Response Burden: What Predicts It and What is its Impact on Response Quality?

Ting Yan
Westat

August 31, 2015

Abstract

Concerns about the burden that surveys place on respondents have a long history in the survey field. As early as the 1920s, survey researchers and practitioners expressed concern about the potential negative impacts of response burden. However, a review of the response burden literature reveals that conceptualizations and measures of burden are still underdeveloped, and as a result findings from empirical research in this area remain equivocal.

Beginning in 2011, the Consumer Expenditure Quarterly Interview Survey (CEQ) fielded survey questions to measure respondents' perceptions of burden and their attitudes and perceptions about the survey. The current study examines these data to understand the factors that are associated with burden and the impact burden can have on data quality. The first phase of this work makes use of structural equation modeling to explore how survey characteristics, respondent characteristics, and respondent attitudes affect respondents' perception of the survey and subsequently perception of response burden. We found that low motivation, recall task difficulties, and challenging survey requests all directly contribute to respondents' perception of burden. In addition, low motivation contributes to response burden through its effect on respondents' negative perception of the survey. We also found that part of the measurement models are not equivalent across modes of data collection but the structural relationships are. The second phase of the study assessed the impact of response burden on CE data quality. We found that perceptions of burden increase the number of Don't Know and Refusal answers, especially for those interviewed on the phone. Removing the most burdened respondents did not significantly affect the expenditure estimates and could potentially lead to cost savings in data collection and post-survey processing. We discuss the implications of these findings for conceptualizations of survey response burden and for tailoring survey designs to balance burden considerations with quality and cost targets.

1. Introduction

Concerns about the burden that surveys place on respondents have a long history in the survey field. As early as the 1920s, survey researchers and practitioners expressed concern about the potential negative impacts of response burden. Bradburn pointed out the challenges of studying burden in his nominal paper on burden: "The topic of respondent burden is not a neat, clearly defined topic about which there is an abundance of literature" (1978: p49). About four decades later, burden is still considered as "*not* a straightforward area to discuss, measure, and manage" (Jones, 2012: p1). A review of the burden literature reveals that conceptualizations and measures of burden are still underdeveloped, and as a result findings from empirical research in this area remain equivocal.

Efforts to measure burden tend to fall into three broad categories. The first category measures properties of surveys/tasks that are believed to impose response burden, such as the length of an interview and the difficulty of the response task (Filion, 1981; Warriner, 1981; Hoogendoorn and Sikkel, 1998; Groves, Singer, and Corning, 1999; Singer et al., 1999; Hoogendoorn, 2004; Rostald, Adler, and Ryden, 2011). This category identifies likely sources of burden rather than measuring burden as perceived by respondents. The second category measures respondents' attitudes and beliefs toward surveys, such as interest in the survey, importance of the survey, and the perception of time and effort spent (Sharp and Frankel, 1983; Hoogendoorn, 2004; Stocke and Langfeldt, 2004; Fricker, Gonzalez, and Tan, 2011; Fricker, Kreisler, and Tan, 2012; Geisen, 2012). These respondent characteristics are potential mediators of the perception of burden, resulting in differential perceptions of burden across respondents, but they are not direct measures of burden themselves. The third category measures burden as perceived by respondents through respondent behaviors such as willingness to be re-interviewed and feeling of exhaustion and so on (Sharp and Frankel, 1983; Stocke and Langfeldt, 2004). This category measures the impact or effect of burden instead of burden itself.

A more important issue, however, is that these very different measurements of response burden reflect both the lack of and the need for a well-developed conceptual framework on burden. Direct measurement of burden is used in two studies. In a web survey to a convenience sample, burden is directly measured via a survey question asking respondents how burdensome answering survey questions on a particular web page was to them (Galesic, 2006). The Consumer Expenditure Interview Survey (CE) is the first and only government-sponsored large scale survey that also directly measures respondents' feeling of burden by asking how burdensome they found the survey was (Fricker, Gonzalez, and Tan, 2011; Fricker, Kreisler, and Tan, 2012; Fricker, Yan, and Tsai, 2014; Yan, Fricker, and Tsai, 2014). Too often, researchers and practitioners rely on loose definitions of burden, or continue to employ interview length as a proxy measure of burden (e.g., Groves et al., 1999; Rolstad et al., 2011). As the earliest conceptualization of burden, Bradburn's view of burden was multidimensional and reflected the influences of interview length, effort required of respondents, the frequency of interviews, and the amount of stress on respondents (Bradburn, 1978). He emphasized that burden is a subjective phenomenon – “the product of an interaction between the nature of the task and the way it is perceived by the respondent” (Bradburn, 1978; p36) – and suggested several possible factors that could influence respondents’ perceptions of the survey task (e.g., interest in survey topic). Haraldsen (2004) outlines a model where the subjective perception of burden is explicitly shown as an intermediate variable explaining the relationship between causes of response burden and data quality. Causes of response burden are further divided into survey properties and respondent characteristics. Haraldsen (2004) presented qualitative test results to shed light on survey properties and their impact on data quality. The middle part—the interaction between survey and respondent characteristics—was not tested at all.

Empirical research on burden shows that burden leads to unit nonresponse (e.g., Groves et al., 1999; Rolstad et al., 2011), panel attrition (e.g., Martin et al., 2001), item nonresponse (e.g., Warrier, 1991), break-offs (Galesic, 2006), and delayed responses (e.g., Giesen, 2012). However, there is no research yet looking into the impact of burden on data quality and statistical estimates.

To sum up, there exist three gaps in the burden research – undeveloped conceptualization of burden, lack of good measurement of burden, and lack of empirical research looking into the impact of burden on data quality and statistical estimates. This research combines Bradburn’s and Haraldsen’s work by defining burden as subjective perception and feelings of burden and attempts to fill the first gap on conceptualization and the third gap on relation between burden and data quality. This research consists of two phases. Phase 1 posits a path model that explicitly models the direct and indirect effects of survey features, respondent characteristics, and respondents’ perceptions of the survey on burden, in hopes of shedding light on which factors (or combination of factors) are most likely to result in response burden. Phase 2 examines the impact of burden on data quality and statistical estimates, while taking into consideration cost of data collection and post-survey processing.

For both phases of work, we take advantage of data from a large federal household survey – the U.S. Consumer Expenditure Survey (CE). Since 2011, CE includes a batch of survey questions at the end of the 5th interview to assess respondents’ perceptions about their survey experience. One of the survey questions directly asks respondents how burdensome the survey was to them, using a four-point fully-labeled unipolar scale. The scale runs from “very burdensome,” to “somewhat burdensome,” “a little burdensome,” and “not at all burdensome.” We combined cases that finished their 5th interview between October 2012 and March 2013 and used them in the analyses.

2. Phase 1Work

Phase 1 makes use of structural equation modeling (SEM) to test a model of burden that includes latent factors related to respondent motivation, respondent characteristics affecting the level of difficulty for answering CE questions, survey features and respondent perceptions of surveys, and to examine the causal relations (direct and indirect) between these factors and burden. The SEM is used to examine the nature of the relationship between these factors and respondents' perceptions of burden. We believe that this type of approach can significantly improve our understanding of burden by identifying characteristics of respondents that are most associated with high levels of burden given a particular survey feature (e.g., length). The ability to predict which respondents are at a greater danger of feeling burdened will help survey organizations to modify their survey protocols so as to reduce the likelihood of particular respondents feeling burdened and to reduce the negative impact of burden on data quality.

We used the SAS procedure PROC CALIS to estimate the model. We excluded cases with missing data on observed variables from the SEM estimation and analysis.

2.1. Burden Model

The SEM framework consists of two inter-related models: (1) the measurement model, which describes the assignment of the observed items (or indicators) to each unobserved latent factor; and (2) the structural model, which describes the relationship among the set of latent factors. Both models are explicitly defined by the analyst, and depicted in a path diagram.

For our structural model (see Figure 1), we examined one factor related to motivation, factors related to survey and task characteristics (task difficulty and survey request), one intermediate factor related to respondent perceptions of the survey, and the key dependent variable – respondents' report of burden. A description of the items used to construct the latent factors included in the structural model follows.

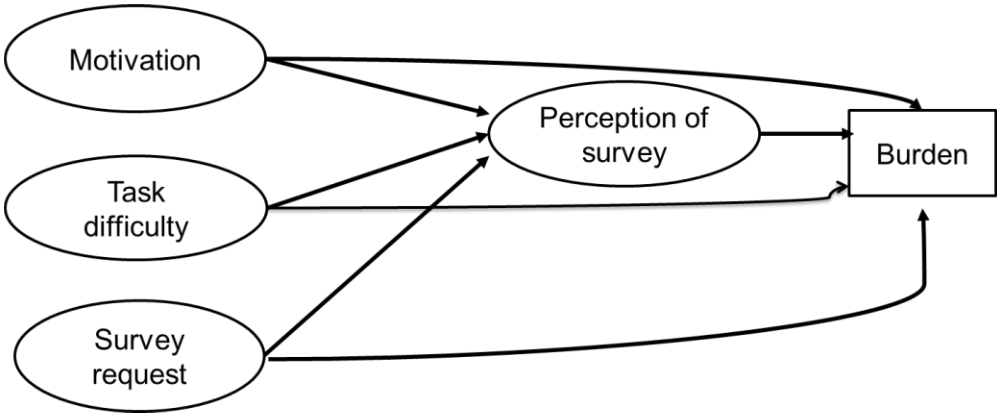

Figure 1. Structural Model of Burden

2.1.1 Measure of burden

Burden was measured through a survey item asking respondents directly how burdensome they felt the survey was. Higher values indicate higher level of burden. In the SEM, burden was treated as an observed variable with no measurement error. (We discuss this issue further in Section 3.4 below.) Shown in Table 1 are the percentages (and counts) of respondents endorsing each answer category to the burden item.

Table 1: Distribution of Responses to Burden Item

	Sample Count	**Percentage**
Very burdensome	645	10.6%
Somewhat burdensome	1,684	27.6%
A little burdensome	1,925	31.6%
Not at all burdensome	1,845	30.3%
Total	6,099	100%

2.1.2 Measure of motivation

Motivation in our model was measured as a latent construct with four indicators that are conceptually related to motivation. The first indicator, *high concern*, was a summary variable drawing from doorstep concerns data and represents the level of concerns expressed by sampled respondents at the doorstep. Higher values on this indicator denote higher level of concerns, and thus, lower level of motivation. The second indicator, *sensitivity of CE*, represents how sensitive respondents considered CE with higher values indicating higher level of sensitivity. The third indicator, *low trust*, denotes the extent to which respondents trusted in the US Census Bureau to safeguard the information they provided. Higher values on this indicator are associated with lower level of trust. The last indicator, *number of refusals expressed*, counts the number of times sampled respondents refused the survey request throughout their entire panel life (that is, across all five survey requests).

2.1.3 Measure of task difficulty

Task difficulty was measured as a latent construct with three indicators. The first indicator, *number of kids in household*, captures the number of children living in the same household with sampled respondents. The second indicator, *number of household members*, represents the number of people living in the household. We selected these two indicators because respondents living in large households (with more kids or more people) likely have more expenditures to report than those in small households, and because proxy reporting for other household members’ expenditures may be more difficult (i.e., burdensome) than simply reporting for oneself. The third indicator, *less than 65 years old*, is a dummy variable coded from age where 1 means that the respondent is less than 65 and 0 means that he/she is 65 or older. Older respondents are considered to have reduced cognitive capacity (Salthouse, 1991) and may find the same task as more difficult and burdensome than their younger counterparts (Krosnick, 1991).

2.1.4 Measure of survey request

Survey request was measured as a latent construct with four indicators. The first indicator, *duration of interviews*, sums up the duration of all interviews (in hours) completed by sampled respondents throughout his/her panel life. The second indicator, *number of interviews completed*, counts of the number of interviews completed by sampled households. The third indicator, *using information book*, counts the number of times respondents used the information book always or almost always. The last indicator, *using records*, counts the number of times respondents resorted to records almost always when answering the expenditure questions.

2.1.5 Measure of perception of survey

Perception of survey was measured as a latent construct with four indicators. The first indicator, *too many rounds of interview*, reflects respondents’ perception about the number of interview requests posed to them. Higher values indicate that respondents considered that they have been asked to participate in too many rounds of interviews. The second indicator, *interview too long*, represents respondents’ perception of the length of the CE, with higher values indicating that they consider the survey too long. The third indicator, *survey not interesting*, looks at respondents’ perception of the CE and high values indicate that respondents considered the CE less interesting. The last indicator, *survey difficult*, represents the extent to which respondents considered the CE to be difficult.

2.2. SEM Results

2.2.1 Model Fit Statistics

We examined several model fit statistics. The Chi-square test indicated a poor fit with the data (x^2 (94)=1874, *p*<.0001), though this measure can be an overly sensitive test of global fit with large sample sizes as we have in our study (Byrne 1998; Kline 1998). The second index of overall fit, the Standardized Root Mean Square Residual (SRMSR), was 0.049. According to O’Rourke and Hatcher (2013), SRMSR values less than 0.055 suggests a good fit and values less than 0.09 are suggestive of fair or adequate fit. Therefore, our SRMSR value indicates a good model fit.

We also looked at two parsimony indices. The Root Mean Square Error of Approximation (RMSEA) was 0.056, indicating a fair or adequate fit (O'Rourke and Hatcher, 2013).[1] The Adjusted Goodness of Fit Index (AGFI) was larger than 0.90, reflecting a good fit (AGFI=0.944). The Bentler Comparative Fit Index, an incremental index, was 0.920 and was above the traditional 0.90 cut-off value, suggesting a good fit.

Looking across all indices, we considered our models to reflect a rather good fit to the data.

2.2.2 Measurement Model

Estimates from our SEM measurement model are shown in Table 2. The unstandardized factor loadings for each item with its associated latent variable are statistically significant and the standardized loadings are generally sizeable (i.e., greater than 0.3). All of the loadings are in the expected direction. For example, *lower respondent motivation* was associated with respondents who consider the survey as sensitive, have low or no trust in the survey organization, express more concerns at the doorstep, and have ever refused the survey request more often. Large households with more kids, large households with more household members, and respondents younger than 65 all positively contribute to *task difficulty*. *Challenging survey request* was positively associated with longer interviews, more interviews completed, using information book, and using records during the interview. Negative *perceptions of the survey* were reflected in respondents complaining having too many rounds of interviews, interviews being too long, survey less interesting, and survey more difficult.

2.2.3 Structural Model

With our measurement model validated, we examined the hypothesized structure of our latent factors. We began by looking at the direct effects of each factor on the other model factors (see Table 3). All factors had significant direct effects on burden in the right direction. Not surprisingly, *lower motivation, more difficult task, more challenging survey requests, and negative impressions of survey* all lead to higher level of perceived burden.

[1] RMSEA values less than .055 indicates a good fit and values less than .09 suggests a fair and adequate fit (O'Rourke and Hatcher, 2013)

Table 2: Model Estimates from the Measurement Models for Pooled Cases

Measurement Model		Standardized Estimates	Unstandardized Estimates	S. E.	*p*-value
Factor	Indicator				
Low Motivation	Level of doorstep concerns	0.340	1.000		
	Sensitivity of CE	0.612	2.108	0.097	<0.0001
	Low trust in survey organization	0.446	1.413	0.072	<0.0001
	Number of refusals expressed	0.168	0.176	0.017	<0.0001
Difficult Recall Task	Number of children in Household	0.844	1.000		
	Number of household members	0.867	0.570	0.018	<0.0001
	Respondent Less than 65	0.334	0.155	0.007	<0.0001
Challenging Survey Request	Duration of interviews	0.886	1.000		
	Number of interviews completed	0.762	1.419	0.042	<0.0001
	Using Information Book	0.350	0.707	0.030	<0.0001
	Using records	0.362	0.639	0.027	<0.0001
Negative Perception of Survey	Too many rounds	0.639	1.000		
	Survey too long	0.621	0.947	0.025	<0.0001
	Survey not interesting	0.584	1.814	0.050	<0.0001
	Survey difficult	0.465	1.171	0.039	<0.0001

The remaining effects shown in Table 3 are also generally in the expected direction. For example, respondents with low motivation were more likely to have a negative impression of the survey than those who were more motivated. Similarly, greater task difficulty was associated with more negative perceptions of the survey, although the effect is only marginally statistically significant. The one puzzling finding is that more challenging survey request (longer interviews, more surveys completed, using information book and records during the survey) was associated with less negative feelings about the survey. One possible explanation for this finding is that individuals are more motivated with a task that is intricate, challenging, and enriching (e.g., Campbell, 1988). Whatever the reason for the direction of the effects, the size of these effects is very small.

Table 3: Model Estimates from the Structural Model of Burden (for Pooled Cases)

		Standardized	Unstandardized		
Structural Model		Estimates	Estimates	S. E.	*p*-value
Factor	Effect on				
Low Motivation	Negative Perception of Survey	0.868	0.905	0.047	<0.0001
	Burden	0.454	1.466	0.227	<0.0001
Difficult Task	Negative Perception of Survey	0.031	0.011	0.006	<0.10
	Burden	0.029	0.032	0.013	<0.05
Challenging Survey Request	Negative Perception of Survey	-0.066	-0.029	0.008	<0.001
	Burden	0.031	0.042	0.016	<0.05
Negative Perception of Survey	Burden	0.337	1.042	0.201	<0.001

We next estimated indirect and total effects of our model factors. Indirect effects are mediated by at least one intervening variable and total effects are equal to the sum of direct and indirect effects. Table 4 summarizes the results of this decomposition of effects.

As shown in Table 4, *low motivation*, *difficult task*, and *negative perception of survey* had significant overall positive effects on burden. Contrary to views commonly held in the survey field, the usual-suspect causes of burden such as *challenging survey request* had no significant overall effects on burden. The direct effects of *challenging survey request* are positive and statistically significant at the 0.05 level. However, the indirect effects of this factor through *negative perception of survey* are negative and statistically significant. As a result, the sum of these two effects essentially cancelled out each other, yielding small and non-significant total effects.

Table 4: Decomposition of Effects of Latent Factors on Burden

	Total Effects	Direct Effects	Indirect Effects
Low Motivation	0.747***	0.454***	0.292***
Difficult Task	0.040**	0.030**	0.011
Challenging Survey Request	0.009	0.031*	-0.022**
Negative Perception of Survey	0.337***	0.337***	0

Note: *p<0.05;**p<0.01; ***p<0.001

2.3 Multiple Group Analysis

We examined whether the burden model is invariant across respondents interviewed in different modes of data collection. In other words, we are interested in examining whether the same burden model holds for people attempted mostly in person and people attempted mostly over the phone. For the purpose of this analysis, for each sampled respondent, we summed up the number of contact attempts made in person and the number of contact attempts made over telephone across all contact attempts and across all waves of interviews. Then we looked at the ratio of the sum of contact attempts in person and the sum of contact attempts over the phone. Based on this ratio, we divided respondents into two groups. A total of 3,584 respondents were classified as the “mostly in-person” group because they were attempted in person more often than over the telephone. 2,515 respondents were grouped together as the “mostly by phone” group as they were attempted over the phone more often than in person. Cases who were attempted equally often in person and by phone were removed from the analysis.

We conducted multiple group analysis (MGA) to test measurement invariance at different levels. The first level of invariance is a model where no constraint was imposed on any parameters across the two groups. If this configural model fits the data, it is used as the base model for model comparison. As shown in Table 5, our configural model fits the data relatively well, evidenced by two model fit statistics (RMSEA and CFI), suggesting that the overall relationships among indicators and factors have the same structure and direction across respondents attempted in different modes of data collection.

Next we constrained all factor loadings to be equal across the two groups of respondents (metric invariance). This metric invariance model is evaluated against the configual invariance model. Although the metric invariance model has a good fit to the data, the model fit between the two nested models (the configural invariance model and the metric invariance model) are statistically different, suggesting that some factor loadings are not invariant across groups. In other words, some relationships between factors and their indicators are not equal across respondents who were attempted mostly in person and those who were attempted mostly over the phone.

We examined the equality constraints on factor loadings one at a time and identified and removed the constraints on seven factor loadings that caused the lack of fit (specifically, loadings for number of children in household, number of household members, respondent less than 65, duration of interviews, number of interviews completed, using Information Book, and survey difficult) so that partial invariance of factor loadings is established. The partial metric invariance model has a better fit than the configural invariance model (smaller RMSEA and larger CFI). Furthermore, the model fit (between the partial invariance model and the configural invariance model) are not statistically significant, indicating that the partial (weak) invariance model is preferred over the configural invariance model.

Given the acceptance of the model with partial metric invariance, we further constrained the relationships among the factors in the model to be equal across respondents attempted in different modes of data collection (structural invariance). The model with structural invariance fits the data well and the model fit between the structural invariance and the partial metric invariance model is not statistically significant, indicating equivalence of relationships among latent factors across modes of data collection.

Table 5. Multiple Group Analysis

Models for Comparison	χ^2	**DF**	***p*-value**	**RMSEA**	**CFI**	$\Delta\chi^2$	**ΔDF**	**p-value**
Configural Invariance	1913	188	<0.0001	0.0549	0.9159			
Metric Invariance	1999	199	<0.0001	0.0545	0.9121	86	11	<0.0001
Partial Metric Invariance	1918	194	<0.0001	0.0540	0.9233	5	6	0.52
Structural Invariance	1924	201	<0.0001	0.0530	0.9234	6	7	0.54

2.4 Phase 1 Conclusions

In Phase 1, we developed and tested a model that assumes that burden is a subjective phenomenon, affected by objective survey features, objective respondent characteristics that are related to task difficulty, respondent motivation, and respondent subjective perception of the CE. We used structural equation modeling to assess how well these data fit latent factors we hypothesized to be important contributors to perceived burden, and then evaluated the impact those factors had on burden.

The results of this study validated our underlying measurement model – our indicators were all significantly related to their associated latent variables in the expected direction. Results of our structural model showed that respondents' motivation, respondents' objective characteristics related to task difficulty, and respondents' subjective perceptions of the survey task had a significant direct impact on burden as well as significant overall effects on burden. The objective survey features themselves had a significant direct impact on burden, but this direct effect is cancelled out by indirect effects through respondent perception of the survey, producing small and non-significant overall effects on burden.

Furthermore, we conducted multiple group analysis to test whether the same burden model holds for people attempted in different modes. We found that some factor loadings are not equivalent for respondents attempted mostly in person vs. those attempted mostly by phone. However, the structural relationships among the latent factors hold whether respondents were attempted most in person or by phone.

A key limitation of this study was that there was only one survey question asking directly about the feeling of burden. As a result, the SEM treated this observed indicator as error-free. Of course, this is a strong assumption, and almost certainly violated. We used an alternative approach suggested in Kline (1998: p264-266) in which we re-specified our model by treating the observed burden variable as an observed indicator of a latent burden factor. We used the Survey Quality Predictor program (http://sqp.upf.edu/) to obtain an estimate of the quality and the error of the burden item, recognizing that the SQP estimate of the error term is at best only an approximation of the true error. We then reran the SEM using the error estimates from the SQP to fix the measurement error term of the observed burden indicator and fixing the loading of the burden indicator on the latent burden construct to be 1. The conclusions remained unchanged.

3. Phase 2 Work

The Phase 2 work has two goals. One is to empirically examine the impact of response burden on data quality and the second is to demonstrate how response burden, once measured, can be used in practice to allow researchers and data users to deal with data quality. As the burden question is asked at the 5th interview, we focused on the quality of data collected from the 5th interview. For this analysis, we excluded cases with missing data on the burden question and on expenditure questions.

We also looked to see if there was any mode difference with regards to the impact of response burden on data quality. Again, for each sampled respondent, we summed up the number of contact attempts made in person and the number of contact attempts made over the phone across all contacts attempted the 5th interview. We then divided respondents into two groups – one group ("mostly in-person") was attempted in person more often than over the phone and the other group ("mostly by phone") was attempted by phone more often than in person. The proportions of cases reporting different levels of burden were shown in Table 6.

Table 6. Distribution of Responses to Burden Question by Mode Group

	Mostly telephone	**Mostly in-person**	**Total Cases**
Very burdensome	13.7%	9.5%	726
Somewhat burdensome	32.4%	24.9%	1823
A little burdensome	32.6%	30.3%	2056
Not at all burdensome	21.3%	35.3%	1990
Total Cases	2426	4169	6595

3.1. Impact of Burden on Indirect Indicators of Data Quality

Ideally we would examine biases in the expenditure data, which is a direct measure of the quality of expenditure data, if true expenditure were available. Given that it is infeasible, if not impossible, to collect true expenditure data from the CE respondents, we used two indirect indicators to evaluate the quality of expenditure data. The CE publishes two variables in the public use paradata files – numdk and numrf – to indicate the number of “Don’t Know” answers and the number of “Refused” answers provided by respondents. The two variables are deleted from the 2013 public use paradata files. As a result, we only used the 2012 data to examine the impact of response burden on the number of “Don’t Know” and “Refused” answers.

One apparent trend from Figure 2 is the positive relation between burden and the number of “Don’t Know” and “Refused” answers; the more burdensome respondents felt, the more “Don’t Know” and “Refused” answers respondents provided. The trends hold for both respondents attempted mostly in-person and for those attempted mostly by the phone. However, there is some indication that the impact of burden on “Don’t Know” and “Refused” answers is stronger for cases attempted mostly by phone than in person.

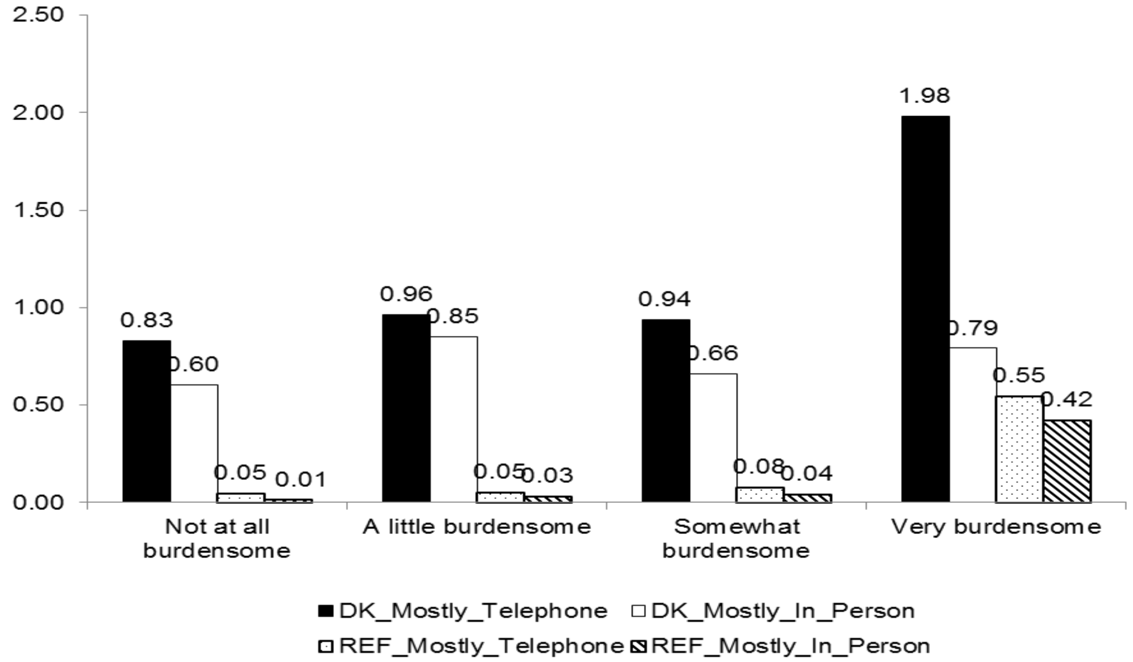

Figure 2. Average Number of “Don’t Know” and “Refused” Answers by Level of Burden and by Mode Group

In order to formally test the interaction between burden and mode groups, we fit two regression models with the number of “Don’t Know” answers as the dependent variable for one model and the number of Refused answers as the dependent variable for the second model. Both models included burden, mode groups, and respondent characteristics that are shown in the survey literature to be related to respondents’ likelihood to provide “Don’t Know” and “Refused” answers (e.g., respondent age, education, family size, number of kids under 18, urbanicity, duration of interview, level of doorstep concerns, whether or not respondents were converted refusers, whether or not respondents used the information book, whether or not respondents used records).

The regression results confirmed the trends observed in Figure 2. After controlling for respondent characteristics related to providing “Don’t Know” or “Refused,” perceived burden still has a significant impact on the number of “Don’t know” answers (F(3,3340)=5.89, p<0.001) and “Refused” answers (F(3,3340)=39.91, p<0.0001). In addition, the interaction between perceived burden and the mode groups is statistically significant for the model predicting the number of “Don’t know” answers (F(3,3340)= 5.93, p<0.001), but not for the model on the number of “Refused” answers (F(3,3340)= 0.80, p=0.49).

3.2. Impact of Burden on Statistical Estimates of Expenditure

We next examined whether or not perceived burden affects statistical estimates. In particular, we examined whether or not removing burdened out respondents (who reported “very burdensome” to the burden question) would affect the resultant expenditure estimates. We selected 13 expenditure variables published in the CE public-use data files and, for each expenditure variable, we calculated the mean estimates using all cases and the mean estimates after removing burdened-out respondents. We used SAS’s procedure PROC SURVEYMEANS to calculate the mean estimates and took into account both the clustering of cases and weights.

Displayed in Table 7 are estimates of mean expenditures (in dollars) when burdened out respondents are included vs. excluded, by mode group. For instance, the mean total expenditure is $8,282 for all respondents attempted mostly in-person, including those who found the survey very burdensome, with a standard error of $506. After removing those who were burdened out from the calculation, the estimate of the mean total expenditure is $8,308 with a standard error of $471, producing a difference of $26 in the mean estimate. Obviously, the confidence intervals for the two mean estimates overlap, suggesting that removing burdened out respondents would only produce a small and non-significant shift in the mean estimate for the total expenditure variable. Looking at the columns labeled as “Difference” in Table 7, it is clear that removing burdened out respondents doesn’t seem to have much impact on the mean estimates for all 13 expenditure variables. In addition, the differences in mean estimates due to excluding burdened out respondents are comparable across the mode groups. In other words, regardless of whether respondents were attempted mostly in person or over the phone, removing respondents who were burdened out would not have an impact on the mean estimates.

Table 7. Estimates of Mean Expenditures (in Dollars) With and Without Burdened-out Respondents, by Mode Group

	MOSTLY IN PERSON			MOSTLY TELEPHONE		
	WITH (n=4169)	WITHOUT (n=3775)	Difference (n=394)	WITH (n=2426)	WITHOUT (n=2094)	Difference (n=332)
Total Expenditure	8,282 (506)	8,308 (471)	26	8,272 (436)	8,206 (449)	-66
Food	1,229 (59)	1,216 (47)	-13	1,249 (56)	1,243 (59)	-6
Alcoholic beverages	69 (4)	71 (4)	3	59 (4)	60 (4)	1
Housing	2,566 (244)	2,565 (235)	-1	2,813 (266)	2,787 (255)	-26
Apparel and services	173 (18)	173 (15)	0	178 (15)	181 (15)	4
Transportation	1,547 (47)	1,566 (50)	19	1,429 (56)	1,416 (51)	-13
Health care	580 (17)	585 (17)	6	564 (21)	565 (21)	1
Entertainment	390 (19)	386 (10)	-3	365 (21)	370 (21)	5
Personal care	49 (5)	49 (5)	1	51 (6)	51 (6)	-1
Reading	20 (1)	20 (1)	0	16 (1)	17 (1)	1
Education	190 (34)	190 (34)	0	217 (39)	204 (31)	-13
Tobacco	62 (4)	65 (3)	3	45 (6)	47 (5)	2
Miscellaneous	95 (8)	100 (8)	5	81 (6)	81 (7)	1
Cash contributions	366 (57)	368 (58)	2	274 (19)	266 (17)	-8
Pensions	947 (78)	953 (74)	6	932 (76)	918 (85)	-14

3.3. Burden and Cost Considerations

The prior two sections demonstrate that respondents who perceived the survey to be very burdensome provided more missing data (more “Don’t Know” and “Refused” answers) and that removing them from the analysis doesn’t change the mean expenditure estimates much. In this section, we considered a hypothetical situation where data from burdened-out respondents were *not* collected and compared it to the real-life situation where effort were spent on these respondents to collect their data.

As shown in Table 8, if burdened respondents were *not* collected, we would end up with a total of 5,870 completed interviews, a reduction of 11 percentage points in the number of completed interviews compared to the real-life situation. However, not collecting data from burdened cases would translate into a reduction of 13 percentage points in the total number of attempts needed at Wave 5 and, in particular, a reduction of 13percentage points in the total number of contact attempts made in-person and 12 percentage points reduction in the total number of contact attempts by phone. Furthermore, not collecting data from burdened cases means that we would not need to convert 234 refusers – a reduction of 31 percentage points in refusal conversion effort – and we would not need to spend 753 production hours to administer the CE to these people (a reduction of 10 percentage points in interviewer production hours). Using only data from 2012, we would have 18% fewer "Don't Know" answers and 60% fewer "Refused" answers to be edited and imputed, if we did not collect data from respondents who felt very burdensome.

Table 8: Cost Implications of Not Collecting Data from Burdened-out Respondents

	All Cases	**No Burdened Cases**	**Differences**	**% Change**
# of Completed interviews	6,596	5,870	726	11%
	W5 Data collection effort			
Total number of attempts at W5	23,652	20,694	2,958	13%
Total number of in-person attempts at W5	10702	9357	1,345	13%
Total number of attempts by telephone at W5	12,950	11,337	1,613	12%
Total number of W5 Refusers converted	744	510	234	31%
Total number of W5 interview hours	7,231	6,478	753	10%
	W5 Post-survey processing effort*			
Total number of "Don't know" to be edited/imputed	2,829	2,321	508	18%
Total number of "Refused" to be edited/imputed	302	120	182	60%

Note: *Post-survey processing effort is based on 2012 data only.

3.4. Phase 2 Conclusions

The phase 2 work fills a gap in burden research by looking into the relation between burden and data quality. We found that respondents who felt very burdensome provided more “Don’t Know” and “Refused” answers to the expenditure questions. The difference in the number of “Don’t Know” answers by level of perceived burden is stronger for respondents attempted mostly over the phone than respondents attempted mostly in-person. We also demonstrated that, regardless of how respondents were attempted, removing burdened-out respondents didn’t have much impact on the mean expenditure estimates. In addition, not collecting data from these respondents would result in savings in the cost of data collection by reducing the total number of attempts needed, interviewer production hours, and number of refusers to be converted. It would also reduce the cost of post-survey processing by reducing the number of “Don’t know” and “Refused” answers to be edited and imputed after data collection.

One limitation of the work lies in that true expenditure data are unavailable; as a result, we wouldn’t be able to examine bias in the expenditure data by level of perceived burden. Indirect indicators of data quality (e.g., Don’t Know and Refused answers) were studied instead. Furthermore, we only examined changes in mean expenditure estimates when excluding vs. including burdened out respondents. Other statistical estimates (e.g., over-time changes in expenditures, regression coefficient estimates) are not examined, which limits the generalizability of our findings.

4. General Conclusions and Discussion

We identified three gaps in the burden research – undeveloped conceptualization, lack of good measurement, and lack of empirical research examining impact of burden on data quality. Prior research on response burden often has relied on inadequate conceptualizations and measures of burden. As a result, it is difficult, based on the small empirical literature that exists, to make firm predictions about survey features or respondent characteristics that are most likely to give rise to burden. In order to tackle the gaps in the research, we focused the first phase of our work on the understanding of the conceptualization of burden and the 2nd phase on the impact of burden on data quality.

The phase 1 work employed the Structural Equation Modeling (SEM) to test a model of burden that includes latent factors related to respondent motivation, respondent characteristics affecting the level of difficulty for answering CE questions, survey features and respondent perceptions of surveys, and to examine the causal relations (direct and indirect) between these factors and burden. The findings support the notion of burden as a subjective, multidimensional phenomenon and showed that respondent motivation, respondent perception of survey, and respondent characters related to task difficulty all have significant overall effects on burden. Survey request features, usually used as a measure of burden, have significant direct effects on burden, but the direct effects are canceled out by indirect effects via respondent perception of survey, leading to small and non-significant overall effects on burden. In addition, we found that, although some relationships between indicators and latent factors are not equivalent across respondents attempted in different modes of data collection, the relationships between latent factors and burden remain the same. This is encouraging as the survey field worries that the mode of data collection affects how respondents feel about the burden of the survey. Our results demonstrate that the modes of data collection do not affect the paths leading to the perception of burden; the same set of factors have the same impact on burden regardless whether respondents were attempted mostly by phone or in person.

Our analysis was made possible because of the CEQ's unique datasets which not only contain information about respondent reactions to the survey but also one survey item directly asking respondents how burdensome they felt. The approach of directly measuring the perception of burden with a single survey item was also used in Galesic (2006), who found that later web pages, pages where respondents spent longer time, and more open-ended questions increased respondents' perception of burden and burdened respondents were more likely to break off than those with a lower level of burden. Our results together with those of Galesic (2006) demonstrate that the approach of directly measuring burden through a single survey item is promising.

We hope that our results encourage other surveys to collect and disseminate similar data on burden. When possible, we strongly encourage other surveys to include at least one survey item directly measuring the level of perceived burden, as what is done in the CE. If it is not possible to add even one survey item to measure burden, we strongly encourage survey practitioners and researchers to take advantage of information tapping into the subjective and attitudinal evaluations and reactions to identify respondents at a greater risk of being burdened out. Our results find that respondents with low motivation and negative evaluation of survey report higher level of burden. We encourage survey practitioners and researchers to make use of existing paradata that reflect either motivation or perception of survey (e.g., the doorstep concerns data) upfront to identify those prone to feeling burdened out and to take steps to increase their motivation or reduce their negative evaluation of the survey.

The Phase 2 work adds to the literature on how measured burden can and should be used in practice to investigate the impact of response burden on data quality. We showed that respondents who answered "very burdensome" to the burden question provided more missing data and took more recruitment effort. Removing these cases doesn't have an impact on the resultant mean expenditure estimates. We encourage future research to take into account these findings when developing tailored or responsive survey designs. One example, for instance, is for longitudinal surveys to reallocate survey resources for the new wave of interview based on the level of perceived burden (measured directly or through proxy indicators such as doorstep concerns). Even in cross-sectional surveys, our results indicate that survey organizations should reconsider their stopping rules when recruiting respondents at different levels of burden.

The results from both phases of work provide significant and critical additions to the survey literature. However, a key limitation of the work is that the burden question is asked only at the fifth interview – the last interview at the end of respondents' panel life. Consequently, burden information is only available for those who completed the fifth interview and *not* available for attriters who dropped out of the panel before the 5th interview. Respondents who completed the fifth interview (and especially those who completed all five interviews) are likely to be more cooperative and more motivated than those who did not. What is more, attrition is likely a consequence of sampled respondents feeling burdened out, as shown in Galesic (2006). As a result, our findings are based on data from a pool of respondents with higher motivation and cooperation, which limits the generalizability of our findings to another pool of respondents with low motivation and cooperation. Furthermore, the CE is a large-scale government-sponsored longitudinal survey and has established its legitimacy and importance. Our findings may not generalize to other surveys in a different setting (e.g., cross-sectional surveys sponsored by a private organization). We do hope that our work will call researchers' attention to the burden research and will motivate more researchers to focus on the burden research.

Acknowledgements

This research was funded by the ASA/NSF/BLS Research Fellowship awarded to Ting Yan. The author thanks ASA/NSF/BLS Research Fellowship for supporting the work. The author thanks the BLS staff for coordinating all logistics and paperworks and the CE for making burden data available. The author likes to thank two BLS collaborators – Scott Fricker and Shirley Tsai – for working with her and assisting her in so many aspects throughout the fellowship. This research can’t be conducted without them.

References

Bradburn, N. (1978). Respondent Burden. *Health Survey Research Methods Proceedings*, 49–54.

Byrne, B. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, D. J. (1988). Task Complexity: A Review and Analysis. *Academy of Management Review*, 13, 40–52.

Filion, F. L. (1981). Importance of Question Wording and Response Burden in Hunter Surveys. *The Journal of Wildlife Management*, 45(4), 873-882.

Fricker, S., Gonzalez, J., & Tan, L. (2011). Are You Burdened? Let's Find Out. *Paper Presented at the Annual Conference of the American Association for Public Opinion Research*, Phoenix, AZ.

Fricker, S., Kreisler, C., and Tan. L. (2012). An Exploration of the Application of PLS Path Modeling Approach to Creating a Summary Index of Respondent Burden. *Paper presented at the Joint Statistical Meeting*, San Diego, CA.

Fricker, S., Yan, T., and Tsai, S. (2014). Response Burden: What Predicts It and Who is Burdened Out? In *JSM Proceedings*, *Survey Methods Research Section*, Alexandria, VA: American Statistical Association, pp. 4568-4577.

Galesic, M. (2006). Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey. *Journal of Official Statistics*, *22*, 313-328.

Geisen, D. (2012). Exploring Causes and Effects of Perceived Response Burden. Paper presented at the International Conference on Establishment Surveys. Paper accessed at http://www.amstat.org/meetings/ices/2012/papers/302171.pdf

Groves, R., Singer, E., and Corning, A. (1999). A Laboratory Approach to Measuring the Effects on Survey Participation of Interview Length, Incentives, Differential Incentives, and Refusal Conversion. *Journal of Official Statistics*, 15, 251-268.

Haraldsen, G. (2004) Identifying and Reducing Response Burden in Internet Business Surveys. *Journal of Official Statistics*, 20, 393-410.

Hoogendoorn, A. W. (2004). A Questionnaire Design for Dependent Interviewing that Addresses the Problem of Cognitive Satisficing. *Journal of Official Statistics*, 20, 219–232.

Hoogendoorn, A.W. and Sikkel, D. (1998). Response Burden and Panel Attrition. *Journal of Official Statistics*, 14, 189–205.

Jones, J. (2012). Response Burden: Introductory Overview Literature. Paper presented at the International Conference on Establishment Surveys. Paper accessed at http://www.amstat.org/meetings/ices/2012/papers/302289.pdf

Kline, R. B. (1998). *Structural Equation Modeling*. New York: Guilford Press.

Krosnick, J.A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5, 213–236

Martin, E., Abreu, D., and Winters, F. (2001). Money and Motive: Effects of Incentives on Panel Attrition in the Survey of Income and Program Participation. *Journal of Official Statistics,* 17, 27-284.

O'Rourke, Norm and Hatcher, Larry (2013). *A Step-by-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling.* SAS Institutes.

Rostald, S., Adler, J., and Ryden, A. (2011). Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-analysis. Value in Health, 14, 1101–1108.

Salthouse, T. A. (1991). *Theoretical Perspectives on Cognitive Aging.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sharp, L.M. and Frankel, J. (1983). Respondent Burden: A Test of Some Common Assumptions. *Public Opinion Quarterly*, 47, 36-53.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathen, T. and McGonagle, K. (1999). The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys. *Journal of Official Statistics*, 15, 217-230.

Stocke, V., and Langfeldt, B. (2004). Effects of Survey Experience on Respondents' Attitudes Towards Surveys. *Bulletin de Methodologie Sociologique*, 81, 5-32.

Yan, T., Fricker, S., and Tsai, S. (2014). The Impact of Response Burden on Data Quality in a Longitudinal Survey. Paper presented at the International Total Survey Error Workshop, Washington, DC.

Warriner, G. K. (1981). Accuracy of self-reports to the burdensome question: survey response and nonresponse error trade-off. Quality & Quantity, 25, 253–269.