on the NSOPF:04 field test, where the average time spent in the instrument was 42 minutes—12 minutes more than the client mandated 30-minute full-scale interview. Of these 42 minutes, on average 35 minutes were spent answering questions (onscreen time) and 7 minutes were spent saving data and loading forms (transit time). Twenty-seven screens, which were calculated to take 7 minutes to administer, were recommended for deletion from the instrument for full-scale administration. In addition, server connection capacity and coding changes were made to reduce interview time by an estimated 5 minutes. The average administration time for the full-scale interview was 29.7 minutes, just under the 30 minute maximum burden required by the client (Heuer et al., 2004).

Taking this a step further, one could apply this process to ensure that the instrument does not exceed client limits for certain subgroups (e.g., undergraduates, graduate students, drop-outs).

Another use for screen timing data that is similar to implement is the design of an abbreviated questionnaire. Abbreviated questionnaires may be used to convert nonrespondents (near the end of data collection) who are unwilling to spend 30 minutes of their time answering questions but can be persuaded to answer questions for 10 minutes. This can be done using individual screen timers from the instrument being administered, as discussed above. Because of the complex skip logic in NSOPF:04, the instrument designers—rather than picking individual items—decided it was critical to keep sections intact when creating an abbreviated instrument and instead used section timers (which are essentially a sum of the individual screen and transit timers for a section). The design team evaluated each section to determine the value of the data collected, choosing three sections that together took less than 10 minutes to administer, based on timing data collected up to that point in the data collection.

**Table 2 Field test questionnaire items recommended for deletion and amount of time (in seconds) to administer: NSOPF:04**

| Screen | Label | Time to administer (seconds) |
|---|---|---|
| Total | | 449.0 |
| Q7 | Part-time faculty: years employed part time | 9.8 |
| Q17B | Holds Ph.D. in addition to professional degree | 0.3 |
| Q17C | Year received doctoral degree | 0.0 |
| Q17C2VS | Doctoral field—verbatim | 0.0 |
| Q17C2CD | Online coding: doctoral field | 0.0 |
| Q17C3 | Online coding: doctoral degree institution (name, city, state) | 0.2 |
| Q17D2 | Online coding: bachelor's degree institution (name, city, state) | 38.5 |
| Q19C | Number classes taught full time/part time at other postsecondary institution | 3.9 |
| Q20 | Non-postsecondary education jobs related to teaching field | 6.1 |
| Q22 | Total number of postsecondary educators employed as faculty | 13.3 |
| Q25 | First postsecondary faculty position—academic rank | 6.9 |
| Q29 | Previous job related to teaching field | 10.7 |
| Q30 | Years teaching in postsecondary institutions | 9.1 |
| Q34A–Q34D | Percentage allotment of other time | 76.4 |
| Q40A–Q40G | Uses of Website | 24.6 |
| Q43A–Q43D | Plan/develop instruction/curriculum/employment opportunities | 51.9 |
| Q44A–Q44F | Training opportunities | 56.0 |
| Q45 | Hours professional training in 2003 | 27.1 |
| Q52AiCAT | Categorical items for nonresponse follow-up to Q52AA–Q52AG | 6.3 |
| Q58 | Primary funding source | 6.9 |
| Q59 | Number of grants/contracts | 7.8 |
| Q60A | Total funding grants/contracts | 3.5 |
| Q60B | Range total funding grants/contracts | 0.7 |
| Q63 | Age expecting to stop working at postsecondary institution | 20.3 |
| Q76A–Q76E | Type of disability | 0.9 |
| Q78 | Number of dependents | 14.1 |
| Q84 | Respondent comments and suggestions | 53.7 |

NOTE: Recommendations for items to be deleted were based on examination of timing reports, use of the item in previous reports, monitoring of interviews, reliability testing, and rankings of the item by project staff and Technical Review Panel members. Estimates of time to administer are based on the total time spent on a given form (onscreen plus transit time), weighted by the percentage of cases who saw the question; hence seemingly impossible times (less than a second) are screens that were reached by a very small subset of cases.

SOURCE: U.S. Department of Education, National Center for Education Statistics, 2004 National Study of Postsecondary Faculty (NSOPF:04) field test.

Timing data can also be used to assess and improve the quality of the data collected. Long administration times for a particular screen may be an indication that the item wording is confusing or the respondent is asking questions of the interviewer. To get a better understanding of problems in the instrument, the instrumentation team for NSOPF:04 created a spreadsheet that included a row for each screen in the field-test instrument. The first column listed the form name and a brief description of the content. Subsequent columns listed the average administration time for the screen (in seconds), the rate of missing data for the item, and the number of help text hits.[3] This information was used first to guide a debriefing with telephone interviewers to shed light on why particular items took longer than the instrument developers expected them to take, or had high rates of help text hits or high rates of item-level missing data. CATI debriefing information was then added to the spreadsheet. Using this information, items were rewritten or definitions were provided onscreen to clarify the item for the full-scale instrument administration. Table 3 provides an example of a portion of the spreadsheet from the NSOPF:04 field test.

Yet another use of timing data is to test different versions of a question, particularly complex ones that take an unusually long time to administer. For example, the BPS:04/06 field test experimented with the efficacy of two coding systems to categorize field of study. Cases were randomly assigned to one of the two coding systems, and the results were evaluated to determine if there was a difference in the amount of time required to complete the coding process. Respondents first entered a text string to describe their major field of study. Then they were presented with either a pair of dropdown boxes (with general and specific categories for coding their major) or a set of categories returned by a keyword search of the database (computer-assisted coding). The manual coding using dropdown boxes took 0.9 minutes on average, compared with 0.4 minutes for the assisted coding. A separate analysis (comparing the results from an expert coder's coding of the text string) showed that the computer-assisted coding produced more reliable results, hence the computer-assisted coding method was used in the BPS:04/06 full-scale study (Wine et al., 2006).

[3] Help text was available for each screen in the instrument by clicking on a "help" button. The help text provided definitions, clarification, or examples, as appropriate for the item. Each help text request was recorded to provide data on the number of help text hits for each screen.

**Table 3 Selected field test questionnaire forms with average administration time, help text usage, and interviewer debrief comments: NSOPF:04**

| Form | Description | Mean time (secs) | Help text usage | Type of change | Comments |
|---|---|---|---|---|---|
| Q31 | Hours worked per week: paid and unpaid tasks at and outside of institution | 80 | 104 | Wording change | Suggested wording (clarify this pertains to work activities): On average, how many hours per week did you spend at each of the following work activities during the 2003 Fall Term?<br>Suggestions: Modify part d to ask "Unpaid professional services related to your work" or "Unpaid work activities outside [institution name]" followed by "Do not include volunteer work not related to your profession." Include specific examples.<br><br>High rate of missing data for parts b, c, d (9%, 12%, 15% respectively) - likely an implicit "0" but could be confusion over what constitutes paid/unpaid work at the institution.<br><br>High rate of help text hits (11%). TIO feedback: checking for the definition of "pro bono;" some sample members (part-time instructors) consider activities such as class preparation, grading papers, etc. as unpaid activities although the question considers these as paid activities; respondents are unsure as how to answer or categorize these activities. Respondent feedback: need more clarification on what is paid/unpaid. Interviewers indicated question is too wordy. |
| Q32 | Percent time: instructional activities, research, other | 63 | 28 | Format and wording changes | Suggestions: Combine with Q33 (on a single screen) to ask a) undergraduate instructional activities, b) graduate instructional activities, c) research activities, d) other activities. For "other activities" include all examples onscreen from the NSOPF:99 questionnaire and everything from Q34. Include more info in help text on what is included in the "other" category, i.e., remind respondents to include such things as outside consulting jobs. Make first sentence wording ("and at other jobs") conditional upon Q31c>0 or Q31d>0.<br><br>Interviewer debriefing: Interviewers reported that they often had to back up to Q31 because sample members had not included some of their work-related activities. Question is too wordy. |
| Q33 | Percent instructional time: undergraduate and graduate/first professional | 28 | 23 | Format change | Suggestions: Combine with Q32 on a single screen.<br><br>Interviewer debriefing: Interviewers reported sample member confusion in making the numbers sum to 100 percent. Ask Q33 immediately after Q32A (on the same screen). Question is too wordy. |
| Q34 | Percent other time: admin, professional growth, service, other activities | 82 | 33 | Delete | Suggestion: Delete for full-scale. (Of Q32-Q34 this is hardest for respondents to answer.)<br><br>Interviewer debriefing: Interviewers reported sample member confusion in making the numbers sum to 100 percent. Ask Q34 immediately after Q32C (on the same screen). Question is too wordy. |

SOURCE: U.S. Department of Education, National Center for Education Statistics, 2004 National Study of Postsecondary Faculty (NSOPF:04) field test.

## Conclusions and Future Research

We have shown that with careful implementation of instrument timers, designers and developers alike will have access to information that can assist them in refining and tailoring future instruments. Timing data helps provide the instrument designer with a holistic view of the respondent interview experience, including not just the respondent who completes the survey in one sitting, but also the user who requires several sessions and perhaps even completes different portions of the interview in different modes. Furthermore Internet latency and other server-side issues can be addressed separately if appropriate timing information is captured for those data.

One area of future research might be determining the utility of capturing additional form level timing data. These data might assist researchers in evaluating and refining form level content, such as allowing multiple items per form. However with additional data being collected there will be additional survey administration overhead for each form. Designers need to weigh the additional overhead against the usefulness of the additional data being collected and the desirability of administering multiple items on a form.

Another area for future study is in computing behaviors of sample populations that are not directly related to the survey data collection itself, but nonetheless have an impact on interview length. For an increasingly technological society, other computing behaviors such as reading and answering electronic mail, instant messaging, and Web surfing will all have an impact on the total time it takes a respondent to complete a Web survey.

Emerging technologies allow for the refreshing of parts of screens as opposed to having to send entire forms for processing and display. This should drastically reduce transit times, and thus decrease overall interview burden. Timer collection techniques and analysis methods will have to be adapted to handle these new methodologies in order to allow instrument designers to make predictive decisions about interview length and item or form level questionnaire administration.

## References


Bassili, J.N. (1996). The How and Why of Response Latency Measurement in Telephone Surveys. In N. Schwarz & S. Sudman (Eds.) *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp. 319-346). San Francisco: Jossey-Bass Publishers.

Bassili, J.N., and Scott, B.S. (1996). "Response Latency as a Signal to Question Problems in Survey Research." *The Public Opinion Quarterly*, *60*(3): 390-399.

Cominole, M., Siegel, P., Dudley, K., Roe, D., and Gilligan, T. (2006). *2004 National Postsecondary Student Aid Study (NPSAS:2004) Full Scale Methodology Report* (NCES 2006-180). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Fazio, R.H., Chen, J., McDonel, E.C., and Sherman, S.J. (1982).Attitude Accessibility, Attitude-Behavior Consistency, and the Strength of the Object-Evaluation Association. *Journal of Experimental Social Psychology*, *18*: 339-357.

Heuer, R.E., Cahalan, M., Fahimi, M., Curry-Tucker, J.L., Carley-Baxter, L., Curtin, T.R., Hinsdale, M., Jewell, D.M., Kuhr, B.D., and McLean, L. (2004). *2004 National Study of Postsecondary Faculty (NSOPF:04) Field Test Methodology Report* (NCES 2004-163). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Mulligan, K., Grant, J.T., Mockabee, S.T., and Monson, J.Q. (2003). Response Latency Methodology for Survey Research: Measurement and Modeling Strategies. *Political Analysis*, *11*(3): 289-307.

Wine, J., Cominole, M., Wheeless, S., Bryant, A., Gilligan, T., Dudley, K., and Franklin, J. (2006). *2004/06 Beginning Postsecondary Students Longitudinal Study (BPS:04/06) Field Test Methodology Report* (NCES 2006-01). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Wine, J., Cominole, M., Wheeless, S., Dudley, K., and Franklin, J. (2005). *1993/03 Baccalaureate and Beyond Longitudinal Study (B&B:93/03) Methodology Report* (NCES 2006-166). U.S. Department of Education. Washington, DC: National Center for Education Statistics.