Effect of Mode Choice and Respondent Characteristics on Data Quality: Profiling Respondents to BEA's Annual Survey of Foreign Direct Investment in the United States

Ricardo Limés [1]

[1] U.S. Bureau of Economic Analysis, 4600 Silver Hill Rd, Washington, DC 20233

Abstract

The U.S. Bureau of Economic Analysis (BEA) conducts the Annual Survey of Foreign Direct Investment in the United States, a mandatory enterprise level survey that collects information on the finances and operations of foreign-owned U.S. businesses. Respondents submit data by mail, facsimile, and BEA's electronic filing system (eFile). Over the last decade BEA has made incremental efforts to increase the eFile usage rate among respondents. The literature indicates that business characteristics and the mode of response can influence the quality of the data reported. This study profiles respondents to BEA's direct investment surveys by collection mode with the goal of better understanding respondent characteristics and mode choices. The paper analyzes the effect of survey response mode and respondent characteristics on the quality of the data reported to BEA. The paper aims to motivate future research that leads to better data collection strategies to increase response rates, improve data quality, and lower survey costs.

Key Words: Enterprise survey, mode choice, respondent characteristics, data quality, BEA, electronic filing

1. Background

The Bureau of Economic Analysis (BEA) prepares official U.S. economic statistics such as the U.S. International Transactions Accounts, the National Income and Product Accounts, and the Input-Output accounts. The Direct Investment Division of BEA conducts seven surveys that are used to produce direct investment transactions and income statistics for the U.S. International Transaction Accounts, direct investment position statistics for the U.S. International Investment Position Accounts, activities of multinational enterprises statistics, and statistics on new foreign direct investment in the United States.

The Annual Survey of Foreign Direct Investment in the United States (Form BE-15) collects financial and operating data from U.S. affiliates that are used to produce BEA's statistics on the activities of multinational enterprises.[1] From 2008 to 2011, affiliates were required to file one of three forms (BE-15A, BE-15B, or BE-15C (EZ)) depending on their size and whether or not they were majority-owned. Majority-owned affiliates with total assets, sales, or net income (or loss) greater than $275 million were required to

[1] A U.S. affiliate is a U.S. business enterprise in which there is foreign direct investment—that is, in which a single foreign person owns or controls, directly or indirectly, 10 percent or more of the voting securities or an equivalent interest.

report on the more detailed A form.[2] To minimize the burden on survey respondents, the less detailed B form was required of mid-sized majority-owned affiliates (that is, those with assets, sales, or net income/loss greater than $120 million but less than or equal to $275 million) and of minority-owned affiliates that had total assets, sales, or net income (or loss) greater than $120 million.[3] Smaller affiliates that had total assets, sales, or net income (or loss) greater the $40 million but less than or equal to $120 million were required to report on the abbreviated C form.

Filers of the BE-15 survey submitted their reports through various modes. They sent paper forms via mail or fax or they completed their reports through BEA's electronic filing (eFile) system. Once they logged into the eFile system, they completed a fillable PDF version of the form (with content identical to the paper forms) and submitted electronically. The fillable PDF had several "soft checks" in which a pop-up notification warned respondents their answers did not meet certain conditions.

This study explores BE-15 survey filers. First, descriptive statistics on response modes will be provided. Then we will see if the response mode chosen by filers had any effect on the quality of the information reported, as measured by the timeliness and error rate of the reports received. Lastly, we will see if other respondent characteristics aside from mode of response had an effect on the quality of the information reported.

The 2008-2011 period was chosen because it is the latest full intra-benchmark period available. Every five years, the benchmark survey of foreign direct investment in the United States (Form BE-12), BEA's most comprehensive survey of foreign direct investment in the United States in terms of both the number of companies covered and the amount of information gathered, is conducted in lieu of the annual BE-15 survey. During a benchmark year, all U.S. business enterprises in which a foreign person owns a 10 percent or more voting interest are required to file, whether or not they are contacted by BEA. In these benchmark surveys BEA takes additional steps, including increased outreach to potential respondents, use of survey sample frame information obtained from other government agencies, and mining of commercial datasets, to ensure complete coverage of the universe of U.S. affiliates. All of these factors usually result in a larger change in the survey frame in benchmark years than would occur in a non-benchmark year. Benchmark years have been omitted in this study to aid in isolating the effects of mode choice and respondent characteristics.

2. Summary Statistics

Electronic filing (eFile) usage increased every year during the 2008-2011 period. The BE-15 eFile usage rate went from 41.3 percent in 2008 to 60.1 percent in 2011. By survey form type, A form filers had a higher eFile usage rate than B form filers, except in 2010, and B form filers had a higher usage than C form filers. A Chi-square test confirmed ($\alpha < 0.01$) that there was an association between survey form type and the mode of response used. However, this association was due to C form filers' relatively low eFile usage rate; there was no association between the mode of response and survey form

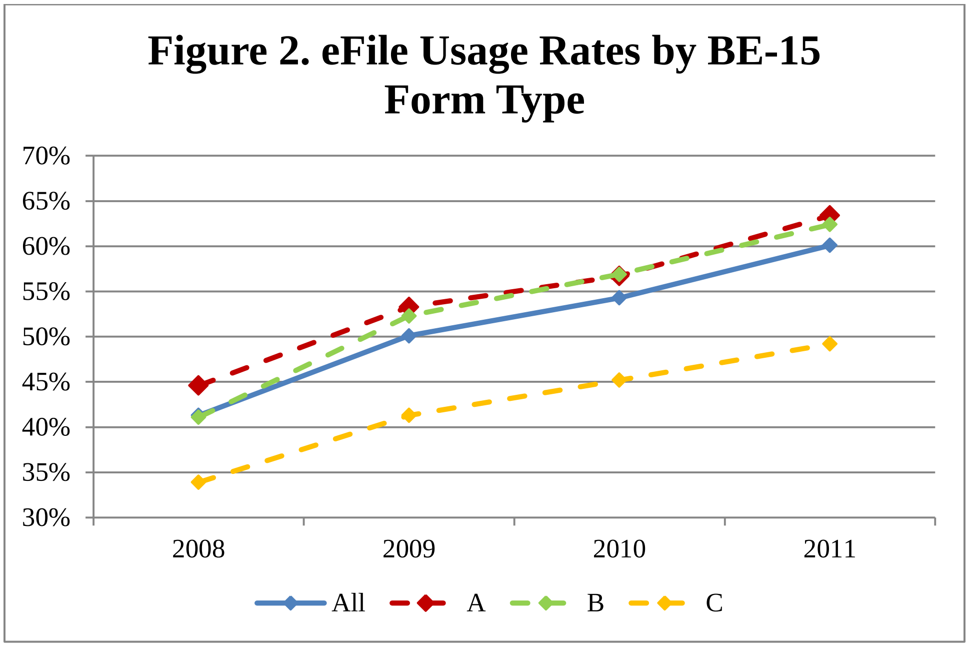
[2] A U.S. affiliate is majority owned if the combined direct or indirect voting ownership interests (or the equivalent) of all the foreign parents of the U.S. affiliate exceed 50 percent.

[3] A U.S. affiliate is minority owned if the combined direct or indirect voting ownership interests (or the equivalent) of all the foreign parents of the U.S. affiliate are at least 10 percent, but not more than 50 percent.

type when only A and B form filers were considered. Table 2 and figure 2 below show eFile usage rates for the 2008-2011 period broken out by survey form type.

Table 2. eFile Usage Rates

	BE-15 Form Type			
Year	All	A	B	C
2008	41.3%	44.6%	41.1%	33.9%
2009	50.1%	53.3%	52.3%	41.3%
2010	54.3%	56.7%	56.9%	45.2%
2011	60.1%	63.4%	62.4%	49.2%



3. Measuring Data Quality

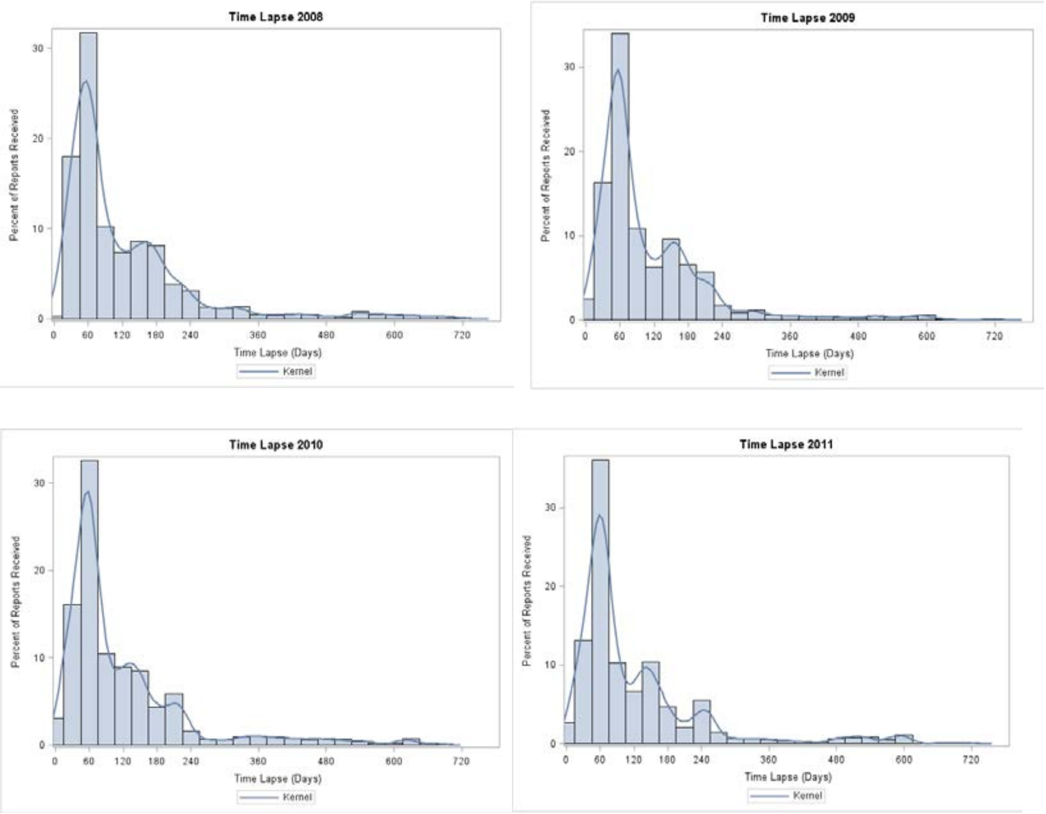
3.1 Timeliness

BEA is interested in receiving reports in the timeliest manner possible to facilitate their processing, to ensure that BEA's data releases are accurate, timely, and complete, and to minimize costs associated with follow-up for non-compliance. To study the timeliness of the reports submitted, the time it took BEA to receive a completed report, from now on referred to as "the time lapse," was calculated as the difference in days between the date the reports were mailed to respondents and the date BEA received a completed report. Time lapses greater than 730 days (2 years) were excluded because reports received after 2 years would not be incorporated into the published statistics for the year covered by the report.

Most reports had a due date of 2 months after the mailing date. However, respondents may request an informal extension of up to 30 days after the initial due date for any form they file. In addition, respondents may request formal extensions of 3 months if they file a BE-15B or BE-15C form, or of 4 months if they file the longer BE-15A form, from their original due dates.[4]

These patterns of due dates and extensions are noticeable when the time lapse variable is plotted on a histogram and a kernel distribution is overlaid on it; figure 3 below shows histograms with a kernel distribution curve for time lapse for each year in the 2008-2011 period. The distribution of the time lapse variable is positively skewed. Over one-third of responses were received between 45 and 75 days after the mailout date, representing the highest proportion of responses received of any 30-day time lapse period; and this period includes the original due date most respondents had. In fact, around one-half of all responses were received by 75 days after the mailout date. From there, the proportion of responses received dropped for the subsequent time lapse periods, but bumped up again between 135 days and 195 days, reflecting the due dates of those respondents that asked for formal extensions. Subsequently, the proportion of responses received continued to drop, with a slight bump up again between about 195 and 255 days, likely because of the compliance mailout sent to delinquent respondents. BEA received about 90 percent of all reports within 255 days, and continued to receive the remaining in the following time lapse periods.

Figure 3. Time Lapse by Year



Next, we consider the impact (if any) the mode of response had on time lapse. Descriptive statistics (mean, median, and lower and upper quartiles) for the time lapse are

[4] Formal extensions must be submitted in writing, and respondents must explain why they are requesting them.

shown below for both eFile and paper submissions. Cases where the time lapse descriptive statistic for eFile users was shorter than for paper filers are highlighted in yellow.

Table 3a. Time Lapse 2008

Form	Mean		Median		Lower quartile		Upper quartile	
Type	eFile	Paper	eFile	Paper	eFile	Paper	eFile	Paper
All forms	125	119	72	76	56	50	163	157
A	138	138	86	98	58	62	176	177
B	115	111	62	66	51	49	153	149
C	99	87	54	44	30	30	144	123

Table 3b. Time Lapse 2009

Form	Mean		Median		Lower quartile		Upper quartile	
Type	eFile	Paper	eFile	Paper	eFile	Paper	eFile	Paper
All forms	108	114	65	70	50	54	148	152
A	115	130	78	84	56	58	149	166
B	96	111	58	65	44	49	141	149
C	109	89	62	63	39	42	162	120

Table 3c. Time Lapse 2010

Form	Mean		Median		Lower quartile		Upper quartile	
Type	eFile	Paper	eFile	Paper	eFile	Paper	eFile	Paper
All forms	117	118	70	67	53	53	147	147
A	131	135	85	82	55	60	162	159
B	102	106	61	63	46	48	120	127
C	91	92	61	62	40	34	121	131

Table 3d. Time lapse 2011

Form	Mean		Median		Lower quartile		Upper quartile	
Type	eFile	Paper	eFile	Paper	eFile	Paper	eFile	Paper
All forms	122	122	73	69	52	59	152	151
A	126	134	81	80	55	62	157	160
B	117	109	65	67	47	53	146	136
C	118	109	61	69	40	48	149	140

From the preceding tables we can see that, except for the 2009 survey year, responses via eFile did not consistently come in faster than paper submittals. However, further analysis can be done to determine if there is a relationship between mode of response and time lapse.

One way to do this would be by performing an analysis of variance test, to see if the variation in time lapse can be explained by the mode of response. To this end an analysis of variance model was constructed with time lapse as the response variable and mode of response (paper or eFile) and survey form type as the explanatory variables. Survey form type was included as an explanatory variable to be able to analyze the effect of mode on time lapse while controlling for form type because from the observed data the type of form used does seem to affect the time lapse, and mode usage varies among form types.[5]

The model indicated that only between 2.7 percent (in 2008) and 0.4 percent (in 2011) of the variation in time lapse is explained by mode of response and survey form type, suggesting that there are other factors with a significant impact on time lapse not accounted for in the model. As suspected, survey form type had a significant effect ($\alpha < 0.01$) on the variation in time lapse during all years. However, the mode used to complete the report was not significant.[6] Summary results are presented in Table 4 below.

Table 4. Results of Analysis of Variance Test

	2008	2009	2010	2011
Model R-Square	0.027	0.012	0.023	0.004
Response Mode				
F (df, df error)	0.61 (1,3237)	4.43 (1, 3241)	0.65 (1, 3056)	0.01 (1, 2981)
Pr > F	0.4333	0.0354	0.4213	0.9099
Survey Form Type				
F (df, df error)	42.91 (2,3237)	18.58 (2, 3241)	35.83 (2, 3056)	5.58 (2, 2981)
Pr > F	<0.0001	<0.0001	<0.0001	0.0038

3.2 Error Rate

Another means of analyzing respondents' data quality is to measure how "clean" the reported data are. BEA's survey processing system (SPS) runs automated edit checks to flag inconsistencies in submitted forms. An edit check is a validity condition that governs the relationships between, and the values that can be taken by, one or more survey items. For the following analysis, an error rate was calculated for each completed report; the error rate was computed as the total number of edit checks failed or triggered divided by the total number of edit checks that were run on that form. This error rate was used as a measure of how clean reported data were instead of just using the total number of edit checks failed because the number of edit checks varies by survey form type and survey year.

Descriptive statistics (mean, median, and lower and upper quartiles) for the error rate are shown below. Statistics for 2008 are excluded from the analysis due to missing source data.[7]

[5] A Chi-square test confirmed that there was an association between survey form type and the mode of response used.
[6] In 2009 it was significant at the $\alpha < 0.05$ level.
[7] Many B forms had missing values for the number of edit checks triggered in 2008.

Table 5a. Error Rate 2009

Form	Mean		Median		Lower quartile		Upper quartile	
Type	eFile	Paper	eFile	Paper	eFile	Paper	eFile	Paper
All forms	0.0530	0.0385	0.0440	0.0330	0.0308	0.0224	0.0628	0.0483
A	0.0567	0.0383	0.0419	0.0330	0.0305	0.0216	0.0661	0.0483
B	0.0450	0.0338	0.0418	0.0308	0.0308	0.0220	0.0549	0.0418
C	0.0551	0.0432	0.0493	0.0359	0.0359	0.0269	0.0717	0.0493

Table 5b. Error Rate 2010

Form	Mean		Median		Lower quartile		Upper quartile	
Type	eFile	Paper	eFile	Paper	eFile	Paper	eFile	Paper
All forms	0.0502	0.0422	0.0448	0.0359	0.0317	0.0242	0.0622	0.0527
A	0.0486	0.0419	0.0419	0.0355	0.0292	0.0241	0.0596	0.0520
B	0.0483	0.0375	0.0462	0.0330	0.0352	0.0220	0.0593	0.0440
C	0.0584	0.0474	0.0538	0.0448	0.0404	0.0291	0.0762	0.0628

Table 5c. Error Rate 2011

Form	Mean		Median		Lower quartile		Upper quartile	
Type	eFile	Paper	eFile	Paper	eFile	Paper	eFile	Paper
All forms	0.0537	0.0434	0.0483	0.0374	0.0352	0.0264	0.0662	0.0534
A	0.0508	0.0444	0.0433	0.0382	0.0305	0.0254	0.0611	0.0560
B	0.0543	0.0402	0.0484	0.0352	0.0374	0.0264	0.0659	0.0505
C	0.0617	0.0448	0.0578	0.0400	0.0444	0.0267	0.0756	0.0533

The preceding tables show that the descriptive statistic for eFile reporters exceeded that of the paper filers in every case. On average, eFile submitted reports triggered 1.5, 0.8, and 1 percentage points more edit checks than paper reports in 2009, 2010, and 2011, respectively.

Does a respondent’s choice of mode explain the differences in error rates; do paper filers provide cleaner answers? Once again, an analysis of variance model was constructed, this time with the error rate as the response variable and mode of response and survey form type as the explanatory variables. Survey form type was included as an explanatory variable to analyze the effect of mode on error rate while controlling for form type because of the different complexities between forms.[8]

The results from the model indicate that mode of response and survey form type explained 6.8 percent (in 2009), 3.5 percent (in 2010), and 3.4 percent (in 2011) of the variation in error rate, indicating that mode and form type did impact the rate but that there were also other factors affecting the rate not accounted for in the model. Both explanatory variables, mode of response and survey form type, had a significant effect ($\alpha < 0.01$) on error rate for all years. Summary results are presented in table 6 below.

[8] From tables 4a – 4c, survey forms C usually have a higher mean error rate than A or B forms.

Table 6. Results of Analysis of Variance Test			
	2009	2010	2011
Model R-Square	0.068	0.035	0.034
Response Mode			
F (df, df error)	200.08 (1,3441)	77.89 (1, 3307)	103.74 (1, 3260)
Pr > F	<0.0001	<0.0001	<0.0001
Survey Form Type			
F (df, df error)	29.84 (2,3441)	25.67 (2, 3307)	10.88 (2, 3260)
Pr > F	<0.0001	<0.0001	<0.0001

4. Revisiting Data Quality (Additional Factors)

From the previous section, which analyzed the quality of the data reported on the survey, it is evident that there were other factors, aside from the survey form type and the mode of response used, that affected how quickly BEA received a response and how clean the data received were. Maybe reporters in a certain industry, or with ultimate beneficial owners (UBO) from a particular country, provided higher quality data? To answer these questions the analysis of variance models used in the previous section were expanded to include the industry of the U.S. affiliate (at the 4-digit BEA ISI code) and the country of UBO as explanatory variables.[9]

4.1 Timeliness

With the addition of industry of affiliate and country of UBO, the model's explanatory power increased to 13.3 percent from 2.7 percent in 2008, to 11.4 percent from 1.2 percent in 2009, to 16.5 percent from 2.3 percent in 2010, and to 13.4 percent from 0.4 percent in 2011. The country of UBO did not have a significant effect on the time lapse, but industry had a significant (α<=.01) effect in all years. Hence, the industry of affiliate explains the variation in time it takes BEA to receive a completed report when controlling for survey form type, mode of response, and country of UBO. Table 7 below presents a summary of the results.

[9] The international surveys industry (ISI) classification system and their code numbers were adapted by BEA from the North American Industry Classification System (NAICS). Additional information on BEA's ISI codes is available from https://www.bea.gov/surveys/iftcmat.htm.

Table 7. Results of Analysis of Variance Test

	2008	2009	2010	2011
Model R-Square	0.133	0.114	0.165	0.134
Response Mode				
F (df, df error)	0.83 (1,2977)	4.11 (1,2979)	0.32 (1,2797)	0.00 (1, 2724)
Pr > F	0.3627	0.0427	0.5708	0.9538
Survey Form Type				
F (df, df error)	54.96 (2,2977)	26.73 (2,2979)	51.15 (2,2797)	12.43 (2, 2724)
Pr > F	<0.0001	<0.0001	<0.0001	<0.0001
Industry of Affiliate				
F (df, df error)	1.39 (187,2977)	1.26 (187,2979)	1.98 (189,2797)	1.78 (187,2724)
Pr > F	0.0005	0.0106	<0.0001	<0.0001
Country of UBO				
F (df, df error)	1.08 (71, 2977)	1.24 (74, 2979)	1.09 (70, 2797)	0.81 (70,2724)
Pr > F	0.3011	0.0857	0.2867	0.8670

After determining that industry had a significant effect on the time lapse, a next step is to determine which industries provided their reports in a timelier manner, and which industries took the longest to submit their reports.[10]

Affiliates classified in ISI codes 5221, depository credit intermediation (banking), and 5224, non-depository credit intermediation, were among the top 5 industries in terms of the shortest average time lapse for both A and B forms. Affiliates classified in 5221 were the second fastest A-form filers (being surpassed by those classified in 5229, financial non-depository branches and agencies) and the fastest B-form filers.

On the other end of the spectrum, reporters in information technology related industries had the longest time lapses. Affiliates classified in 3344, semiconductors and other electronic components manufacturing, were among the top 5 industries in terms of the longest average time lapse for both A and B forms. Those classified in 3344 had the second longest time lapse among A-form filers (being surpassed by those classified in 5415, computer systems design and related services) and the longest time lapse among B-form filers.

[10] For the following analysis, only industry-form type combinations which had 20 or more reporting affiliates were considered. To increase the number of industry-form type combinations that made the threshold, observations for all years were grouped together. In the end, there were 66, 36, and 22 industries that made the threshold for A form, B form, and C form-filers, respectively.

4.2 Error Rate

With the addition of industry of affiliate and country of UBO, the model's explanatory power increased to 17.0 percent from 6.8 percent in 2009, 17.2 percent from 3.5 percent in 2010, and 19.1 percent from 3.4 percent in 2011; the mode of response used, the survey form type filed, the industry of the affiliate, and the country of UBO explain almost 20 percent of the variance in error rate. The country of UBO did not have a significant effect on the error rate in 2009 and 2010,[11] but it was significant in 2011 (α<.01). Industry of affiliate, on the other hand, had a significant effect on error rate in all three years analyzed. Summary results are presented in table 8 below.

Table 8. Results of Analysis of Variance Test

	2009	2010	2011
Model R-Square	0.170	0.172	0.191
Response Mode			
F (df, df error)	191.04 (1, 3178)	85.41 (1, 3046)	112.32 (1, 2999)
Pr > F	<0.0001	<0.0001	<0.0001
Survey Form Type			
F (df, df error)	29.37 (2, 3178)	21.77 (2, 3046)	11.66 (2, 2999)
Pr > F	<0.0001	<0.0001	<0.0001
Industry of Affiliate			
F (df, df error)	1.48 (188, 3178)	1.98 (190, 3046)	2.17 (189, 2999)
Pr > F	<0.0001	<0.0001	<0.0001
Country of UBO			
F (df, df error)	1.18 (74, 3178)	1.32 (71, 3046)	1.84 (72, 2999)
Pr > F	0.1451	0.0381	<0.0001

After determining that industry of affiliate had a significant (α<.01) effect on the error rate in all three years, the next step is to examine which industries had the lowest average error rates, and which had the largest.[12]

Affiliates classified in 5221, depository credit intermediation (banking), were among the top five industries with the lowest average error rates for both eFile and paper A form filers, having the third and second lowest, respectively. Those classified in 4244, grocery and related product wholesalers, had the lowest average error rate among eFile A form

[11] In 2010 it was significant at the α < 0.05 level.

[12] For the following analysis, only industry-mode-form type combinations which had 20 or more reporting affiliates were considered. To increase the number of industry-mode-form type combinations which made the threshold, observations for all years were grouped together. In the end, there were 37 industries for eFile-A form filers, 34 industries for paper-A form filers, 20 industries for eFile-B form filers, 13 industries for paper-B form filers, 6 industries for eFile-C form filers, and 10 industries for paper-C form filers that made the threshold.

filers; and those classified in 5224, non-depository credit intermediation, had the lowest average error rate among paper A form filers.

On the other end of the spectrum, affiliates classified in 3254, pharmaceuticals and medicines, were among the top five industries with the highest average error rate for both eFile and paper A form filers. Those classified in 2132, support activities for oil and gas operations, had the highest average error rate among eFile A form filers; and those classified in 5231, securities and commodity contracts intermediation and brokerage, had the highest average error rate among paper A form filers.

5. Conclusions

The first section of this study showed that eFile usage increased every year in the 2008-2011 period, and that eFile usage increased for all survey form types. Additionally, it showed that there is an association between mode of response used and survey form type, with C form filers consistently making less use of eFile. [13]

The second section compared eFile users' responses with those of paper filers in terms of timeliness and error rate. It showed that the response mode used did not have a significant impact on the time it took BEA to receive a report from the mailout date. In terms of error rate, paper submittals were consistently cleaner than reports submitted via eFile, and this was true for all survey form types. Importantly, the analysis of variance model confirmed that mode of response had a significant impact on the error rate of the data reported. The fact that paper submitted data was cleaner could be explained by the "pre-editing" process, in which BEA survey analysts can do a quick review of the forms that arrive in the mail before the forms are sent to be keypunched and entered into the SPS, but there could be other factors causing the paper submitted data to be cleaner.[14]

The analysis of variance models used to explain the variation in time lapse and error using the mode of response and the survey form type had low explanatory power; in order to increase it, the industry of affiliate and the country of UBO were added as explanatory variables. The explanatory power of the resulting models substantially increased, notably due to the significant effect industry had in explaining both the timelines and error rate of the reports received. Affiliates classified in banking stood out for their comparatively short time lapses and low error rates, possibly reflecting the industry's regulatory environment which makes it especially adept at fulfilling reporting requirements.

As an extension to this study, further research could analyze the link between eFile usage and size, as measured by the enterprises' employment, assets, sales, net income, or other metrics, particularly within the same survey form type. Also, the time it took respondents

[13] Previous research has shown that larger establishments (as measured by the number of employees) are more likely to make use of web-based data collection instruments. Even though BEA's direct investment surveys are enterprise level surveys, C form filers are smaller (in terms of assets, net income, and/or sales) than filers of A or B forms. See for example: Phipps and Jones (2010)

[14] Previous research has shown that in some complex enterprise surveys, especially those for which the web questionnaire is made to look as similar as possible to the paper counterpart (as is the case for the BE-15 survey) and for which the web instrument has few or no edit checks, data quality is not always improved with web-based data collection. See for example: Erikson (2009)

to complete the survey could be analyzed to determine if it has any effect on the error rate of the data reported. Are the data from early filers cleaner? What about the data of those that request extensions? Answers to these or other related questions could lead to better data collection strategies that increase response rates, improve data quality, and lower survey costs.

References

Dillman, Don A. Mail and Internet Surveys: The Tailored Design Method 2nd Edition (New York: John Wiley & Sons, 2000) 356.

Erikson, Johan. "Going Web-only in a Complex Enterprise Survey – Experiences and Lessons Learned". Statistics Canada's International Symposium Series: Proceedings (Component of Statistics Canada Catalogue no. 11-522-X) 2009. (http://www.statcan.gc.ca/pub/11-522-x/2008000/article/10987-eng.pdf)

Oliver, B. and Thompson, K. "An Analysis of the Mixed Collection Modes for Business Surveys at the U.S. Census Bureau". Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference. (https://fcsm.sites.usa.gov/files/2014/05/C4_Oliver_2013FCSM.pdf)

Phipps, P. and Jones, C. "Using Establishment Characteristics to Predict Respondent Mode Preferences in the Occupational Employment Statistics Survey". Presented at the Joint Statistical Meeting of the American Statistical Association in 2010. (https://www.bls.gov/osmr/pdf/st100350.pdf).