Articles

Response Mode and Bias Analysis in the IRS Individual Taxpayer Burden Survey

J Michael Brick, Roy Nord, Karen Masken, George Contos

Tags: survey practice

DOI: 10.29115/SP-2010-0025

Survey Practice

Vol. 3, Issue 5, 2010

The Internal Revenue Service (IRS) periodically conducts surveys to measure the time and money that individuals spend on pre-filing and filing activities in response to the requirements of the U.S. federal tax system. The survey data are used as an input to the Individual Taxpayer Burden Model (ITBM). The ITBM is a micro-simulation model that is based on econometrically estimated relationships between compliance burden and the tax characteristics available from the associated tax returns of the taxpayers.

The objectives for the model are: to assess the impact of programs on taxpayer burden, to assess the role of burden in tax administration, to fulfill IRS obligations to the Office of Management (OMB) for information required by the Paperwork Reductions Act, and to improve services to taxpayers. Official forecasts of total compliance burden are produced for each fiscal year. In addition, estimates of average compliance burden for each calendar year by tax form are published in the taxpayer instructions as a guide for the taxpayers. Finally, the model supports tax policy decision making through "what-if" type analysis, which allows the IRS to understand better the effect of changing rules or laws or processes.

SURVEY DESCRIPTION

The survey's target population was individual taxpayers who filed a Tax Year 2007 federal income tax return (Form 1040, 1040A or 1040EZ) during the 2008 processing year and were at least 18 years old at the time the survey went into the field. Taxpayers living abroad were excluded. In an effort to contend with memory decay, the sampling was conducted at several points in time over the course of the 2008 processing year and selected returns were surveyed in three waves.

For all waves of taxpayers selected for the survey, the same strata definitions and sampling rates were used. The first wave (11,786 taxpayers) included returns filed between January and April 2008, and the sample was selected at the end of May 2008. The second wave (2,428 taxpayers) included returns filed between May and October 2008, and was selected at the end of November

2008. Finally, the third wave (391 taxpayers) covered returns filed between October and December 2008, and was selected in January of 2009. The IRS contracted with IBM to administer the survey and they began contacting taxpayers in August of 2008 and concluded surveying in May of 2009.

The survey questionnaire was divided into seven sections that each represented a primary pre-filing or filing activity that taxpayers may have spent time or money on in completing their tax returns. The sections addressed the following activities: Record Keeping; Gathering Tax Materials; Tax Planning; Form Completion; and Form Submission. The survey also covered two categories of out-of-pocket costs: Paid Preparation Costs and Other Costs (such as the purchase price of tax preparation software). Each section of the survey included a series of questions intended to enhance memory recall of items the taxpayer should consider for the respective section. At the end of each section, the taxpayer was asked to provide a time or money estimate for the particular focus of that section. For example, collecting forms and publications, obtaining books or guides, and learning about the economic stimulus payment were all activities the taxpayers were prompted to consider when providing an estimate for time they spent gathering tax materials.

SAMPLE DESIGN

The sample design was based on data from the previous survey of taxpayer burden. It was a stratified design that crossed preparation method (three categories) and the complexity of the return (five categories). The three preparation methods used were: prepared by a paid professional, self prepared using tax preparation software, and self prepared by hand. The five complexity categories (based on elements of the return) were: low, low-medium, medium, high-medium, and high. To develop the complexity categories we placed the various tax items into categories based on the record keeping intensity, tax planning activities, and overall complexity of extracting that information from the individual's financial books. More specifically, the lower categories included items that are recorded and reported at an aggregate level. The medium category includes items that require additional record keeping and are reported to the IRS separately. Many of the items included in the medium category require attaching worksheets documenting how the totals were determined. Finally, the higher categories included items that may require a separate record keeping system or a process with potentially separate rules for each item. Tracking records across years is an additional component for most in the highest category. This resulted in a final sample design with fifteen strata, with different sampling rates used in each stratum.

DATA COLLECTION

A vendor was used to obtain the most current address and telephone information (if available) for the sampled taxpayers. The data collection protocol depended on whether the sampled taxpayer could be matched to a telephone number. Telephone numbers were found for approximately 76

percent of the sampled taxpayers and these were classified as 'telephone matches'; the remainder are 'nonmatches.' Both groups ("matches" and "non-matches") were sent an initial mailing providing a detailed description of the purpose of the survey along with a letter from an IRS executive emphasizing the importance of the study and ensuring that the information collected would not be used for enforcement purposes. It also included a one-dollar bill as "an attention getter" and indicated that respondents would receive $25 if they completed the survey.

In the initial mailing, the telephone matches were informed they could wait for a call from the survey administrator (who used a Computer Assisted Telephone Interviewing (CATI) system) or complete the survey on-line by going to a specified URL. The initial mailing was staggered, allowing for telephone contact to be attempted soon after the taxpayers received the mailing. The telephone protocol called for at a minimum of 25 attempts to contact a potential respondent. The attempts were systematically spread to different hours during day and evening, weekdays and weekends. In addition to the advanced mailing, if needed, the surveyor sent up to three follow-up letters and three postcard reminders. If the telephone protocol resulted in no response, these taxpayers were switched to a modified mail protocol, although the contractor continued attempting contact over the telephone.

The "non-matches" group members were sent a letter that provided the web address (URL) and were told a mail questionnaire was being sent. If needed, the contractor also sent up to five follow-up paper questionnaire mailings and three postcard reminders (a week "one" postcard, a week "eleven" postcard, and final postcard two weeks before the end of the collection period).

RESPONSE AND MODE ANALYSIS

As shown in Table 1, of the 14,605 sampled cases 6,968 responded for an overall response rate of 47.7 percent. Three-fourths of all sampled cases were telephone matches (11,129/14,605=76.2 percent). The response rate for the matched cases was 51.6 percent; the response rate for the nonmatches (3,476) was 35.2 percent. The difference in response rates are a function of many factors, such as the stratum from which the taxpayer was selected and other characteristics of the taxpayer. One potentially important factor, even controlling for these characteristics, is the ability to telephone the sample cases to obtain responses for those cases that can be matched to a telephone number. However, this factor is confounded by the fact that typically the population of persons that can be matched to a telephone number differs from the population of those that cannot after controlling for the stratum and other demographic characteristics. For example, the matching cases are often less likely to have moved in the last few years and may have a more tangible relationship with others in their area, and these people tend to respond to surveys at a higher level. This is analogous to many RDD surveys, where the response rate for telephone numbers without an address match is 10 to 15

percentage points lower than for those with matching addresses.

Table 1 Response Rates by Assigned Protocol.

	Initial sample size	Number of respondents	Response rate
Overall	14,605	6,968	47.7%
Survey protocol			
Telephone matches	11,129	5,745	51.6%
Nonmatches	3,476	1,223	35.2%

Table 2 shows the number of completed surveys by the initial match status and the mode used by the respondents to complete the survey. One interesting finding is that a surprisingly high percentage responded by the web, with 30 percent of the responses from the telephone match group completed on-line. The mail mode also contributed substantially for the telephone matched sample. The vast majority of the nonmatch sample responded by mail, although 17 percent of the completed surveys were done on-line. Overall, 28 percent of all the responses were completed on-line, which is higher than in other of the data collection efforts that have been reported in the literature.

Table 2 Number of Complete Surveys by Assigned Protocol and Response Mode.

Assigned protocol	Response mode	Complete surveys	Percent
Telephone match	Telephone	2,748	48%
	Mail	1,282	22%
	On-line	1,715	30%
	Total	5,745	100%
Nonmatch	Mail	1,019	83%
	On-line	204	17%
	Total	1,223	100%

Although the sample cases were assigned the survey protocol based on whether a telephone number could be found, it is still interesting to briefly examine the characteristics of the respondents by the mode they used to respond to the survey. Demographically, the web respondents are younger and more educated. Age is based on the number of years they filed a tax return, and 24 percent of the web respondents filed 10 years or less while 20 percent of the mail respondents and 13 percent of the telephone respondents were in this category. Of the web respondents, 55 percent reported that they had at least a college degree while 43 percent of the mail respondents and 40 percent of the telephone respondents reported that they had college degrees. As expected, nearly all the web respondents have access to the web at home or work (97 percent), while 84 percent of mail respondents and 78 percent of telephone reported access. This response profile with younger, more connected, and more educated respondents choosing the web at a higher rate is not unusual.

We also examined the burden outcomes (time and cost) and the auxiliary variables available from the original tax forms by the chosen response mode. In terms of time burden (record keeping, tax planning and total time), the telephone respondents reported spending more time than the mail and web respondents. For total time, the telephone respondents reported 31 hours compared to the 27 and 26 hours reported by the mail and web respondents. For burden cost (paid professional, other, and total), the mail respondents, on average, reported greater costs than the web or the telephone respondents (e.g., for professional costs the mail respondent average was $462, the web average was $367, and the telephone respondent average was $348). We suspect, but do not have concrete evidence since this was not an experiment, that these differences in burden outcomes are related to the demographic differences and population differences rather than being directly related to the mode choice of the respondents.

Finally, we wanted to compare survey costs for each of the response modes. Unfortunately, the only information provided to the IRS by the survey contractor was the number of days the survey was in the field (showing the importance of advance planning of the nonresponse bias study). Using this as a proxy for cost, we found the median number of days for those who responded via the web (11 days) to be almost half the median of either the telephone or mail respondents (21 days each).

The auxiliary variables we examined were known for both respondents and nonrespondents from the taxpayer's original filing. Table 3 gives the response mode distribution for a few of these auxiliary variables. Consistent with the previous analysis, the web respondents were younger, more likely to file electronically and use self-software, and were less likely to live in rural areas. These, and other auxiliary variables, are discussed in the next section.

Table 3 Auxiliary Variable Distribution by Response Mode.

Auxiliary variable	Response mode		
	Telephone	Mail	Web
Average age	55.5	51	45.5
w/ dependents	45.4	44	43.3
No dependents	60.3	54.6	46.9
Average income			
Married	116,807	134,896	116,552
Not married	36,657	40,928	38,664
Rural	46%	41%	38%
Schedule C present	23%	21%	23%
Schedule D present	41%	33%	37%
Filed electronically	62%	54%	66%
Preparation method			
Paid preparer	75%	67%	58%
Self-paper	9%	14%	9%
Self-software	16%	19%	34%

NONRESPONSE BIAS ANALYSIS

Our nonresponse bias analysis was based on preliminary results for the first sampling wave. As shown in Table 4, response rates varied widely across the strata, indicating the potential for nonresponse bias. Taxpayers who utilized a paid preparer had lower response rates than taxpayers who prepared their own returns and the more complex the return was, the higher the response rate tended to be. This is in keeping with the literature that suggests that people with a vested interest in the subject will respond to surveys at a higher rate (Groves and Couper 1998).

Table 4 Preliminary Response Rates for First Sampling Wave.

Strata definition	Initial sample size	Number of respondents	Response rate
Paid, Low	535	157	29%
Paid, Low-Medium	2,081	665	32%
Paid, Medium	1,657	637	38%
Paid, Medium-High	1,820	713	39%
Paid, High	2,107	807	38%
Self, Low	368	146	40%
Self, Low-Medium	418	167	40%
Self, Medium	83	45	54%
Self, Medium-High	73	32	44%
Self, High	22	13	59%
Soft, Low	575	196	34%
Soft, Low-Medium	777	298	38%
Soft, Medium	591	271	46%
Soft, Medium-High	537	252	47%
Soft, High	142	75	53%
Total	11,786	4,474	38%

While there is a great deal of literature on nonresponse bias analysis and adjustments, much of it assumes that there is only a single outcome variable of interest (e.g., Curtin et al. 2000; Ekholm and Laaksonen 1991). The issue faced in this particular survey was that there are seven separate outcome measures of comparable interest. We explore the consequences of nonresponse adjustments for a vector of outcome variables of interest, not just one. We hypothesized that different outcome variables would likely require different nonresponse adjustments and that adjustments based on one outcome variable may adversely affect estimates of other outcome variables. We also hypothesized that raking the weights would address this issue and provide better results overall (specifically the bias for all the statistics could be controlled). To test these hypotheses, we compared a number of different weighting schemes utilizing post-stratification and raking to determine the statistical properties of the estimates.

AUXILIARY DATA

To aid in the nonresponse bias analysis, tax return information and some demographic data from other external sources were available for all sampled taxpayers. In conducting the nonresponse bias analysis, we used the following variables from the tax return: adjusted gross income, preparation method, complexity, presence of schedules C and D, balance due, and whether the return was electronically filed. We also made use of several demographic variables: gender, filing status (as a proxy for marital status), age of primary taxpayer, age of youngest child, region and urbanicity (based on Zip code).

METHODOLOGY

The first step was to develop separate regression models for response and for

each outcome variable of interest. Treating each of these as dependent variables, we used the same twelve auxiliary variables as independent variables and determined which were significant in each of the respective models. As shown in Table 5, all of the auxiliary variables proved to be significant in at least one of the models.

We then developed separate poststratification weights for each outcome variable as if that variable was the only outcome variable of interest. The respective adjustment cells were determined by the most significant auxiliary variables (based on Type III sums of squares) in the respective regression model. Poststratification forced the estimate based on the respondents to exactly equal the known population counts of filers for the cells of the specific variable used. As a control, we also developed weights based on a general model that used only adjusted gross income, since that was the one variable significant in all models. We then used eleven of the auxiliary variables and developed raked weights (we dropped urbanicity because there were too many adjustment cells for the program to run when it was included). Raking is a process that iteratively adjusts the weights to match each of the marginal population totals until the weights converge. One way of thinking of raking is as poststratification applied iteratively to more than one dimension. Battaglia et al. (2009) describe the basics of raking, software for doing raking, and some of the issues that may be encountered when it is used. In all, we developed ten different sets of weights. For this study, we used the raking algorithm in WesVar – free software available at (http://www.westat.com/westat/statistical_software/WesVar/index.cfm).

Table 5 Models for Missingness and Selected Survey Responses.

Source	Respond	Total burden**	Total time	Total cost	Record keeping time	Tax planning time	Paid prep cost	Other cost
Adjusted gross income	X	X	X	X	X	X	X	X
Due a refund			X	X	X	X	X	
Complexity	X	X	X	X	X	X	X	X
Presence of Schedule C			X		X			
Presence of Schedule D	X		X	X	X	X	X	
Preparation method	X	X	X	X			X	X
Electronically filed	X	X						X
Filing status / Gender	X		X		X			
Age of respondent	X		X	X	X	X	X	
Age of youngest child	X		X	X	X	X	X	
Region	X	X	X	X	X		X	X
Urbanicity	X			X			X	

X - Significant with Pr<=0.05

**Time monetized at $20/hr

Next, we compared point estimates, bias, and variances under each weighting scheme. For the survey outcome variables, we assume that the point estimate using the post stratification weight developed based on the model for that particular outcome is the 'best' or minimal bias estimate. Under all weighting schemes, we found that the estimates were very close to the 'best' estimate and were not statistically significantly different.

For the auxiliary variables, the true population value is known so the bias analysis was straightforward. We found that the point estimates for each of the auxiliary variables was almost always biased under each of the weighting schemes, except for raking. Under raking, only a few of the point estimates remained biased. The literature suggests that if auxiliary variables are associated with both response and the outcome variable of interest then using them in weight adjustments generally reduces bias (Little 1986). Since this is the case, we assume that the raking reduces bias in the survey variable estimates – though the true bias is unknown. The increase in the standard error for the raking estimator was generally negligible as well.

Our last step was to compare the variances of the estimates for each of the weighting schemes. We compared the variance inflation factors due to weight adjustments, which were computed as one plus the coefficient of variation of the weights squared. As one would expect, the raking scheme had the highest amount of variance inflation. The inflation factor for raking was 1.31, compared to 1.06 for the general weighting scheme. With the exception of the Other Cost scheme (1.05), the remaining inflation factors fell between the raking scheme and the general scheme.

Finally, we looked at the ratio of the raking variance to the various post stratification schemes. We found that variance for total time was actually lower in the raking scheme than in the scheme for total time. For the other scheme comparisons, the variance under the raking scheme was higher, but not strikingly, ranging from a factor of 1.04 to 1.20.

CONCLUSION

The individual taxpayer burden survey was a multi-mode survey undertaken by the IRS between August 2008 and May of 2009. All taxpayers were contacted by an advance mailing that invited them to complete the survey on-line. If the taxpayer's telephone number could be obtained from commercial vendors, they were called to complete the survey by telephone; the nonmatches were mailed a questionnaire to complete. The overall response rate was about 48 percent, with a much higher response rate for those taxpayers with matching telephone numbers. One of the surprising outcomes was the relatively high percentage of the respondents who chose to complete the survey on-line rather than by telephone (if a telephone number was obtained) or by mail. While this survey may not be typical of household surveys, the results do show that, at least in this case, a substantial number of respondents were interested in the

offer of the web survey.

We also conducted a nonresponse bias analysis that focused on the bias associated with weighting schemes. We compared strategies that were specifically designed to reduce the bias for several different variables, as well as one strategy that is a standard general type of approach. The strategies for the specific variables were identified by running a series of regressions with the specific variable as the dependent variable. While this approach has been advocated as a method of reducing both bias and variance (Little 1986), it is essentially a univariate approach. Our analysis showed that none of the variable specific strategies performed well for many of the other outcome variables of interest.

We also investigated a raking approach where the dimensions were created using the auxiliary variables that were identified in the regression analysis. The raking strategy worked well in controlling the bias for all the outcome variables. One of the concerns about raking with many dimensions is that it will result in high variation in the weights and increase the variance of the estimates. However, the raking for this study did not substantially increase the variation in the weights despite the large number of dimensions. The low bias and moderate variance associated with the raking strategy suggests that this method is very beneficial compared to the other strategies considered. Accordingly, raking will be used to control the nonresponse bias in this study.

While the procedures used in our study are general, there are some features that may limit its utility in other applications. First, this study benefited from a rich set of auxiliary data pertinent to the characteristics being estimated. Without these data, the bias analysis would not have been as informative. Second, the estimation procedures used to create the various weights are complex and require careful review, especially the raking methods. Researchers should be familiar with these techniques before proceeding. Battaglia et al. (2009) discuss some of the problems that might arise in raking and diagnostics that can be used to avoid errors.

REFERENCES

Battaglia, M., D. Hoaglin, and M. Frankel. 2009. "Practical Considerations in Raking Survey Data." *Survey Practice*. http://www.surveypractice.org.

Curtin, R., S. Presser, and E. Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64: 413–28.

Ekholm, A., and S. Laaksonen. 1991. "Weighting via Response Modeling in the Finnish Household Budget Survey." *Journal of Official Statistics* 7: 325–37.

Groves, R., and M. Couper. 1998. *Nonresponse in Household Interview Surveys*. John Wiley & Sons, Inc.

Gupta, A., and J. O'Hare. 2000. "Practical Microsimulation Models." In *Economic Analysis: Microsimulation Modeling in Government*. North Holland.

Guyton, J., J. O'Hare, M. Stavrianos, and E. Toder. 2003. "Estimating the Compliance Cost of the U.S. Individual Income Tax." *National Tax Journal* 56: 673–88.

Little, R. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review* 54: 139–57.