Using Paradata to Understand Business Survey Reporting Patterns

Eric B. Fink

Joanna Fane Lineback

November 5, 2013

2013 Federal Committee on Statistical Methodology Research Conference

United States Census Bureau
Economic Statistics

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

Outline

- Motivation for Research
- Adaptive Design and Paradata
- Definitions
- Background on the Annual Survey of Manufactures (ASM)
- Results
- Future Research

Motivation for Research

- With increasing nonresponse and declining budgets for survey programs, it has become imperative to find cost efficiencies wherever possible.

- One possible intervention to mitigate these trends is the implementation of an effective adaptive survey design program.

Adaptive Design and Paradata

- Adaptive design attempts to utilize survey information to enable managers to make decisions during data collection, with a goal to balance survey costs and errors.

- One such source of survey information comes in the form of the data collected about the survey process referred to as paradata (Couper, 1998).

Definitions

- Establishment - a single physical location where business is conducted or where services or industrial operations are performed.

- Enterprise - a business organization consisting of one or more domestic establishments that were specified under common ownership or control. The enterprise and the establishment are the same for single-unit firms. Each multi-unit company forms one enterprise.

ASM – Background and Sample Design

- The ASM is a mandatory response survey providing statistics on employment, payroll, operating expenses, etc.

- There are two sampling strata: mail and nonmail.
  - The ASM mail sample includes approximately 50,000 establishments of which about 20,000 are selected with certainty, and about 30,000 are selected with probability proportional to a composite measure of establishment size.

  - Although the nonmail stratum contained approximately 180,000 individual establishments in 2011, it accounted for less than 7 percent of the estimate for total value of shipments at the total manufacturing level.

- A new sample is selected at five-year intervals beginning the second survey year subsequent to the Economic Census.

ASM – Data Collection

- Paper forms for ASM and the Company Organization Survey (COS) are mailed together for those units in sample for each survey. The ASM and COS are not conducted in Economic Census years.

- Respondents can choose to report by mail or electronically via either the Economic Census Surveyor software for multi-unit organizations or Web-based Centurion software reporting for single-unit organizations.

- In 2011, every enterprise in the sample received a paper form.
  - For the 2012 Economic Census if all 2011 ASM and COS responses for an enterprise were electronic, paper forms will not be sent.

- Responses are due within 30 days of receiving the form.

ASM – Nonresponse Follow-up

- Follow-up commences approximately two months after the initial mailout, and is usually in the form of a mailed letter.
- After the first reminder, there are three additional reminders sent, approximately once a month, until a case is considered a nonrespondent.
- For some very large establishments that are deemed important for estimation purposes, follow-up may occur via telephone.
- Currently, data collection persists for the ASM until the project runs out of time or money.

Analysis Questions

1. What is the time between mailout, downloading Surveyor software, and uploading data?

2. What is the cumulative response rate?

3. What is the cumulative total quantity response rate?

4. How much money are we spending on each stage of data collection relative to the achieved response rate?

5. What changes in reporting trends do we notice since the previous survey cycle?

6. What are the characteristics of early versus late responders?

7. Are there strong predictors for switching from paper to electronic reporting? From electronic to paper?

Data Sources

- 2011 ASM.
  - Surveyor
    - The Surveyor stores data at the company level, including company ID, a timestamp for downloading and uploading the software, and mode the company last reported their data to the Census Bureau.
  - Business Register (BR)
    - A centralized business database where information for enterprises, establishments, and employer identification numbers (EINs), and various administrative data are stored. Survey response data are also housed in the BR. Economic Census and ASM data are stored here, such as payroll, number of employees, etc.
- 2010 American Community Survey (ACS) 5-year estimates.
  - ACS ZIP-code level microdata, including measures such as income, length of commute to work, etc.

Limitations on the Analysis

- Because establishments report using the company ID, it is impossible on the paradata file to distinguish between multiple establishments downloading or uploading the software under a company, and a single individual establishment downloading or uploading the software multiple times. Thus, our analysis is restricted to only the initial download/upload event. The Surveyor analysis has only 2,014 cases as establishments report by company ID.

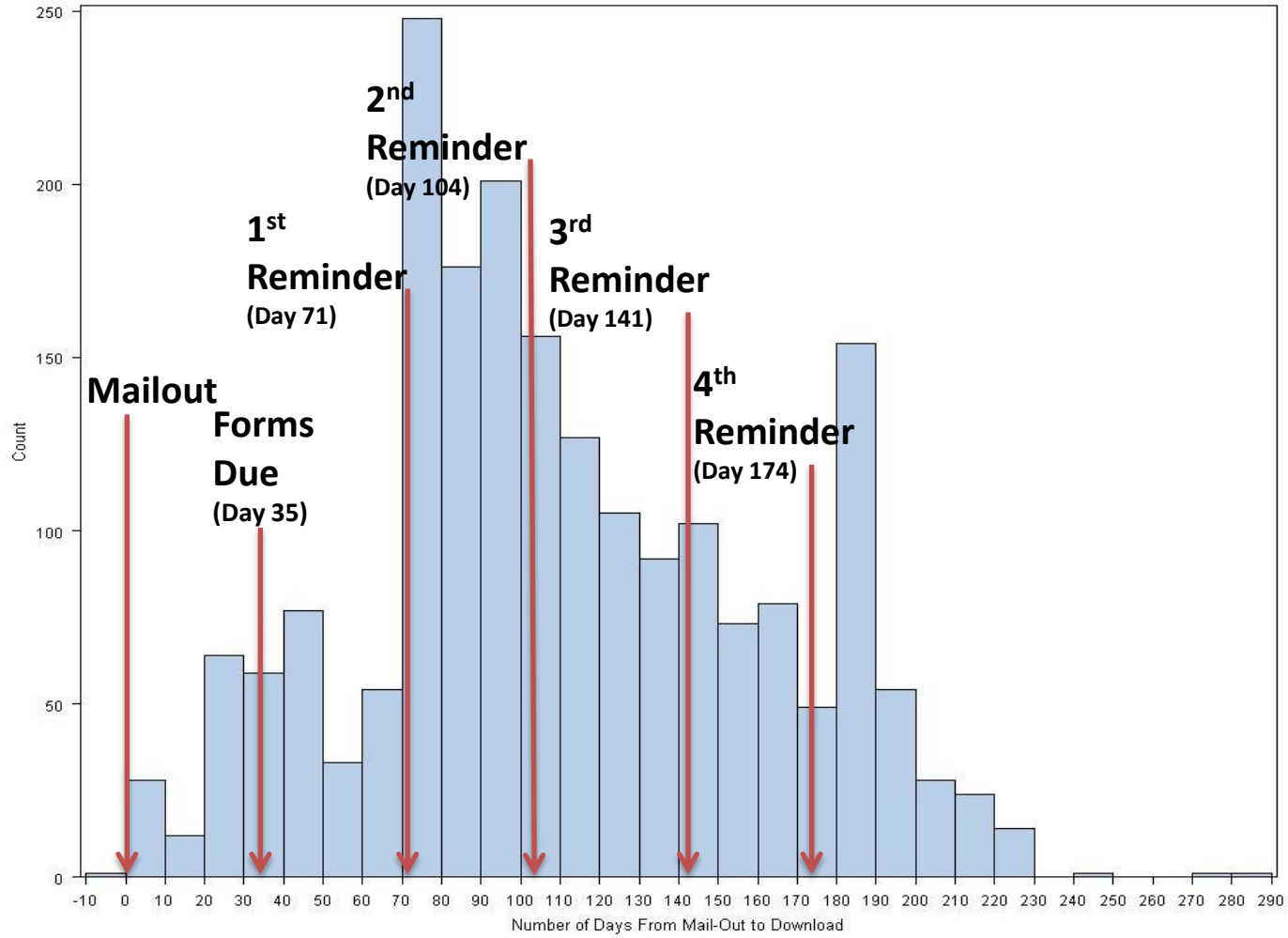
- We were forced to use variants of the check-in rate, rather than proper response rates, as we currently do not have the status flags to indicate if a respondent had supplied sufficient information to be deemed a response.

Limitations (contd.)

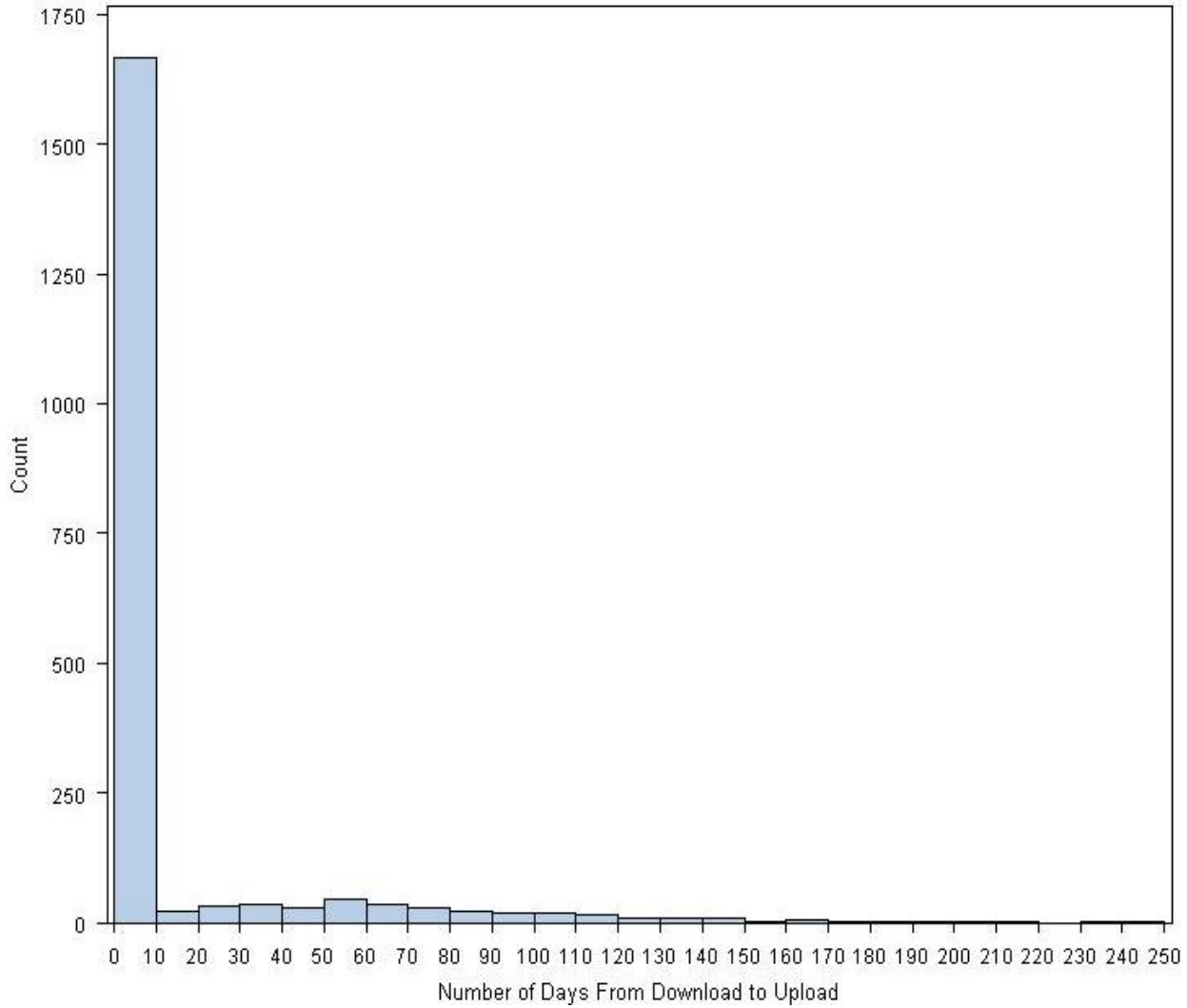
- The costs here only reflect mail form and phone call costs (direct labor, overhead, and outgoing calls). At this point, we are unable to reasonably estimate cost by survey or by survey activity such as form design, sample selection, or data processing.

- It is not always possible to separate ASM and COS costs because they are conducted jointly. As the ASM is a much more involved survey instrument in that it asks much more than does COS, a reasonable simplifying assumption for this paper is that where we are given costs for both ASM and COS, a vast majority of the resources are being utilized for ASM.

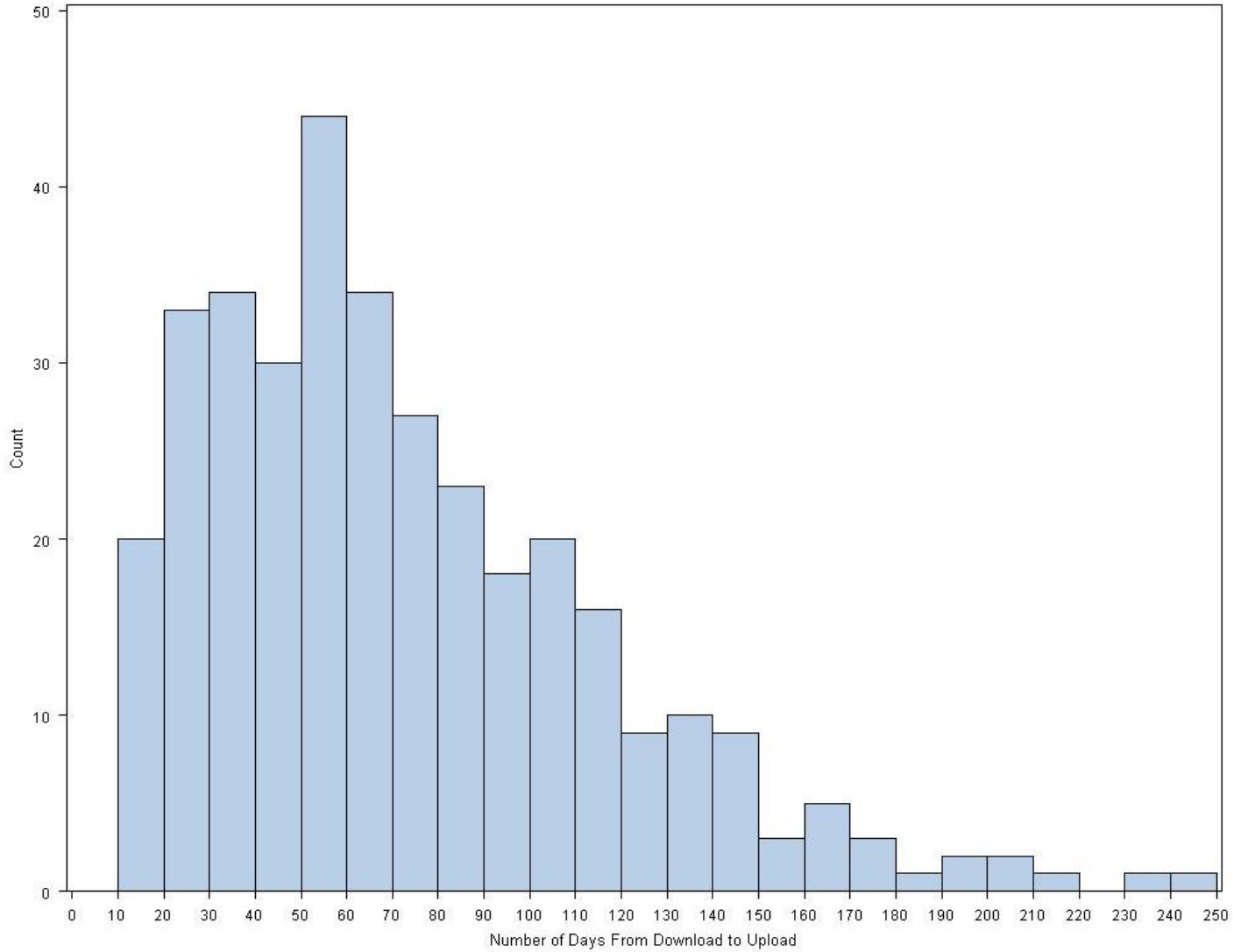
Number of days from when the form was mailed to initially downloading the Surveyor package for the 2011 ASM



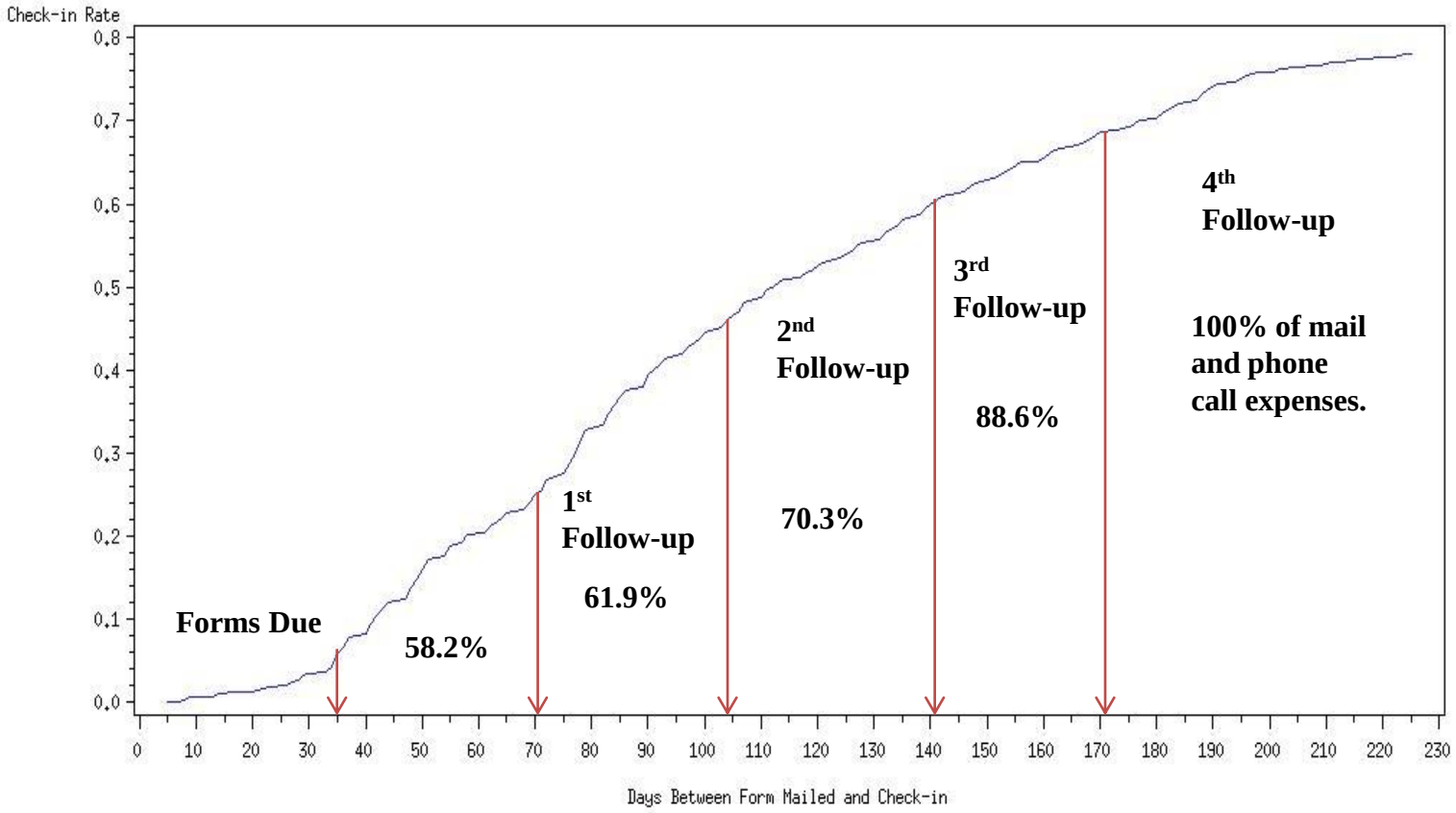
Number of days between the initial download to upload of the Surveyor software for the 2011 ASM



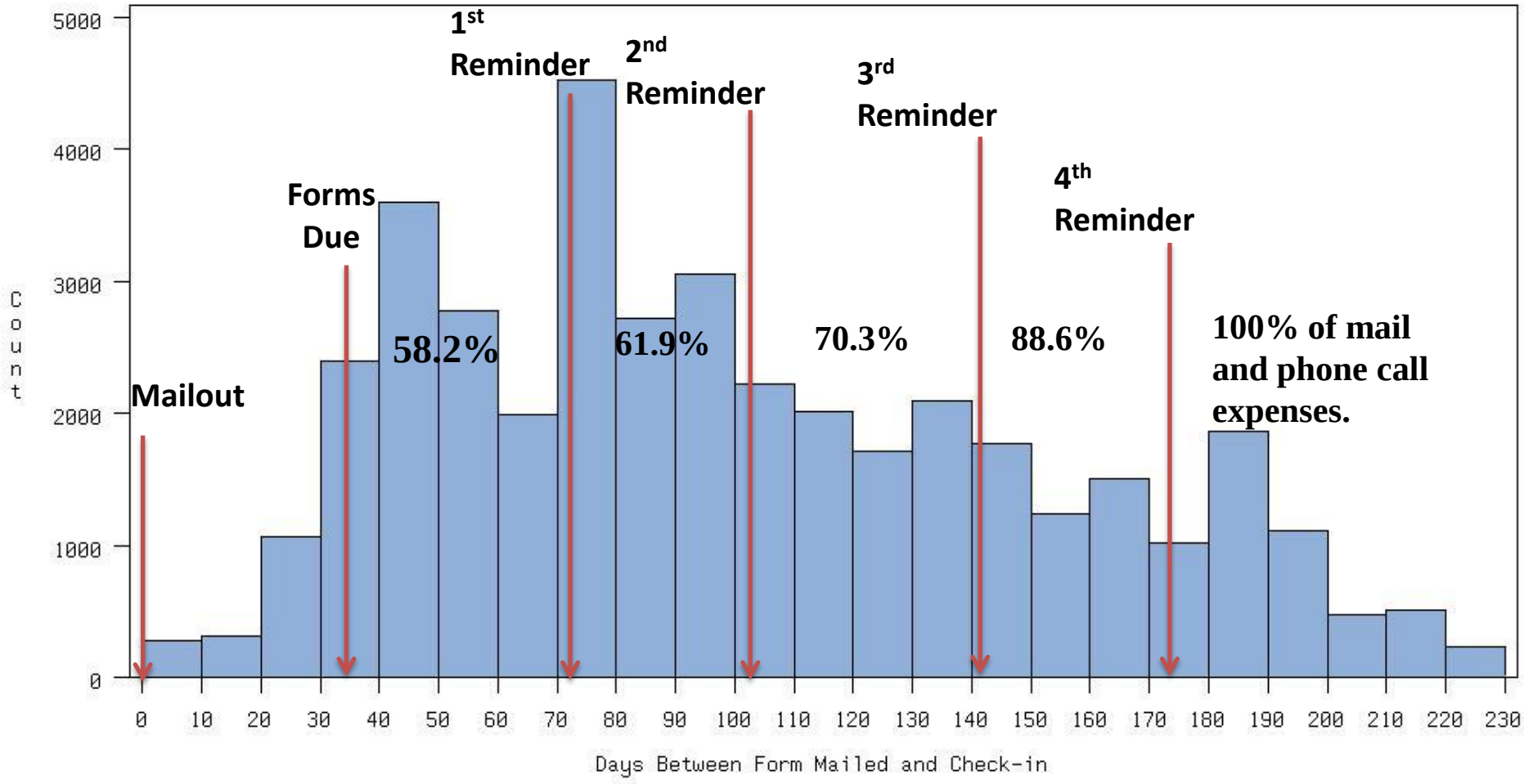
Number of days between the initial download to upload of the Surveyor software for those respondents taking longer than 10 days to upload their data for the ASM



The Check-in Rate For the 2011 ASM From When Forms Were Initially Mailed to Respondents.



The number of check-ins for the 2011 ASM



Total Quantity Response Rate

The TQRR is the proportion of the estimated, weighted total of data item *t* reported by the active tabulation units in the statistical period or from sources determined to be equivalent-quality-to-reported data (expressed as a percentage).

The TQRR is computed as follows:

$$TQRR = \left[\frac{\sum_{i=1}^{N_T}(r_{ti} + q_{ti}) * t_i}{\sum_{i=1}^{N_T} f_i t_i}\right] * 100$$

Where:

- r_{ti} is the indicator variable for reported data for tabulation unit *i* and data item *t*,
- q_{ti} is the indicator variable of "equivalent quality" data for tabulation unit *i* and data item *t*,
- t_i is the design-weighted data value of item *t* for tabulation unit *i*,
- f_i is the nonresponse weighting adjustment factor for tabulation unit *i*, and
- N_T is the total number of eligible tabulation units.

Proxy Total Quantity Response Rate (p-TQRR)

To compute this rate, we assumed a conservative administrative data rate of 8%. We then computed the denominator as:

$$\left[\sum_{i=1}^{N_T} w_i t_i\right] * 1.08$$

where:

w_i is the design weight of tabulation unit *i*,
t_i is the 2011 ASM annual payroll value for tabulation unit *i*, and
N_T is the total number of eligible tabulation units.

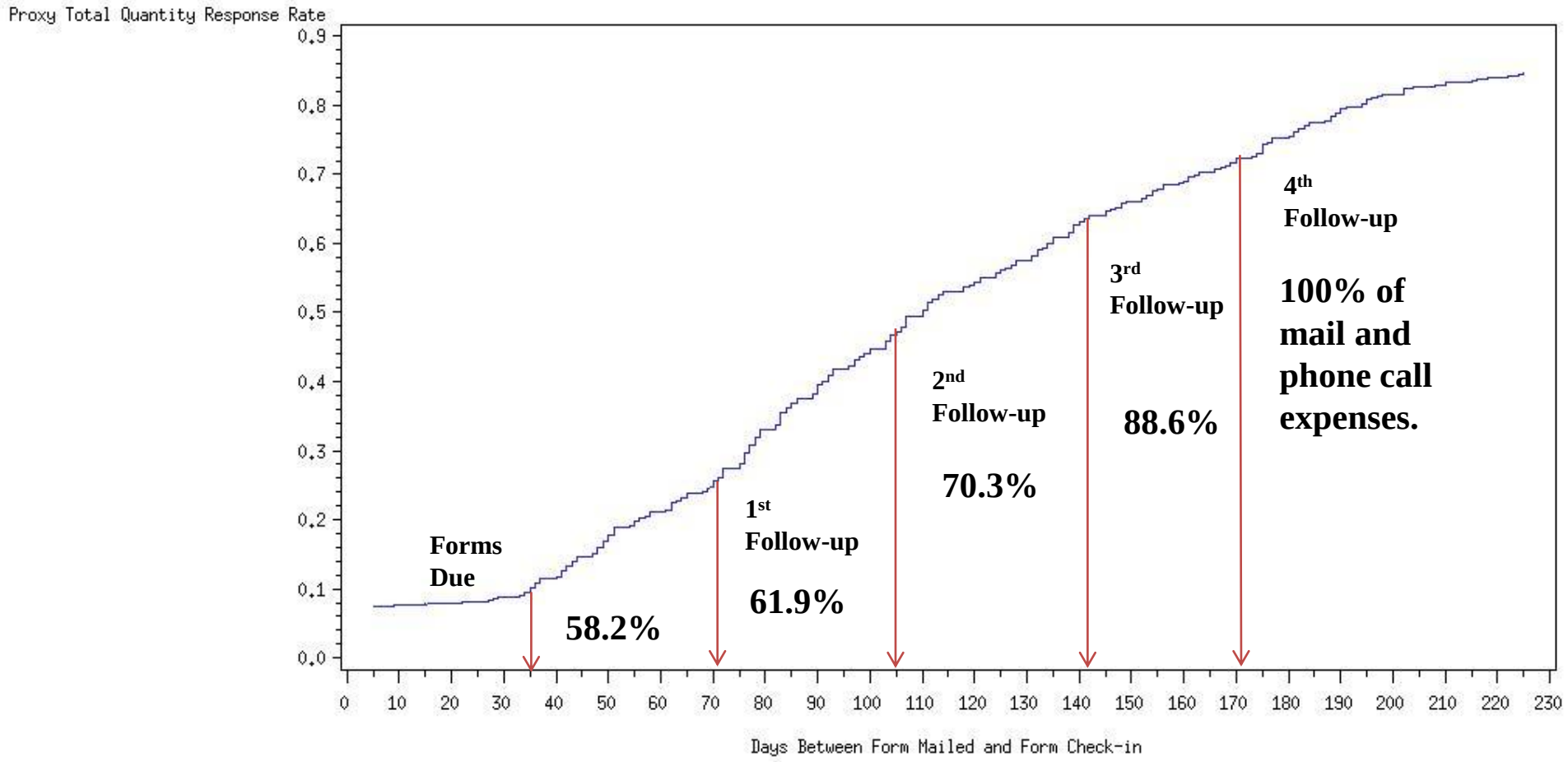
The numerator can then be computed as:

$$q + \sum_{i=1}^{N_T} w_i r_{ti} t_i$$

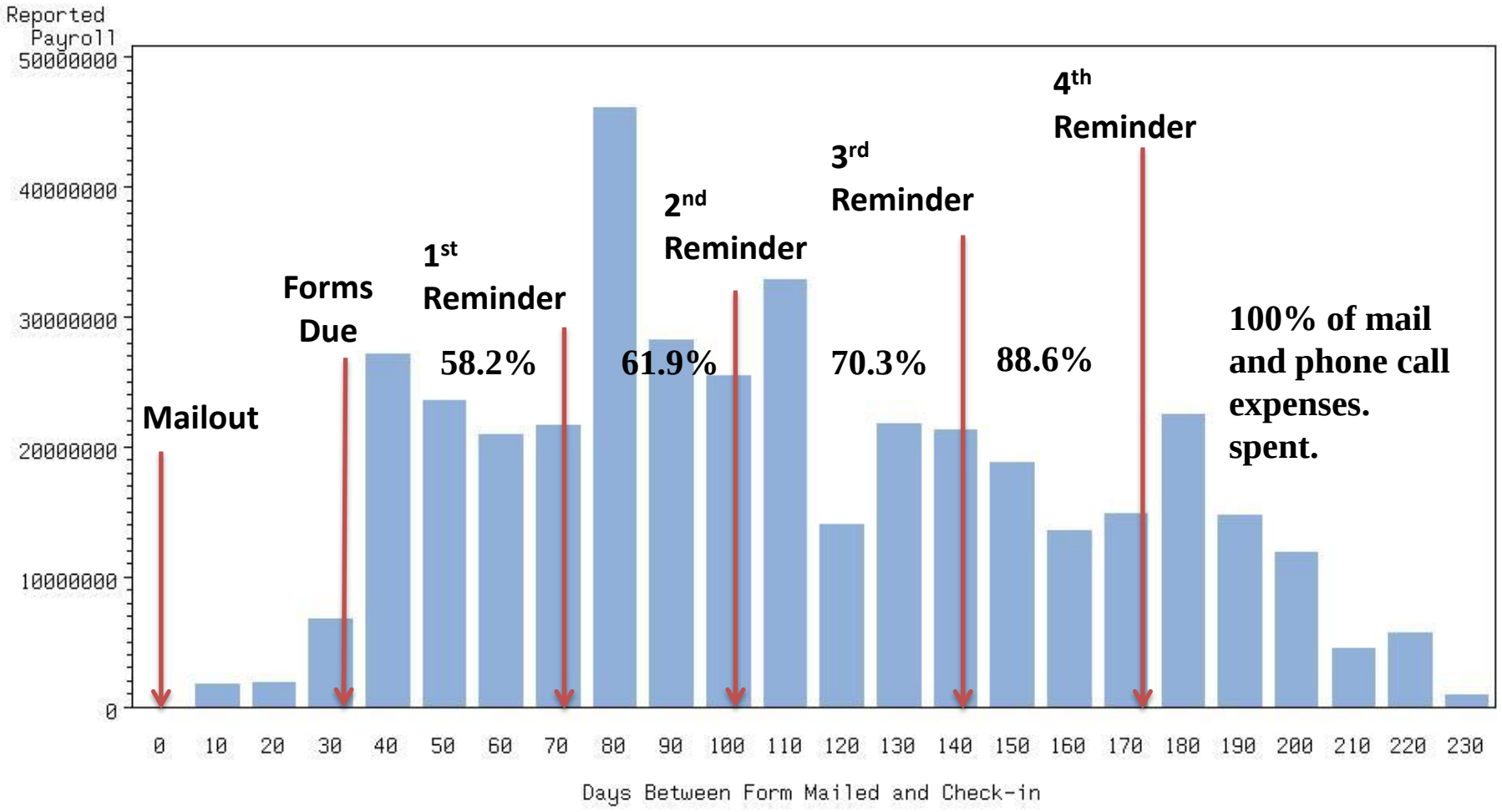
where:

q is the administrative data value calculated from the denominator.

A Proxy Total Quantity Response Rate For the 2011 ASM From When Forms Were Initially Mailed to Respondents



Reported Payroll For 2011 ASM



Mode Change Model

Parameter	df	Estimate	Standard Error	Wald Chi-Square	p-value
Days between mail out and form return	1	0.00223	0.000553	16.29	< 0.0001
Number of establishments	1	– 0.00024	0.000122	3.98	0.0461
Average percent of personal income imputed	1	– 0.0384	0.0102	14.11	0.0002
Average minutes to work	1	0.0212	0.00972	4.76	0.0291

Results of logistic regression modeling the propensity to switch from paper to electronic. There is a 52.4% concordance rate in the model.
Source: 2011 ASM, 2010 ACS 5-year estimates.

Weighted Frequencies of Establishments That Switched Response Mode From 2010 to 2011 in the ASM.

Type of Establishment	Switch from electronic to paper	Switch from paper to electronic
Multi-unit	2154.15	5083.66
Single unit	3177.16	6137.47

Source: 2010, 2011 ASM.

Confidence Intervals Comparing the Weighted Frequency of Establishments Switching Modes by 3-Digit NAICS

3-Digit NAICS	Lower Bound for Paper	Upper Bound for Paper	Lower Bound for Electronic	Upper Bound for Electronic
310	0.00	4.07	0.00	8.31
311	220.70	472.86	479.93	864.01
312	0.00	61.75	88.26	221.20
313	20.38	80.92	30.78	110.88
314	15.51	239.17	80.98	397.56
315	10.88	135.10	99.91	371.55
316	0.00	25.63	4.93	33.25
321	148.79	383.83	391.64	706.18
322	34.38	115.18	98.74	239.06
323	278.54	694.70	719.78	1308.00
324	22.69	152.39	38.06	104.26
325	101.01	286.51	420.26	733.64
326	107.43	237.27	543.55	895.77
327	245.87	557.87	543.40	968.82
330	0.00	6.34	N/A	N/A
331	55.91	196.61	143.01	268.43
332	837.88	1417.00	1985.00	2798.00
333	253.50	534.38	894.23	1339.00
334	79.05	312.27	305.30	564.84
335	38.71	111.53	126.67	261.35
336	91.77	247.99	219.04	400.16
337	159.01	451.27	322.13	712.57
339	339.90	854.92	502.06	1013.00

Source: 2010, 2011 ASM.

Discussion

- Most Surveyor software packages are initially uploaded the same day they are initially downloaded.
  - This originally suggested to us most establishments/companies have the information readily accessible to enter into the software.
  - However, most establishments do not respond until 70 – 190 days after receiving their paper form for ASM, indicating establishments took the time to record requested information on a paper form, downloaded Surveyor, entered the necessary information, and then uploaded the software.
- On the surface, the nonresponse follow-ups seem effective. There is a relatively monotonic increase in check-rates and well as the proxy TQRR.
- Furthermore, it does not appear that differential investment in each phase of follow-up is yielding any discernible increase in the rates.
- Based on these results, there are no obvious actionable items to recommend for the ASM.

Future Research

- The results lend themselves to many questions best answered via careful experimental design.
  - Why is there a due date of 30 days, when clearly most establishments take at least 60 for ASM?
  - What would happen if we altered follow-up procedures?

- We are continuing to incorporate the cost information into our analysis, most notably we hope to be able to show how resources are being allocated throughout the survey life cycle in hopes that we can find ways to improve efficiency. We will conduct this analysis on key subgroups such as geographic areas, industries, and multiunit and single unit establishments.

- We will also be expanding the above logistic regression model into a multinomial model to understand business characteristics of those establishments not changing response mode.

Contact Information

- eric.fink@census.gov