RESULTS FROM THE INCENTIVES FIELD TEST FOR THE CONSUMER EXPENDITURE INTERVIEW SURVEY

March 23, 2018

Ian Elkin, Brett McBride, Barry Steinberg

Consumer Expenditure Surveys Program Report Series

Executive Summary

From July 2016 through December 2016, the Consumer Expenditure Surveys (CE) program tested the effect different incentive delivery methods and incentive amounts have on survey costs, response rates, and data quality for the CE Interview Survey (CEQ). The results of this test will be used to inform both the Large Scale Feasibility test of the Gemini Redesign plan and the overall Gemini Redesign project.

Treatment and Control Groups			
	$5 Token Incentive (unconditional)	**$40 Survey Incentive (conditional)**	**$20 Records Use Incentive (conditional)**
All	**$5**	**$40**	**$20**
No Token	None	**$40**	**$20**
No Record	**$5**	**$40**	None
Control (1st Interviews)	None	None	None

The report findings, summarized below, are that respondents in the incentive test groups were more likely to respond to the survey, report similar quality of data, use records, and were, generally, more cooperative during the interview process as reported by the interviewers. However, these gains were offset by operational problems, questionable cost effectiveness of the incentives, and a higher (but not statistically significant) nonresponse bias in the incentive groups compared to the control group.

Findings

Cost Evaluation

CE staff hypothesized that the introduction of incentives would achieve a marked reduction in contact attempts as they/we/etc. expected that the incentives, mostly sent prior to attempted interview contact, would make the respondent more amenable to consent to the interview request, requiring less interviewer persuasion; however, minimal reduction was shown in the data.

Thus, given the lack of reduction in contact attempts for each test group, the analyses did not show a reduction in costs per interview to offset the increase in costs associated with implementing and administrating incentives to respondents.

Feasibility

Potential operational problems emerged through a quantitative analysis of test data. Some respondents reported not receiving the incentives. Many who did receive the debit cards provided by the financial institution did not activate the cards. Help desk staff debriefings revealed respondent difficulty in using the debit cards themselves.

Effect on Response

Analyses did show that estimation response rates were higher for all three incentives test groups, regardless of whether it was a first or second interview, when compared to the control group. However, differences in response rates between the test groups and the control group were modest.

Nevertheless, the increase in response rates seen in the first and second interviews disappeared by the third interview, as response rates for all three test groups fell to levels similar to the control group. Thus, it appears that the lack of an incentive offering in the second interview did not have a profound effect on respondent participation in that interview, but the lack of an effect diminished as the CUs participate in subsequent interviews.

Effect on Sample Composition

Chi-Square tests showed that there was a similar sample composition across test and control groups for every socio-demographic group with the exception of household tenure. For tenure, the All Incentive group had a lower percentage of owners than both the Control group and the No Records group. Nonresponse bias analysis was also performed on each of the three incentives groups as well as the Control group. There was some indication of nonresponse bias, but not enough to be considered statistically significant using 95% confidence intervals[1].

[1] Using Procsurveymeans in SAS, 95% confidence intervals were calculated for several major expenditure categories and is described later in the report as is the formulation for calculating bias.

Effect on Data Quality

Examining the quality of data in terms of the number of expenditures needing editing, it appeared the incentives helped improve data quality as a slight decrease in the proportion of expenditures needing editing was observed. For example, fewer than 8% of expenditures reported by CUs receiving incentives needed to be imputed compared to over 9% of expenditures reported by CUs not receiving incentives. On the other hand, the provision of incentives did not result in a statistically significant increase in the total amount of expenditures reported by any cohort of the test groups, with the exception of CUs with an income between $100,000 and $150,000 in the No Token group. The general finding of similar expenditure totals was not surprising in light of the different cohorts of the test and control groups having similar income levels.

Contents

I. Introduction

In 2009, the Bureau of Labor Statistics (BLS) Consumer Expenditure Survey (CE) initiated the multi-year Gemini Redesign Project for the purpose of researching, developing, and implementing an improved survey design to improve data quality through a verifiable reduction in measurement error – particularly error caused by underreporting. As part of the Gemini Redesign Project, the primary objective of the Incentives Field Test was to implement and test incentives as a means to address underreporting as proposed in the Gemini literature. A secondary objective of the Incentives Field Test was to test incentives as a way to increase respondent motivation and respondent cooperation. The survey research literature on incentives and the results of various CE tests[2] have shown that providing incentives to respondents is an effective method of increasing respondent cooperation and response rates (Gfroerer et al., 2002).

Incentives can be cash or non-cash, and are distributed as prepaid, promised, or performance-based. For the majority of federal household survey incentive tests reviewed in To (2014), incentives are distributed to respondents after successful survey completion; however, some studies did distribute prepaid incentives prior to the administration of the survey (either the full incentive or a token incentive) to all potential respondents. Past research has also shown that monetary incentives, particularly prepaid incentives, perform better than non-cash incentives under most conditions (Caporaso et al., 2016), but their effectiveness varies depending on the amount of the incentive, how it is administered, and respondent characteristics (Singer et al., 1999). To's (2014) research also showed that on average, monetary incentives being tested or in use in Federal surveys are approximately $50 per household, with some in excess of $100 per household.

From July 2016 through December 2016, the CE tested the effects that different incentive delivery methods and incentive amounts have on survey costs, response rates, and data quality for the CE Interview Survey (CEQ The results of this test will be used to inform both the Large Scale Feasibility test of the Gemini Redesign plan and the overall Gemini Redesign project.

[2] Gemini Incentive Structure Review: Summary of Incentive Experiences (To, 2014), CE Interview Incentives Test (Goldenberg et al., 2009), and CE Diary Incentives Test (McGrath et al. 2007) reports

II. Research Objectives

The main objective of the Incentives Field Test was to test varying incentive levels and distribution methods in the CE. Steps to conduct the test included: developing a plan for operationalizing and implementing incentives for the CE Interview Survey while keeping changes within the scope of the proposed Gemini Redesign structure; researching and recommending incentive amounts; proposing incentive distribution methods (including methods to capture respondents that generally do not respond to classic incentives); and making a recommendation regarding incentive implementation based on test results.

The Incentives Field Test attempted to answer the following research questions:

1. *Were surveying costs reduced by using incentives through the initial collection period, and for subsequent interviews?*
2. *What are the operational issues related to implementing incentives for CE Interview Survey data collection?*
3. *How do incentives affect the following:*
    a. *Respondent participation through the initial collection period, and for subsequent interviews?*
    b. *Composition of the sample through the initial collection period and for subsequent interviews?*
    c. *High-level expenditure reporting rates and data quality?*
    d. *Perceived respondent burden?*

III. Study Design

The Incentives Field Test, fielded from July 2016 through December 2016 as part of regular CE Interview Survey data collection, consisted of three test groups (All Incentives, No Token, and No Records), which were offered an incentive option, and a control group that did not receive an incentive option.

Test Groups:

1. All Incentives –

- Monetary survey incentive of $40 debit card conditional on completion of the first interview.
- Monetary unconditional token incentive of $5.
- Monetary records use incentive of $20 debit card, distributed by mail after completion of the interview, conditional on the use of at least one receipt, paper record, or electronic record.
- Debit card for survey incentive and token incentive distributed with Advance Letter mentioning incentive.

2. No Token –
    - Monetary survey incentive of $40 debit card conditional on completion of the first interview.
    - Monetary records use incentive of $20 debit card, distributed by mail after completion of the interview, conditional on the use of at least one receipt, paper record, or electronic record.
    - Debit card for survey incentive distributed with Advance Letter mentioning incentive.

3. No Record –
    - Monetary survey incentive of $40 debit card conditional on completion of the first interview.
    - Monetary unconditional token incentive of $5.
    - Debit card and token incentives distributed with Advance Letter mentioning incentive

Consumer Units (CUs) were randomly assigned to one of three test groups (starting sample size 1,350 each) or to the control group (starting sample size 1,950) and the test was conducted throughout all Census Regional Offices to create a nationally representative subsample. The main survey incentive, received by all respondents, was a debit card sent with the Advance Letter (via USPS First-Class mail[3]) and activated upon completion of the interview by providing the respondent with the debit card PIN.

[3] Goldenberg et al., (2009) – Goldenberg et al. reported that research completed after the start of the original CE Incentives field test showed little or no effect of Priority Mail on response rates for a mail survey with personal visit nonresponse follow-up and an incentive (Beckler and Ott, 2006), and only a small effect when Priority Mail was used to send letters and incentives for refusal conversions in a telephone survey (Brick et al., 2005). So, USPS First-Class mail was utilized to test mailing distribution mechanisms that were identical to what is currently used in normal CE data collection and to reduce costs.

Additional CUs residing in the household were sent an activated debit card upon completion of their interviews; however, they did not receive the token cash incentive, if they were otherwise eligible. All test groups were asked all of the usual CE questions during the visit as they were a part of normal CE collection.

To prepare the interviewers to administer the Incentives Field Test, the Census Bureau developed in-class training that lasted for half-a-day.

IV. Findings

The findings cover a number of different analyses related to incentives in the CE Interview Survey. Findings are divided into five sections, with each based on a research objective of the test.

Research Objective 1 – Cost Evaluation

Were costs reduced by using incentives through the interview period, and for subsequent interviews?

The analysis that follows presents cost information to the extent possible given the test was fielded as part of normal CE collection, and Census systems do not differentiate between regular cases and test cases at a cost level.

With most initiatives, cost is a key factor with a goal of either remaining cost neutral or reducing costs while improving key metrics. Since there are costs associated with purchasing, distributing, and administering the monetary incentives, a subsequent offsetting reduction in the cost to obtain a completed survey is needed to achieve this goal.

Average Number of Contact Attempts

Table 1: Average number of contact attempts, in-person contact attempts, and telephone contact attempts, interview 1

	N	Mean	Median	Mean (In-Per.)	Median (In-Per.)	Mean (Tel.)	Median (Tel.)
All Incentives[4]	766	4.7	4	4.3	3	5.9	5
No Token[5]	747	4.9	4	4.3	3	6.8	6
No Record[6]	781	4.8	4	4.2	3	6.4	5
Control	1,054	5.1	4	4.3	3	6.8	6

The number of contact attempts is often used as a proxy for analyzing the cost effectiveness of an applied treatment, such as incentives, during a field test. Without access to a clear cost structure per interview or per contact attempt, this method has been used to analyze the cost effectiveness of incentives in the CE Interview survey. Analysis shows that overall contact attempts for the three incentives test groups were marginally lower (less than 0.4 contact attempts) than for the control group, and did not greatly vary between test groups; these results were not significant at the 95 percent confidence interval. Analysis of telephone contact attempts yields similar results with a noted slight reduction in telephone contact attempts for the test groups that received the token incentives while in-person contact attempts were essentially the same.

It was hypothesized that the introduction of incentives would achieve a reduction in contact attempts as the incentives, mostly sent prior to attempted interview contact, were thought to make the respondent more amenable to consent to the interview request, requiring less interviewer persuasion; however, this reduction is only minimally shown in the data. The marginal decrease may be due to a misconception in what causes increased contact attempts. If the number of contact attempts is driven by inefficiencies in how the interviewers make their contacts, then it is possible that incentives have only a minimal effect in reducing that number, regardless of the cooperativeness of respondents. In addition, anecdotally, interviewer debriefings suggested an increase in interviewer willingness to make multiple placement

[4] Two-sided Wilcoxon-Mann-Whitney Rank Sum Test shows no significant differences between All Incentives group and the Control at the .05 level of significance.
[5] Two-sided Wilcoxon-Mann-Whitney Rank Sum Test shows no significant differences between No Token group and the Control at the .05 level of significance.
[6] Two-sided Wilcoxon-Mann-Whitney Rank Sum Test shows no significant differences between No Record group and the Control at the .05 level of significance.

attempts due to the presence of the incentive; therefore, it is possible that this change in interviewer behavior may have increased the number of contact attempts in the test groups.

Table 2: Average number of contact attempts, in-person contact attempts, and telephone contact attempts, interview 2

	N	Mean	Median	Mean (In-Per.)	Median (In-Per.)	Mean (Tel.)	Median (Tel.)
All Incentives[7]	593	5.3	4	4.7	4	6.1	4
No Token[8]	582	5.4	4	4.5	4	6.5	5
No Record	605	5.4	4	4.8	4	6.0	5
Control	837	5.5	4	4.9	4	6.1	5

For the second interview, overall mean contact attempts increased across all three test groups and the control group compared to mean contact attempts in the first interview. The same was true for in-person and telephone contact attempts with the exception of the All Incentives test group. In addition, the marginal reduction in overall contact attempts, as compared to the control, seen in the first interview continued to dissipate with the results being significant at the 95 percent confidence interval for the test group that received the full set of incentives and the test group that did not receive the token incentive. As such, the larger increase in contact attempts between interviews in the test groups, compared to the control group, may be attributable to respondents, who initially agreed to do the survey, due to the incentive, becoming more reluctant after discovering that the incentive did not carry over to the second interview.

Thus, given the lack of reduction in contact attempts, the analyses do not show a lasting carry-over effect; thus, indicating a reduction in costs per interview that offsets the increase in costs associated with implementing and administrating incentives to respondents is not present.

Research Objective 2 – Feasibility

What are the operational issues related to implementing incentives for CE Interview data collection?

An objective of the Incentives Field Test was to determine if there were any operational issues associated with purchasing, distributing, and administrating incentives in CE Interview Survey data collection.

[7] Two-sided Wilcoxon-Mann-Whitney Rank Sum Test shows significant differences between All Incentives group and the Control at the .05 level of significance.

[8] Two-sided Wilcoxon-Mann-Whitney Rank Sum Test shows significant differences between No Token group and the Control at the .05 level of significance.

Potential operational issues were uncovered through a quantitative analysis of test data regarding receipt of the incentives and of debit card paradata provided by the financial institution supplying the debit cards, as well as a qualitative analysis of information provided during interviewer and help desk staff debriefings.

Proportion Reporting No Incentive Receipt

*Table 3: Percentage of respondents reporting debit incentive receipt**

	N	Received	Not Received	Refused
All Incentives	766	70.7%	28.3%	1.0%
No Token	747	61.6%	38.4%	0.0%
No Record	781	70.2%	28.8%	1.0%
Total	2,294	67.0%	32.3%	0.7%

*Of complete first interviews in incentive groups (2016 Q3&Q4)

Analysis found that relatively high percentages of CUs in the incentive groups reported not receiving the debit cards which came with the advance letters addressed to the resident[9]. (See Table 3.) Approximately 10 percentage points fewer CUs in the No Token group reported receiving a debit card[10] compared to CUs in the All Incentives group, a statistically significant difference[11]. Most respondents indicated at the start of the interview that they had received the advance letter. However, among those in the incentive groups that received the letter, one in five said they did not receive the debit card. It is possible that in some CUs, the resident who received the advance letter was not the eventual survey respondent.

Debit Card Activation Rate

*Table 4: Debit card activation rate **

	N	Act. Rate
All Incentives	766	59.9%
No Token	747	52.4%
No Record	781	55.4%

*Of complete first interviews in incentive groups (2016 Q3&Q4)

[9] Among those in the incentive test groups, even among those reporting receipt of the advance letter, 21.5% reported that the debit card incentive was not received.

[10] 26% of CUs in this group reported receiving receipt of the advance letter but not the debit card incentive.

[11] One-tailed Z test at shows significant differences between All Incentives group and the Control at the .05 level of significance.

CUs in the test group who did not receive the token cash incentive were more likely to report not receiving the debit card (Table 3 in the previous section), and upon receipt, were also less likely to activate and use the debit card. (See Table 4). Anecdotal evidence suggests the token cash incentive may alleviate some of the "wariness" on the respondent's part to accept the legitimacy of the mailed incentive. Regardless, the overall low activation rate is a concern since there is an administrative cost associated with unactivated debit cards, which can lead to increased survey costs.

Proportion of Respondents Requesting Replacement Debit Cards

Respondents were expected to have received a debit card prior to the first interview. After completing the interview, they were asked if they had received the card and, in the event they hadn't, were offered a replacement card. Furthermore, respondents who used at least one record or receipt during the first interview, were expected to have received an additional debit card before the second interview. If they reported they hadn't received the card, they were offered a replacement.

Table 5: Proportion requesting another debit card

		Percent of those CUs that requested a replacement debit cards for…	
Incentive Group	Number of CUs completing the second interview	…the $40 conditional incentive	…the $20 records incentive
All Incentives	352	3.4%	3.4%
No Token	350	4.3%	3.1%
No Record	368	4.6%	NA
Total	1,070	4.1%	3.3%

Despite the fact that many CUs who completed the interview and may have also used records did not receive the promised debit cards, very few CUs requested replacements (less than 5 percent in all incentive conditions), regardless of the type of debit card incentive.

However, some respondents may have requested replacement cards. Thirty-six calls were made to the help desk reporting non-receipt of a debit card.

Respondent Complaints

Although experiences varied across CUs that participated in the three test groups, overall, interviewers noted that the respondents reacted positively to the presence of incentives.

Interviewers reported that respondent complaints were often similar to the regular complaints concerning the CE Interview Survey, such as excessive burden, invasiveness of some questions, and anti-government sentiment. However, interviewers did report some complaints specific to the Incentives Field Test. These complaints included lost or thrown away debit cards due to respondent mishandling, uncertainty over how or when to use the debit card, concern regarding the legitimacy of the incentive offer, specifically for the token cash incentive, and dissatisfaction with the service of the help desk line.

Many of these issues may be addressed through methods such as mailing the incentive using a Priority Mail envelope, additional instruction on when and how to redeem the incentive, and better trained help desk employees with additional tools at their disposal. However, issues such as anti-government sentiment and, for a small portion of the sample, the inability to use the debit card, as well as "normal" issues that surveys must overcome may be more difficult to solve.

Help Desk Calls

A help desk at the Census National Processing Center was established to resolve any operational issues reported by respondents using the debit card incentives and to field any questions related to the incentives themselves. Respondents were able to call the help desk during business hours and help desk staff were trained to resolve a myriad of debit card related issues.

Table 6: Help desk call reasons

Reason for Call	Call Reason Percent
Debit card will not work	27.2%
ATM related	23.5%
What is my balance?	6.2%
What is my PIN?	17.1%
Other	26.1%

Help desk staff were required to log each call that they received noting the reason for the call and its resolution. There were 404 calls to the help desk during data collection which was approximately 0.2 help desk calls per completed interview.

Overwhelmingly, the calls were in regards to issues related to the use of the debit card. While approximately a quarter of the calls were related to issues with the features (content?) of the debit card itself, almost one-half were related to using the debit card at ATM's or Point of Sale. Thus, robust respondent materials and interviewers and help desk staff, knowledgeable with the mechanics of the debit cards, are essential components for the use of debit cards as an incentive delivery mechanism.

Research Objective 3a – Effect on Response

How do incentives affect respondent participation through the interview period, and for subsequent interviews?

Response rates are often associated with survey quality. Historically, high response rates have been thought to increase the likelihood that the survey respondents represent the target population, thereby lowering potential nonresponse bias. Yet, research by Groves and Peytcheva (2008) has shown that there may be an inconsistent relationship between survey response rates and nonresponse bias. However, in the CE Interview survey, high response rates during the first interview are extremely important because they augur relatively high response rates in subsequent interviews.

Overall Response, Refusal, and Noncontact Rates

For the Incentives Field test, the overall response rates were calculated using the CE estimation response rate definition: the total number of complete interviews divided by the total number of eligible interviews (completed interviews plus Type A non-interviews using data using data processed through BLS systems). Type A non-interviews include interviews that were not completed due to factors, such as respondent refusal and inability to contact the respondent.

Table 7 shows that estimation response rates were higher for all three incentives test groups, regardless of whether it is a first or second interview, compared to the control group.

Table 7: Response, refusal, and noncontact rates, interview 1

	N	Res. Rate[12]	Diff[13] (% Points)	Ref. Rate[14]	Diff (% Points)	NC Rates[15]	Diff (% Points)
All Incentives[16]	1,110	68.4%	4.3	23.5%	-3.3	8.1%	-1.0
No Token	1,154	64.7%	0.6	26.2%	-0.6	9.1%	0.0
No Record[17]	1,130	69.8%	5.7	22.4%	-4.4	7.8%	-1.3
Control	1,643	64.1%		26.8%		9.1%	

However, differences in response rates compared to the control group were modest, especially for the test group that did not receive the token incentive. The response rates for CUs in the test groups receiving the token incentive were a minimum of 4.3 percentage points higher than that of the control group and the differences were significant at the 95 percent confidence interval. No significant difference in response rate was found between the test group that did not receive the token and the control group.

Due to their higher response rates, the refusal and noncontact rates for the three test groups were lower than those rates for the control group. While the refusal rate for the test group not receiving the token was only marginally lower than for the control group, the other two test groups which received the token incentive were several percentage points lower. Additionally, the noncontact rates for the test groups that received a token incentive were lower than the rate of the other test group. It appears the effect of incentives in reducing the refusal rate was stronger than in reducing the noncontact rate..

[12] Response Rate
[13] Difference
[14] Refusal Rate
[15] Noncontact Rate
[16] One-tailed T-test at shows significant differences between All Incentives group and the Control at the .05 level of significance.
[17] One-tailed T-test at shows significant differences between No Record group and the Control at the .05 level of significance.

Table 8: Response, refusal, and noncontact rates, interview 2

	N	Res. Rate	Diff (% Points)	Ref. Rate	Diff (% Points)	NC Rate	Diff (% Points)
All Incentives[18]	1,106	64.1%	5.4	29.9%	-3.7	6.0%	-1.7
No Token[19]	1,130	62.7%	4.0	30.4%	-3.2	6.9%	-0.8
No Record[20]	1,115	64.3%	5.6	28.7%	-4.9	7.0%	-0.7
Control	1,707	58.7%		33.6%		7.7%	

Additionally, although response rates for the test groups and the control group fell and refusal rates increased from the first to second interview, the response rates for the test groups were statistically significantly higher than the response rate for the control group. (See Table 8.)

Table 9: Response rates, interview 3

	N	Res. Rate	Diff (% Points)	Ref. Rate	Diff (% Points)	NC Rate	Diff (% Points)
All Incentives	1,037	59.2%	1.2	35.0%	-1.7	5.8%	0.5
No Token	1,102	60.4%	2.4	34.7%	-2.0	4.9%	-0.4
No Record	1,095	59.9%	1.9	32.9%	-3.8	7.2%	1.9
Control	1,551	58.0%		36.7%		5.3%	

However, the higher response rates seen in the first and second interviews for the test groups virtually disappeared by the third interview, falling to levels similar to the control group. (See Table 9.) Thus, it does not appear that a lack of an incentive offered in the second interview has a profound effect on respondent participation through the first two interviews periods, but the effect that offering incentives has on response rates does seem to diminish as the CUs participate in subsequent interviews, which brings into question the lasting effect of the treatment.

[18] One-tailed T-test at shows significant differences between All Incentives group and the Control at the .05 level of significance.
[19] One-tailed T-test at shows significant differences between No Token group and the Control at the .05 level of significance.
[20] One-tailed T-test at shows significant differences between No Record group and the Control at the .05 level of significance.

Attrition Rate

Table 10: Attrition rate, interview 1 to interview 2

	Att. Rate	Diff (% Points)
All Incentives	10.6%	-5.5
No Token	9.0%	-3.9
No Record	10.8%	-5.7
Control	5.1%	

In the second interview, response rates declined for all three test groups and the control group. However, the attrition rate, defined as the rate at which CUs that completed the first interview become non-responders in the second interview, is higher for all test groups than the attrition rate for the control group, suggesting that there may be correlation between respondents receiving incentives and their propensity to drop out between waves of the survey when the incentive is longer administered.

Mode of Interview

Table 11: Mode of interview, interview 1

	Telephone	In-Person	Mix
All Incentives	16.5%	80.0%	3.5%
No Token	17.5%	79.7%	2.8%
No Record	20.1%	76.2%	3.7%
Control	21.9%	75.5%	2.6%

Research by McGrath (2005) and Safir and Goldenberg (2008) has shown that data collected through in-person interviews are generally of a higher quality than data collected via the telephone. Thus, to the extent to which decisions of interview mode are determined by the respondent, it can be construed that incentives may make the respondent more receptive to an in-person interview suggesting an increase in data quality.

Although the test group rates for in-person interviews were not found to be statistically significantly higher than the in-person interview rate of the control group, the differences shown in Table 11 suggest that respondents in the incentives test groups were somewhat more likely to agree to an in-person interview than respondents in the control group.

Doorstep Concerns

*Table 12: Comparison of CU doorstep concern themes (not mutually exclusive) by incentive group**

DS theme	Interview 1 (N=4,646)	Interview 1 All Incentives (N=1,038)	Interview 1 No Token (N=1,064)	Interview 1 No Record (N=1,041)	Interview 1 Control (N=1,503)
Not interested/hostile	20.2%	19.1%	23.0%	17.0%	21.2%
Time	39.3%	36.4%	37.1%	39.3%	42.3%
Survey voluntary/privacy	30.1%	26.5%	31.9%	29.1%	32.0%
Gatekeeping	5.1%	4.8%	5.3%	4.6%	5.8%
Prior interview**	4.6%	5.1%	4.4%	2.8%	5.7%
Other	12.4%	12.9%	13.3%	11.1%	12.3%
No concerns***	38.0%	41.0%	37.2%	40.0%	35.1%

*among those contacted by an FR regardless of interview completion status

**includes some concerns (e.g. intends to quit survey) that are applicable to first interviews

***coded to indicate no doorstep concern indicated across all CU contacts

Interviewers record any doorstep (DS) concerns that may be expressed by individuals who they recruit to participate in the CEQ. Doorstep concerns are measured from a predefined Census list of actions or statements commonly made by respondents during survey recruitment that shed light on how or why they do not participate in the survey. These seventeen indicators of concern[21], although not mutually exclusive, can give an idea of what someone is thinking when they agree (or refuse) to take part in the CE Interview Survey. Related indicators were grouped into a smaller number of themes to facilitate the interpretation of how they varied by incentive group. As an example, if interviewers selected 'Not interested / does not want to be bothered,' 'Hang-up / slams door on FR' or 'Hostile or threatens FR' as concern indicators, these were grouped into the 'Not interested/hostile' theme based on the similarities of those indicators.

Table 12 above shows that there were small differences among all the groups in the frequency which each doorstep concern theme was reported by CUs to interviewers. In all three incentive groups, the plurality of CUs reported no doorstep concerns across all interviewer contacts for the first interview. In contrast, in the control group, the plurality of CUs expressed the 'time' DS concern theme (42 percent); CUs reported 'no concerns' second most frequently (35 percent). Research by McBride and Tan (2014) found that CUs

[21] Until 2014 there were 23 indicators but the Census Bureau removed 6 as part of a redesign of the Contact History Instrument (CHI). We were still able to use the same grouping method used by Kopp and colleagues (2013 "An Exploratory Study on the Association of Doorstep Concerns with Three Survey Quality Measures for the CE Interview Survey") as the remaining indicators contributed to each theme.

mentioning 'no concerns' had four times lower odds of first interview non-response than CUs reporting any concerns, and the lower non-response rate among incentive groups here bears this out.

Research Objective 3b – Effect on Sample Composition

How do incentives affect the composition of the sample through the interview period and for subsequent interviews?

If the sample is not representative of the general population, the potential exists for nonresponse bias in both expenditure data and other data. Analyses in this section show that overall there is little to no difference in the demographic characteristics of the respondents in the various test groups. Since the CUs were randomly selected for the test groups and the control group, it is expected that the sample composition across all groups is similar. However, since certain demographic subgroups may respond differently to an incentive, an analysis of sample composition was conducted. .

Demographics of Responding Sample

Table 13: Demographics – race of the reference person

	White	Black	Asian	Other[22]
All Incentives	81.4%	12.3%	4.2%	2.1%
No Token	79.2%	13.7%	4.4%	2.7%
No Record	80.3%	12.2%	4.4%	3.1%
Control	80.0%	11.3%	5.4%	3.3%

The percentage of black respondents in all three test groups was higher than the percentage in the control group. (See Table 13.) A higher proportion of white reference persons can be seen in the two test groups receiving the token incentive. Overall the variation among the test groups and the control group is minimal, suggesting that the presence of an incentive did not alter the racial composition of the sample.

[22] Other includes Pacific Islander, Native American, Native Hawaiian, Guamanian or Chamorro, Samoan, Multi-race, Other

Table 14: Demographics – Hispanic origin (reference person)

	Hispanic origin
All Incentives	13.1%
No Token	14.0%
No Record	11.7%
Control	13.4%

In addition, there is no meaningful pattern in the percentage of respondents of Hispanic origin in the test groups compared to the control group. (See Table 14.) The percentage of Hispanic reference persons in the test group receiving no token incentive is slightly higher than the control group, while the percentages of Hispanics in the remaining two test groups are slightly lower, suggesting that the presence of incentives did not affect the proportion of Hispanics participating in the survey.

Table 15: Demographics – age of reference person

	Under 25	25-34	35-44	45-54	55-64	65 & Older
All Incentives	5.7%	16.3%	16.1%	18.9%	20.0%	23.0%
No Token	5.9%	14.5%	15.1%	19.0%	20.9%	24.7%
No Record	4.0%	14.3%	19.1%	19.6%	17.4%	25.6%
Control	5.9%	14.1%	19.8%	18.0%	17.3%	24.9%

The incentives offered did not have a noticeable effect on the distribution of reference persons by age in the survey. (See Table 15.) While there is a higher proportion of 25 to 34 year olds in all three test groups compared to the control group, the difference is minimal. Otherwise there is little evidence to suggest that incentives influenced the participation of reference persons in different age groups.

Table 16: Demographics – education attainment of reference person

	No Degree	HS Degree	Some College	Bachelor's Degree	Post-Grad
All Incentives	11.8%	22.5%	32.8%	21.8%	11.1%
No Token	11.2%	22.6%	30.9%	23.5%	11.8%
No Record	8.7%	23.7%	33.2%	20.2%	14.2%
Control	10.8%	21.8%	30.7%	22.9%	13.6%

There are slightly higher levels of survey participation among reference persons at lower levels of educational attainment; however, these differences are minimal and do not follow a discernable pattern. As with previous demographic characteristics, the analysis does not show a correlation between education level and participation in the test groups.

Table 17: Demographics - CU size

	One	Two	Three	Four	Five or More
All Incentives	29.5%	33.8%	14.8%	13.3%	8.6%
No Token	29.9%	34.0%	15.4%	11.9%	8.8%
No Record	27.3%	34.8%	16.8%	11.5%	9.6%
Control	29.2%	34.6%	13.7%	11.5%	11.0%

The incentives offered did not have a noticeable effect on the distribution of respondents by CU size in the survey. While there is a higher proportion of three-person CUs in all three test groups compared to the control group, the difference is minimal. Otherwise there is no evidence to suggest that incentives influenced the participation of respondents by CU size in the survey.

Income

Table 18: CU before-tax income (imputed) by incentive group

	N	Mean	Median
All Incentives[23]	766	$67,995	$49,000
No Token[24]	747	$67,764	$51,788
No Record[25]	781	$75,715	$50,000
Control	1054	$73,422	$51,221
Total	3,348	$71,453	$50,300

Although the two test groups receiving the token incentive had lower median income amounts than the control group, neither amount was statistically significantly different from the control group's. Nor did the higher median amount for the No Token group differ significantly from the control group's.

[23] Wilcoxon Two-Sample test shows no significant differences between All Incentives group and the Control at the .05 level of significance.
[24] Wilcoxon Two-Sample test shows no significant differences between No Token group and the Control at the .05 level of significance.
[25] Wilcoxon Two-Sample test shows no significant differences between No Record group and the Control at the .05 level of significance.

Non-response Bias: Demographic Comparisons

The analysis in the previous section showed that the summary statistic proportions for several underlying demographic characteristics appear to be similar across all groups. It is crucial to statistically determine that the distributions of socio-demographic characteristics between the incentive groups and the control group are similar and representative, otherwise bias may be introduced. A two-step procedure was used to determine whether the respondents in these four groups had the same or different distribution of demographic characteristics. In the first step, an "omnibus" chi-square test of independence was run on each demographic variable to see whether there were any differences in distributions between the incentive groups and control group. Then if any differences were detected, a second test would be run to determine which specific demographic categories and which specific test groups were different. An example of this two-step process is shown below:

In the omnibus chi-square test, the null hypothesis was that all three incentive groups and the control group (four test groups) had the same proportional distribution of categories comprising a particular demographic characteristic, and the alternative hypothesis was that at least one of the groups was different. For example, if the demographic variable was "marital status of the reference person" and the variable had five different categories (Married, Widowed, Divorced, Separated, Never Married), then the null hypothesis and the alternative hypothesis would be as follows:

H_0:	p11=p21=p31=p41, and p12=p22=p32=p42, and p13=p23=p33=p43, and p14=p24=p34=p44, and p15=p25=p35=p45
H_a:	H_0 is not true (i.e., at least one "=" sign is really a "≠" sign)

Here p_{ij} is the proportion of all respondents in the i-th test group that are in the j-th marital category. Thus, if the test statistic was statistically significant, the null hypothesis would be rejected, and a conclusion would be drawn that the proportion of respondents in at least one category in one marital test group differed from the proportions of similar respondents in one or more of the other incentive groups or control group. No conclusion, however, could be drawn concerning the specific marital category(ies) nor the specific pair(s) of test groups that were different. Those would be identified from a second test using

contrasts (a linear combination of variables whose coefficients add up to zero, allowing comparison of different test groups) from a logistic regression model in which C(4,2)-1 or {4!/(4-2)!2!} -1=5 pairs of test groups would be compared to each other. This would apply to all five marital status categories totaling 25 (=5x5) individual chi-square statistics.

The distributions of respondents for the following demographic characteristics: age, gender, race, educational attainment, marital status, CU size, income, region, urbanicity and tenure were used in the analysis. The above mentioned omnibus Chi-Square test for independence with (r-1) * (c-1) degrees of freedom was used for each demographic category. The "r" represents the number of categories for each demographic variable and the "c" represents the number of incentive/control groups which is always four for this section.

Ho: Each of the 3 incentive groups and control have an equivalent distribution
Ha: at least one of the incentive groups or control is different

Table 19. *Chi Square test for Socio Demographic Distributions*

Demographic Group	# of Group Categories	P value
Sex	2	0.3740
Marital Status	5	0.1833
Race	6	0.6627
Region	4	0.3333
Urban	2	0.3287
Tenure	2	0.0028**
Education Level	4	0.1863
Income Designation	8	0.6056
Family Size	7	0.8053
Age Designation	6	0.1592

As shown on Table 19, all of the tests comparing the three incentive groups and control group showed that the p values were well above the critical value at $\alpha = 0.05$ and not significant for all demographic characteristics, with one exception, household tenure. This is shown by a "**" next to its p value of 0.0028 in Table 19. Table 20 below shows the All Incentives test group had a statistically significant lower percentage of owners compared to the No Record incentive group and the Control group, respectively (56.9 percent vs. 65.7 percent and 63.7 percent). Their corresponding Wald Chi Square p-values were 0.0037 for the All Incentives vs. Control contrast and 0.0006 for the All Incentives vs. No Record contrast, highlighting their statistical differences.

Table 20. *Tenure Respondent Distribution*

	Owner	Renter	Total CUs	Owner Distribution
All Incentives	436	330	766	56.9%
No Token	464	283	747	62.1%
No Record	513	268	781	65.7%
Control	671	383	1,054	63.7%

Demographic Comparisons by Response Rates

In addition to the distributions of demographic characteristics, which examined only respondents, unweighted response rates were examined across selected groups of the sample. These subgroups were defined by CU size, household tenure, region, PSU size,[26] and urbanicity. Chi-Square tests for equality of proportions[27] were used to determine how the control group and the incentive groups compare in response rates.

Ho: Proportion #1 = Proportion #2
Ha: Proportion #1 ≠ Proportion #2

A Bonferroni[28] technique for multiple comparisons is used for this analysis.

Table 22 provides a summary of p values associated with the comparison of response rates among all test groups, while the actual response rates for the test groups are summarized in Table 21. P-values less than 0.05 are shown with a "**" and those below 0.0166 using the Bonferroni technique are underlined as well. Several demographic subgroups were shown to have statistical differences using this approach, particularly those distinguished by tenure. Owners in the No Record incentive group showed a statistically significant higher response rate (71 percent) than owners in the other incentive groups and the control group (from 64 percent to 65 percent). In addition, renters in the All Incentives test group were found to have a statistically significant higher response rate compared to renters in the other incentive groups and the control group. The response rates for all three incentive groups were also significantly

[26] "S" PSUs are self-representing PSUs from a metropolitan CBSA having a population greater than 2,500,000. An "N" PSU is a non-self-representing PSU from a metropolitan or micropolitan CBSA having a population less than 2,500,000. An "R" PSU is a rural PSU not from a CBSA.

[27] Since the response variable has only two categories (response or nonresponse), and the incentive and control groups are labels rather than levels of a second variable, the chi-square test of independence is called a test for equality of proportions. Therefore, the resulting value follows a chi square distribution with one degree of freedom.

[28] Although $\alpha = 0.05$ was chosen, since multiple comparison testing is being performed, the true "critical alpha" is more realistically α =0.05/3=0.0166 making it tougher to conclude statistical significance. This is known as the Bonferroni technique

higher than that of the control group for 4-person CUs, with response rates from 75 percent to 80 percent while the Control group reported at under 62 percent. Statistical significance was also found for 3-person CUs with two of the three incentive groups showing significantly higher response rates than the Control group. For CUs from both "N" PSUs and Urban areas, all three incentive group response rates were higher than the control group rate, with statistically significant differences between the Control group and the All Incentives and the No Records test groups. "N" PSUs and Urban areas overlap extensively geographically so they are correlated; still CUs in both test groups respond at a higher rate than CUs in the control group.

Table 21. *Response Rates by Demographic Subgroup*

		All Incentives	No Token	No Rec	Control
Region	*Northeast*	67.5%	58.7%	66.0%	59.7%
	Midwest	70.1%	66.3%	71.7%	68.1%
	South	69.5%	63.7%	67.6%	62.1%
	West	66.6%	72.4%	70.5%	65.4%
Family Size	*1 Person CU*	66.9%	62.1%	62.3%	60.8%
	2 Person CU	66.2%	60.3%	67.3%	65.7%
	3 Person CU	66.5%	71.4%	74.4%	61.5%
	4 Person CU	79.1%	79.5%	75.0%	61.7%
	5 Person CU	74.6%	72.4%	82.8%	70.2%
	6 Person CU	68.4%	69.6%	76.9%	79.5%
	7+ Person CU	75.0%	80.0%	87.5%	65.5%
PSU Size	*S PSUs*	65.7%	62.5%	65.3%	61.4%
	N PSUs	71.7%	67.2%	70.9%	64.3%
	R PSUs	59.7%	67.2%	75.9%	75.6%
Urbanicity	*Urban CUs*	68.3%	65.7%	67.8%	62.3%
	Rural CUs	69.4%	63.4%	73.8%	71.1%
Tenure	*Owner CUs*	64.0%	64.6%	70.9%	65.1%
	Renter CUs	75.5%	66.4%	65.4%	61.4%

Table 22: *Test for equality of proportions between incentive group response rates and control group response rates: P values*

		All Incentives vs. Control	No Token vs. Control	No Rec vs. Control	All Incentives vs. No Token	All Incentives vs. No Rec	No Token vs. No Rec
Region	*Northeast*	0.0707	0.8244	0.1406	0.0614	0.7544	0.1185
	Midwest	0.6174	0.6310	0.3636	0.3615	0.7013	0.1953
	South	0.0182**	0.6219	0.0787	0.0834	0.5597	0.2446
	West	0.7612	0.0605	0.1643	0.1430	0.3151	0.6347
Family Size	*1 Person CU*	0.0710	0.6840	0.6531	0.1908	0.2116	0.9644
	2 Person CU	0.8497	0.0877	0.5867	0.0812	0.7451	0.0367**
	3 Person CU	0.3090	0.0421**	0.0060**	0.3302	0.1044	0.5351
	4 Person CU	0.0010**	0.0013**	0.0151**	0.9400	0.4450	0.4183
	5 Person CU	0.5589	0.7713	0.0827	0.7910	0.2803	0.1816
	6 Person CU	0.3553	0.3782	0.8054	0.9364	0.5241	0.5604
	7+ Person CU	0.5527	0.3922	0.2285	0.7805	0.4936	0.6714
PSU Size	*S PSUs*	0.1365	0.6962	0.1770	0.3126	0.9033	0.3765
	N PSUs	0.0028**	0.2372	0.0072**	0.0890	0.7516	0.1610
	R PSUs	0.0374**	0.2616	0.9662	0.3857	0.0585	0.2965
Urbanicity	*Urban CUs*	0.0031**	0.0924	0.0064**	0.2278	0.8190	0.3269
	Rural CUs	0.6837	0.0809	0.5549	0.2021	0.3447	0.0281**
Tenure	*Owner CUs*	0.6537	0.8433	0.0110**	0.8147	0.0063**	0.0114**
	Renter CUs	<0.0001**	0.0950	0.1940	0.0033**	0.0012**	0.7451

For most demographic characteristics (Table 21) the response rates for CUs in the incentive groups were higher than for CUs in the control group, but CUs in the All Incentives and No Record groups more frequently responded at a statistically significantly higher rate than CU's in the No Token group when compared to the Control group. (See Table 22.)

Research Objective 3c – Effect on Data Quality

How does the presence of incentives affect high-level expenditure reporting rates and data quality?

One of the primary measures of data quality for expenditure surveys is complete and accurate reporting of expenditures for the CU. Incentives can affect data quality positively by encouraging behaviors associated with good expenditure reporting, such as records use, which can result in more accurate expenditure estimates and potentially higher aggregate expenditures due to less underreporting. However, incentives can also affect data quality negatively through bad expenditure reporting due to satisficing. Consequently, the analysis that follows examines how the inclusion of incentives in the CE affected factors associated with data quality.

Total Expenditures

Table 23: Total expenditures (ZTOTAL) reported by each test group and survey wave

Interview 1	n	Mean	Median
All Incentives[29]	766	$13,958	$10,654
No Token	747	$14,563	$11,102
No Record	781	$14,300	$11,264
Control	1054	$14,208	$11,115
Total	3,348	$14,252	$11,023

Interview 2	n	Mean	Median
All Incentives	352	$14,235	$9,908
No Token	350	$13,590	$10,888
No Record	368	$14,037	$10,799
Control	2,094	$14,025	$10,157
Total	3,164	$14,001	$10,308

Analysis of mean and median total expenditures indicated no large difference in expenditures among the CUs in the incentive test groups. Those receiving the full set of token, survey, and records incentives appeared to report the lowest total expenditures in the first interview, though this was not significantly different from amounts reported by those in the control group.

[29] Two-sided Wilcoxon Two-Sample test shows no significant differences between All Incentives group and the Control at the .05 level of significance.

Expenditures by Summary Expenditure Categories

Table 24: Z summary averages by group, interview 1

	All Incentives (n=766)	No Token (n=747)	No Record (n=781)	Control (n=1,054)
Housing-related (ZHOUSING)	$4,507.88	$4,895.20	$4,704.17	$4,942.84
Rent/Mortgage (ZSHELTER)	$2,799.04	$3,090.11	$2,845.45	$3,093.97
Food (ZFOODTOT)	$2,089.68	$2,061.28	$2,082.12	$2,064.94
Transportation (ZTRANPRT)	$2,482.37	$2,376.75	$2,522.70	$2,378.82
Health (ZHEALTH)[30]	$1,126.76	$1,183.80	$1,054.34	$1,177.30
Entertainment (ZENTRMNT)	$608.45	$761.42	$694.10	$686.69
Personal Insurance, Pensions (ZPERLINS)	$1,742.57	$1,528.99	$1,677.52	$1,492.68
Apparel and Services (ZAPPAREL)[31]	$336.16	$332.57	$309.56	$307.20

CU's in the No Token group reported the highest mean total expenditures among the three test groups. This was primarily driven by higher housing-related expenditures, though the differences in total and housing-related expenditures between the No Token group and the other groups were not statistically significant. (See Table 24.) Differences in mean expenditures between the control and test groups were significant only for the health (higher for control) and apparel and services (lower for control) categories.

Table 25: Z summary averages by group, interview 2

	All Incentives (n=352)	No Token (n=350)	No Record (n=368)	Control (n=2,094)
Housing-related (ZHOUSING)	$4,510.05	$4,415.85	$4,341.97	$4,631.32
Rent/Mortgage (ZSHELTER)	$2,888.62	$2,682.96	$2,663.60	$2,933.16
Food (ZFOODTOT)	$2,138.53	$2,080.11	$2,075.50	$2,005.56
Transportation (ZTRANPRT)	$2,396.51	$2,387.07	$2,311.81	$2,277.04
Health (ZHEALTH)	$1,012.07	$1,048.72	$1,078.42	$1,135.07
Entertainment (ZENTRMNT)	$729.27	$696.32	$622.13	$631.47
Personal Insurance, Pensions (ZPERLINS)	$1,929.06	$1,528.36	$1,980.36	$1,615.28
Apparel and Services (ZAPPAREL)	$325.72	$318.42	$262.93	$301.48

Among CUs completing the second interview, the group receiving the full set of incentives reported category averages that were similar to or higher than the other incentive groups for seven of the eight

[30] Two-tailed Wilcoxon Two-Sample test shows significant differences between incentives groups and the Control at the .05 level of significance.

[31] Two-tailed Wilcoxon Two-Sample test shows significant differences between incentives groups and the Control at the .05 level of significance.

categories shown in Table 25 (the exception being health which was lower across the board), none of the differences being statistically significant. Aside from higher average expenditures for housing-related and health categories, the control group reported lower averages than the test group receiving all incentives in all of the other categories examined here, and the amounts were significantly lower for the food, transportation, and insurance categories.

Expenditures Needing Editing

The provision of incentives may improve the quality of data collected, because respondents feel their input is more valued. Conversely, incentives may convince those who otherwise would not have participated to complete the survey, and these less-willing respondents may not be as diligent in providing high-quality data. This question was examined by looking at editing rates shown in the tables below.

*Table 26: Percent expenditures that had to be allocated or imputed by group, interview 1**

	CUs	N (Expns)	% edited	% allocated	% imputed	% combined
All Incentives	766	60,436	16.7%	8.8%	7.2%	0.3%
No Token	747	60,232	17.3%	8.7%	7.8%	0.3%
No Record	781	63,203	16.8%	9.1%	7.0%	0.1%
Control	1,054	83,669	18.6%	8.5%	9.3%	0.2%

*From processed, post-EES data

For first interview expenditure data, editing rates were marginally higher - by 1 percent – 2 percent - for reports from CUs in the control group than for those from the incentive groups. Imputation rates ranged from 7.0 percent of expenditure reports from the test group not receiving a token incentive to 9.3 percent for reports from the control group. In contrast, the incentive groups showed no statistically significant differences in allocation and combined edit rates.

*Table 27: Percent expenditures that had to be allocated or imputed by group, interview 2**

	CUs	N (Expns)	% edited	% allocated	% imputed	% combined
All Incentives	352	27,505	16.8%	8.4%	7.6%	0.2%
No Token	350	27,837	16.7%	8.5%	7.7%	0.1%
No Record	368	29,867	16.6%	8.2%	7.4%	0.2%
Control	2,092	158,882	18.3%	8.5%	9.1%	0.2%

*From processed, post-EES data, excludes 2 control group CUs with no MTAB expenditure data

For second interview data, editing rates again were marginally higher for reports from the control group (18.3 percent) than from all three incentive groups (approximately 16.7 percent). Imputation rates were similarly 1.5 percentage points higher for the control group than for the incentive groups (9.1 percent compared to approximately 7.5 percent). Again, there were no statistically significant differences in allocation and combined edit rates.

Interview Quality by Demographics

Analyses were also performed using total expenditures to measure whether spending in specific demographic subgroups was affected by incentives. Twenty-eight subgroups were selected (four levels of education, eight income groups, six age groups, six race groups, two gender groups, and two tenure groups). Each subgroup was analyzed to see if the total expenditures reported by CUs receiving incentives in each subgroup were statistically significantly different from total expenditures by CUs in the control group. Using unweighted means, there was only one occurrence of statistical significance in a subgroup. A minimum sample size of thirty CUs in the incentive groups and control group was used as the threshold for which the results from a statistical test would be deemed credible. There were several occurrences where there was a statistically significant difference between the mean total expenditures of an incentive group and the control group, but in the majority of them there was insufficient sample. As such those cases were deemed not credible.

To be statistically significant at $\alpha = 0.05$, a 95 percent confidence interval was constructed for the mean total expenditures of the control group and each of its corresponding incentive groups. If the intervals did not overlap, the means would be considered statistically significantly different. As stated above, statistical significance was found for only one comparison. This occurred for the test group of CUs not receiving a token incentive with income between $100,000 and 150,000 (N=51). The mean total expenditures for this test group were $26,335 with a standard error of $2,610, while the control group (N=67) had mean total expenditures of $17,288 with a standard error of $946. Only one other group yielded results that approached statistical significance. This was for Asians in the test group receiving all incentives (N=34) with mean total expenditures of $11,847 with a standard error of $1,049. The corresponding control group (N=56) reported mean total expenditures of $17,031 and a standard error of $1,611. Those CUs receiving the incentive reported significantly lower expenditures but not at the 95% confidence level.

An analysis similar to the one carried out for total expenditures was done for the number of expenditure entries reported by CUs. Using the same set of subgroups as for total expenditures above, only CUs in the Renters subgroup that received all incentives showed any statistical difference from the CUs in the control group. This test group had a mean number of entries of 25.4 per CU with a standard error of 0.65 while the control group had a mean of 23.0 per CU with a standard error of 0.53. Even though the difference in number of entries is statistically significant at the $\alpha = 0.05$ level, there is no practical value to be derived from the result.

Non-response Bias: Bias in Weighted Sample Means

A logistic regression model was developed in an effort to measure nonresponse bias for several major expenditure categories. The model used CHI (contact history instrument) data which included the following types of indicator variables: variables related to respondents' behavior concerns or reluctance, variables related to difficulty in contacting for personal interviews, variables related to difficulty in contacting by telephone, and the variable with the number of contact attempts. The logistic regression model used both respondent and nonrespondent information to generate propensity scores. Since there are no known expenditures for the actual nonresponders, it was necessary to create a proxy for them and propensity scores were the method of choice. Propensity score[32] analysis attempts to estimate the effect of a treatment by accounting for variables that predict receipt of the treatment. For this test, response was considered receiving the treatment. The propensity score provides a single metric that summarizes all of the information from given explanatory variables. Responders that have modeled propensity scores closer to nonresponders (for this analysis having higher propensity scores) using the explanatory variables were referred to as "pseudo nonresponders". Ranking was then applied to the propensity scores from respondents to generate ranks for each CU ranging from 0-99. These resulting ranks determined which CU's were the "pseudo responders" and which were the "pseudo nonresponders". The actual response rate (~66 percent) from the data determined the cutoff value for the ranks separating the pseudo responders from the "pseudo nonresponders". CUs with ranks less than or equal to sixty-six were classified as pseudo responders while those with ranks greater than sixty-six were classified as "pseudo nonresponders". A means procedure was then used to calculate weighted means and standard errors for the expenditures by incentive and control group. The nonresponse bias formula used is

$$Bias\ (\bar{X}_R) = \frac{(\bar{X}_R - \bar{X}_{PR})}{\bar{X}_R} \times 100\%$$

[32] Guo, S. and M. Fraser., (2010)

Where:
$Bias$ ($\bar{X}_R$) is the nonresponse bias % of the weighted sample mean,
$\bar{X}_R$ is the weighted mean of all respondent expenditures,
$\bar{X}_{PR}$ is the weighted mean of the pseudo respondent expenditures.

Table 29 provides a summary of results for each of the test groups and control group for total expenditures and several other major expenditure categories. For total expenditures, the nonresponse bias was 1.2 percent for the control group, 5.9 percent for the test group receiving all incentives, 5.8 percent for the No Token test group and 3.9 percent for the test group not receiving a records incentive.

Table 28: Calculation of nonresponse bias for some major expenditure groups

All Respondents Control	Sample Size	Mean Quarterly Expenditures	Standard Err of Mean	Nonresponse Bias %
Total Expenditures	1,054	$14,440	$428	1.2%
Housing	1,054	$4,955	$172	0.4%
Food	1,054	$2,056	$42	-1.5%
Transportation	1,054	$2,423	$169	1.8%

Pseudo Respondent Control	Sample Size	Mean Quarterly Expenditures	Standard Err of Mean
Total Expenditures	677	$14,265	$487
Housing	677	$4,931	$218
Food	677	$2,081	$56
Transportation	677	$2,378	$209

All Respondents No Token	Sample Size	Mean Quarterly Expenditures	Standard Err of Mean	Nonresponse Bias %
Total Expenditures	766	$14,170	$460	5.9%
Housing	766	$4,565	$129	5.0%
Food	766	$2,121	$54	7.4%
Transportation	766	$2,526	$193	-0.4%

Pseudo Respondent No Token	Sample Size	Mean Quarterly Expenditures	Standard Err of Mean
Total Expenditures	522	$13,338	$464
Housing	522	$4,338	$145
Food	522	$1,965	$58
Transportation	522	$2,526	$256

All Respondents All Incentives	Sample Size	Mean Quarterly Expenditures	Standard Err of Mean	Nonresponse Bias %
Total Expenditures	747	$14,791	$574	5.8%
Housing	747	$4,937	$218	4.3%
Food	747	$2,077	$57	6.2%
Transportation	747	$2,505	$206	10.4%

Pseudo Respondent All Incentives	Sample Size	Mean Quarterly Expenditures	Standard Err of Mean
Total Expenditures	519	$13,934	$625
Housing	519	$4,727	$261
Food	519	$1,948	$64
Transportation	519	$2,245	$196

All Respondents No Record	Sample Size	Mean Quarterly Expenditures	Standard Err of Mean	Nonresponse Bias %
Total Expenditures	781	$14,447	$409	3.9%
Housing	781	$4,753	$148	3.7%
Food	781	$2,096	$50	2.3%
Transportation	781	$2,581	$205	9.0%

Pseudo Respondent No Record	Sample Size	Mean Quarterly Expenditures	Standard Err of Mean
Total Expenditures	527	$13,877	$480
Housing	527	$4,579	$180
Food	527	$2,048	$64
Transportation	527	$2,350	$212

Table 28 implies that the "pseudo responders" have slightly lower total expenditures than "pseudo non-responders" which is not unexpected since CUs with greater wealth (and corresponding higher expenditures) historically have lower response rates. This is most apparent for the All Incentives test group which has a lower homeownership rate. However, the standard errors were relatively large so the results did not show statistical significance using 95 percent confidence intervals. Similar patterns existed for many of the other major expenditure groups but were also not statistically significant due to large standard errors.

Records Use

One of the incentives was provided to prompt respondents to use records. In Tables 29 and 30, we examined whether the objective of this incentive was met, by dividing the entire sample (incentive and control groups) into those offered the record-use incentive and those that were not.

Table 29: Any record use by incentive group, interview 1

	CUs	% Using Records
Record-Use Incentive Groups[33]	1,511	83%
No Record-Use Incentive (includes Control)	1,832	60%
Total	3,343	70%

CUs completing first interviews were examined in Table 29. The table shows that 83 percent of CUs used at least one record when offered an incentive to do so, compared to 60 percent among CUs not offered any record incentive, including those in the control group. Interviewers designated whether a CU had used records in any section of the survey. This difference was statistically significant.

Table 30: Average and median number of sections with record use by incentive group, interview 1

	CUs	# Sections with Records (Mean)	# Sections with Records (Median)
Record-Use Incentive Group[34]	1,511	3.1	2.0
No Record-Use Incentive (includes Control)	1,832	2.1	1.0
Total	3,343	2.5	1.0

Table 30 above shows that there was an average of 3.1 sections where records were used by CUs in the record use incentive groups, and an average of 2.1 sections recorded record use in the comparison group.

[33] Two-sided Wilcoxon-Mann-Whitney rank-sum test. Excluding the Control group did not change the findings of a significant difference (68 percent of those in the No Record incentive group used records). - Significant at 95% confidence interval

[34] Two-sided Wilcoxon-Mann-Whitney rank-sum test shows significant differences between No Record group and the Control at the .05 level of significance.

This difference was statistically significant, and suggests that providing a record incentive may lead to records being used across more sections than had record incentives not been offered.

Table 31: Record use by section by incentive group, interview 1

	Record-Use Incentive Group		No Record-Use Incentive		
	CUs Using Records	% Using Records	CUs Using Records	% Using Records	Difference (Incent - None)
Section 1 - General Housing Characteristics	126	8%	128	7%	1%
Section 2 - Rented Living Quarters	116	8%	84	5%	3%
Section 3 - Owned Living Quarters	212	14%	199	11%	3%
Section 4 - Utilities and Fuels	897	59%	722	39%	20%
Section 5 - Construction, Repairs …	133	9%	127	7%	2%
Section 6 - Appliances …	109	7%	97	5%	2%
Section 7 - Household Item Repairs …	86	6%	57	3%	3%
Section 8 - Home Furnishings	101	7%	85	5%	2%
Section 9 - Clothing	295	20%	236	13%	7%
Section 10 - Rented and Leased Vehicles	41	3%	56	3%	0%
Section 11 - Owned Vehicles	228	15%	223	12%	3%
Section 12 - Vehicle Operating Expenses	256	17%	186	10%	7%
Section 13 - Insurance Other than Health	456	30%	363	20%	10%
Section 14 - Hospitalization and Health Insurance	367	24%	314	17%	7%
Section 15 - Medical and Health Expenditures	260	17%	197	11%	6%
Section 16 - Educational Expenses	71	5%	56	3%	2%
Section 17 - Subscriptions …	157	10%	111	6%	4%
Section 18 - Trips and Vacations	151	10%	109	6%	4%
Section 19 - Miscellaneous Expenses	161	11%	120	7%	4%
Section 20 - Expense Patterns	63	4%	48	3%	1%
Section 21 - Work Experience and Income	325	22%	279	15%	7%
Section 22 - Assets and Liabilities	41	3%	31	2%	1%

Looking at the specific sections in which respondents used records, we see in Table 31 that respondents in the record incentive groups used records at higher rates than respondents in groups not receiving record

incentives in almost every section. This was most pronounced in the utilities section where records were used by 59 percent of CUs offered incentives compared to only 39 percent of those not offered incentives.

Table 32: Record type used by incentive group, interview 1

	Record-Use Incentive Group		No Record-Use Incentive		
	CUs Using the Record	% Using	CUs Using the Record	% Using	Difference (Incent - None)
Bills	805	53.3%	589	32.2%	21.1%
Checkbook	413	27.3%	387	21.1%	6.2%
Personal finance	296	19.6%	268	14.6%	5.0%
Receipts	272	18.0%	167	9.1%	8.9%
Home File	9	0.6%	11	0.6%	0.0%
Contracts	66	4.4%	58	3.2%	1.2%
Bank statements	209	13.8%	163	8.9%	4.9%
Other	180	11.9%	165	9.0%	2.9%
None	55	3.6%	59	3.2%	0.4%

Finally, looking at the types of records used by these groups, small differences were found. Those receiving the incentive used bills at higher rates (53 percent) than those not receiving an incentive for record use (32 percent). All other record types were used at equivalent or slightly higher rates when record incentives were provided. All of these findings were significant at the 0.01 level for all record types with the exception of the Home File[35], Contracts and None types.

Survey Time

Interview time can be a measure of respondent burden with an increase in time potentially reflecting increased burden on the respondent. Consequently, it is important to balance any increase in interview time with improvements in data quality.

[35] The Home File is a filing folder that allows the respondent to keep receipts and records for the following interview.

Table 33: Total survey time (minutes)

	Mean	Median	Med. Diff
All Incentives	84.3	78.5	2.6
No Token	84.2	81.4	5.5
No Record	84.5	76.8	0.9
Control	81.3	75.9	

Analysis shows that interviews in all three incentives test groups were longer than interviews in the control group. Median interview length was highest in the two test groups that were offered a record use incentive. The incentive likely encouraged a higher level of records use, resulting in either more expenditures being reported or indicating more time used to locate the appropriate records and find the desired expenditures. In addition, the interviewer had to devote time to administering the incentive once it had been earned. The difference in median time between the test group that did not receive a record use incentive and the control group was considerably smaller than the differences between the other two test groups and the control group.

Table 34: Survey time (minutes) – FRONT Section

	Mean	Median	Med. Diff
All Incentives	10.9	6.4	1.8
No Token	10.4	6.3	1.7
No Record	10.6	5.9	1.3
Control	9.6	4.6	

At the beginning of the interview, in the FRONT section of the instrument where CU eligibility is determined, all three incentives test groups' time-in-section was longer than the control group's. The longer interview length is likely associated with the time interviewers had to devote to cases where the respondents had thrown away or lost their survey debit card incentive.

Table 35: Survey time (minutes) – BACK Section

	Mean	Median	Med. Diff
All Incentives	7.4	5.9	1.8
No Token	7.6	6.2	2.1
No Record	6.5	4.9	0.8
Control	5.7	4.1	

Similarly, all three incentive test groups took more time than the control group in the BACK section at the end of the interview where closeout occurs and the interviewer attempts to schedule the next interview. The longer survey time is likely associated with the time the interviewer must take activating the debit card incentive and inputting information into the survey instrument in regards to the records use incentive.

Table 36: Survey time (minutes) – Sections 1-22

	Mean	Median	Diff
All Incentives	61.5	55.8	-0.1
No Token	61.9	55.9	0.0
No Record	63.1	55.5	-0.4
Control	61.2	55.9	

There was very little difference in the actual survey time associated with recording expenditures and income between the test and control groups. This suggests that, although record use increase (Table 29), there is a possibility that the number of records used only increased marginally. Time spent recording expenditures and income may also be affected by the presence of satisficing by respondents who would normally be nonresponders, but are in the survey due to the presence of an incentive.

V. Conclusion

The analyses for this report found that respondents in the incentive test groups were more likely to respond to the survey, report similar quality of data, use records, and be more cooperative during the interview process. However, these gains were offset by operational issues, questions regarding their cost effectiveness, and potentially a higher nonresponse bias. Fixes for many of the operational issues, such as incentives being lost or thrown away, debit cards not working, and issues regarding the help desk, can be implemented with relative ease, but the questions of cost effectiveness remain.

Subsequently it is the recommendation of the team that the following be operation issues be addressed and additional analysis regarding the effect on incentives on nonresponse bias before proceeding with incentives in the Large Scale Feasibility Test:

1. Increase visibility of the Advance Letter and the incentives. One potential solution is to use Priority Mail envelopes which allow for increased visibility, which in theory, would increase

awareness of the survey, reduce the number of incentives that are accidently disposed of, and lend credence to the authenticity of the incentive.

2. Create more robust respondent materials for handling and cashing debit cards. Although debit cards are used widely, anecdotal evidence shows that certain demographic groups may have trouble handling the debit cards. Robust ancillary respondent material that more thoroughly explains how and where to use the debit cards would reduce respondent confusion and potentially decrease interview attrition.
3. Address how help desk calls are handled. The help desk line used during the test was unable to effectively assist respondents, specifically regarding use of the debit cards, leading to increased respondent dissatisfaction. Improving help desk center training, specifically in the area of debit card question resolution and developing a searchable database for the help desk staff to refer to in resolving frequently asked questions is imperative. In addition, since a majority of the help desk calls were in regards to debit card operational issues, employing the issuing bank's help desk line would increase the likelihood of the respondent's question being answered correctly and in a timely manner. At this time, Census policy states that all help desk calls are required to go through the Census Bureau; however, it is recommended that CE pursue a waiver for this policy.

References

1. To, N., (2014). Gemini Incentive Structure Review: Summary of Incentive Experiences, Internal Bureau of Labor Statistics Document.
2. McGrath, D., B. Chopova, C. Gomes, and N. McDermott, (2007). The Effects of Incentives on the Consumer Expenditure Diary, Internal Bureau of Labor Statistics Document.
3. Goldenberg, K., L. Tan, and D. McGrath, (2009). The Effects of Incentives on the Consumer Expenditure Interview Survey, Internal Bureau of Labor Statistics Document.
4. Caporaso, A., A. Mercer, D. Cantor, and R. Townsend, (2016). Monetary Incentives and Response Rates in Household Surveys: How much gets you how much?, 2016 Consumer Expenditures Survey Methods Symposium.
5. CE Incentives Team, (2006). Wrap-up of Procedures during the Test, Internal Bureau of Labor Statistics Document.
6. Ruffin, J., (2006). Incentives Wrap-up Report, Internal Census Bureau/Bureau of Labor Statistics Document.
7. CE Demographic Surveys Division, (2006). DSD's Report on the CE Incentive Test, Internal Bureau of Labor Statistics Document.
8. Gfroerer, J., J. Eyerman, and J. Chromy, (2002). Redesigning an Ongoing National Household Survey: Methodological Issues, Substance Abuse and Mental Health Services Administration, Office of Applied Studies.
9. Singer, E., J. Van Hoewyk, N. Gebler, T. Raghunathan, and K. McGonagle, (1999). The Effect of Incentives on Response Rates in Interviewer-mediated Surveys. Journal of Official Statistics.
10. Groves, R., and E. Peytcheva, (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis, Public Opinion Quarterly.
11. Kopp, B., B. McBride, and L. Tan, (2013). An Exploratory Study on the Association of Doorstep Concerns with Three Survey Quality Measures for the CE Interview Survey, Internal Bureau of Labor Statistics Document.
12. McBride, B. and L. Tan, (2014). Quantifying CHI Doorstep Concerns as Risk Factors of Wave 1 Nonresponse for the CE Interview Survey. Internal Bureau of Labor Statistics Document.

13. McGrath, D., (2005). Comparison of Data Obtained by Telephone versus Personal Visit Response in the U.S. Consumer Expenditures Survey.
14. Safir, A. and K. Goldenberg, (2008). Mode Effects in a Survey of Consumer Expenditures.
15. Guo, S. and M. Fraser, (2010). Propensity Score Analysis. Statistical Methods and Applications
16. Beckler, D. and K. Ott, (2006). Indirect Monetary Incentives with a Complex Agricultural Establishment Survey, 2006 Proceedings of the American Statistical Association.
17. Brick, M., J. Montaquila, M. Collins Hagedorn, S. Brock Roth, and C. Chapman. (2005). Implications for RDD Design from an Incentives Experiment, Journal of Official Statistics, 21, 2005, pp. 571-589.

.