ABSTRACT

Title of Document:	TESTING FOR PHASE CAPACITY IN SURVEYS WITH MULTIPLE WAVES OF NONRESPONDENT FOLLOW-UP.
	Taylor H. Lewis, Doctor of Philosophy, 2014
Directed By:	Professor Frauke Kreuter, Professor Partha Lahiri, Joint Program in Survey Methodology

To mitigate the potentially harmful effects of nonresponse, many surveys repeatedly follow up with nonrespondents, often targeting a particular response rate or predetermined number of completes. Each additional recruitment attempt generally brings in a new wave of data, but returns gradually diminish over the course of a fixed data collection protocol. This is because each subsequent wave tends to contain fewer and fewer new responses, thereby resulting in smaller and smaller changes on (nonresponse-adjusted) point estimates. Consequently, these estimates begin to stabilize. This is the notion of phase capacity, suggesting some form of design change is in order, such as switching modes, increasing the incentive, or, as is considered exclusively in this research, discontinuing the nonrespondent follow-up campaign altogether. This dissertation consists of three methodological studies proposing and assessing various techniques survey practitioners can use to formally test for phase capacity. One of the earliest known phase capacity testing methods

proposed in the literature calls for multiply imputing nonrespondents' missing data to assess, retrospectively, whether the most recent wave of data significantly altered a key estimate. The first study introduces an adaptation of this test amenable to surveys that instead reweight the observed data to compensate for nonresponse. A general limitation of methods discussed in the first study is that they are applicable to a single point estimate. The second study evaluates two extensions, each with the aim of producing a universal, yes-or-no phase capacity determination for a battery of point estimates. The third study builds upon ideas of a prospective phase capacity test recently proposed in the literature attempting to address the question of whether an imminent wave of data will significantly alter a key estimate. All three studies include a simulation study and application using data from the 2011 Federal Employee Viewpoint Survey.

TESTING FOR PHASE CAPACITY IN SURVEYS WITH MULTIPLE WAVES OF NONRESPONDENT FOLLOW-UP.

By

Taylor Hudson Lewis

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Frauke Kreuter, Co-Chair
Professor Partha Lahiri, Co-Chair
Research Professor Richard Valliant
Research Professor James Wagner
Professor Michael Rendall

© Copyright by
Taylor Hudson Lewis
2014

Acknowledgements

In a strange juxtaposition of emotions, it is with both relief and sadness that I say completing this dissertation marks the end of a major chapter in my life. My graduate studies in the Joint Program in Survey Methodology (JPSM) began in the summer of 2006 when I took a course on questionnaire design immediately preceding my official entrance into the master's program that fall. Although I took one year off between completing the master's program in May 2009 and beginning the doctoral program in August 2010, I have considered myself a JPSM graduate student for the past eight years, my lengthiest tenure at any single educational institution. It has been a truly formative period. I have benefited tremendously from the outstanding teaching and mentoring received along the way from JPSM faculty. I took away a lot from each and every course, particularly Steve Heeringa's Analysis of Complex Survey Data and Richard Valliant's Case Studies in Sampling and Weighting. Both were breakthrough courses helping me establish a solid grasp on the necessary adaptations when applying traditional statistical methods on complex survey data and the theory underpinning nonresponse adjustment techniques, respectively. It therefore came as no surprise to learn that lecture materials for both courses were subsequently developed into textbooks—the Heeringa et al. (2010) and Valliant et al. (2013) references appearing in this dissertation—both of which have been invaluable resources and will remain as such in the future.

Meeting the requirements of a PhD is a formidable enough challenge, let alone doing so while employed full-time. This was only possible because of my good

fortune to work for understanding and accommodating immediate supervisors. In particular, I would like to acknowledge Gary Lukowski and Kimya Lee of the U.S. Office of Personnel Management and Meena Khare of the National Center for Health Statistics (NCHS) for their unwavering support as I pursued my graduate studies, a pursuit that often forced me to abide by a nontraditional work schedule or be granted generous telecommuting arrangements. I also owe a debt of gratitude to the various individuals at the U.S. Office of Personnel Management who helped secure my approval to use data from Federal Employee Viewpoint Survey to evaluate the methods proposed as a part of this research.

I am very appreciative that Frauke Kreuter and Partha Lahiri agreed to serve as co-chairs of my dissertation committee. Frauke has been a strong advocate on both a personal level and about the ideas I began positing as this dissertation took shape. She taught me to confidently address each of the program's hurdles and suppress any internal doubt as to whether there was sufficient time to prepare by reminding me on several occasions how "life is *always* busy." I am also grateful for her unrivaled skills at leading a discussion, which, at certain pivotal times during this process, helped keep things relevant and head off impertinent questions and suggestions. Partha has been very patient and generous with his time, always willing and able to sit down at length to help me work through highly technical, mathematical issues. He never hesitated to point out oddities and inconsistencies in my notation, which has helped engender a deeper appreciation for precision in that regard, something I will definitely carry forward on my continued journey as a researcher.

I would also like to thank the other three committee members, Richard Valliant, James Wagner, and Michael Rendall. I thank Richard for suggesting that additional theory would provide needed insight into how a nonresponse-adjusted point estimate can change over the course of a data collection period, a suggestion prompting development of much of the material in Chapter 2. I thank James for productive discussions during my visits to Ann Arbor and his careful review of an earlier draft. And I thank Michael for taking time away from his multiple, demanding roles with the university to serve as the Dean's Representative.

Lastly, and most importantly, I acknowledge my wife, Katie, for being incredibly selfless during this process. Without (too much) complaint, she has tolerated my extended absences from home in the form of innumerable weekend hours spent at the office and nearby coffee shops. She almost exclusively planned and chronicled our numerous expeditions, many abroad and often tactfully scheduled as an incentive for reaching the major milestones of this endeavor. This kept me motivated and focused on the next milestone. (A trip to Puerto Rico, our first as a family, is the award awaiting a successful dissertation defense!) I also genuinely benefitted from her delicious, health-conscious cooking. When life gets busy, it is tempting to opt for quick and convenient food alternatives, alternatives all too frequently chock full of unpronounceable ingredients and devoid of proper nutrition. Katie keeps tabs on me to ensure I am eating well, and I consider myself extremely lucky to be fed regularly by my favorite cook in the entire world.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

Few surveys are immune to unit nonresponse, which occurs when sampled individuals fail to respond to a survey request. Indeed, response rates have been declining in both the United States and abroad (Atrostic et al., 2001; de Leeuw and de Heer, 2002; Curtin et al., 2005). Groves (2003) argues the domestic trend is a confluence of the rise in single-person households, access impediments such as caller ID and gated communities, and a general increase in reluctance to participate in surveys. This, in turn, has led to rising costs, as increased effort must be expended merely to maintain a survey's historical response rate mark (Curtin et al., 2000). For instance, Groves (2003) reports that the number of interviewer hours required to secure an interview increased some 30 – 40% during the late 1990s for the General Social Survey, the National Comorbidity Study, and the National Survey of Family Growth. While these trends are alarming, there is much evidence refuting the tacit assumption that a higher nonresponse rate is systematically linked to less accurate estimates (Merkle and Edelman, 2002; Groves and Peytcheva, 2008).

The typical protocol for data collection in surveys involves making a sequence of follow-ups on those who have yet to respond, which can take on various forms depending on the survey's mode—reminder mailings, additional telephone calls, or revisits to a residence, to name a few. Each follow-up attempt tends to prompt more survey completes, which we can conceptualize as incoming *waves* of data. On the surface, more follow-ups are desirable, as they serve to reduce the nonresponse rate,

but they come at a cost and extend the data collection field period, delaying subsequent stages of the survey process, such as the reporting and analysis stages. And from a purely practical standpoint, empirical evidence (e.g., Table 1 in Potthoff et al.,1993) suggests returns diminish with each subsequent wave; that is, fewer and fewer completes are attained, impinging smaller and smaller changes upon key estimates.

Descriptive statistics about the nonrespondent follow-up campaign can be subsumed under the concept of *paradata*, a term coined by Couper (1998) to denote process data generated as a byproduct of data collection. Paradata analyses have burgeoned since that time (Kreuter and Casas-Cordero, 2010; Kreuter, 2013). The number of follow-up attempts is one example paradata measure summarizing the level of effort expended to achieve a response. Given the count is known for the entire sample, researchers have evaluated its ability to adjust for nonresponse. Potthoff et al. (1993) reweighted survey data in a telephone survey based on an assumed relationship between the number of callbacks and an outcome variable. Rao, Glickman, and Glynn (2004) evaluated the effect of incorporating the number of follow-up attempts as a continuous predictor variable in an imputation model. Like any candidate variable, its utility hinges on a strong relationship with both the probability of responding *and* the key survey outcome variables (Little and Vartivarian, 2005).

A related class of research has focused on comparing and contrasting the response distributions and associated covariate compositions across some distinction of "early" versus "late" wave respondents (Curtin et al., 2000; Keeter et al., 2006; Billiet et al., 2007; Peytchev et al., 2009; Sigman et al., 2012). In some instances, the objective is to evaluate whether estimates derived from early respondents differ notably from estimates derived using the ultimate set of respondents, early *and* late. A natural feature of these types of these studies is that they tend to measure relative bias, not absolute bias. Estimates using all respondents may not differ much from estimates using only the early wave respondents, but the former is still subject to bias. In other instances, the objective is to assess whether late respondents can proxy for ultimate nonrespondents in some form of nonresponse adjustment. Sometimes the hypothesized relationship holds (Bates and Creighton, 2000), but the technique can backfire when the mechanisms of noncontact differ from nonresponse (Lin and Schaeffer, 1995).

To mitigate the increased costs associated with efforts to stem further declines in response rates, Groves and Heeringa (2006) argue for researchers to employ principles of *responsive survey design*, in which paradata is utilized in real-time to inform data collection decisions and, if necessary, change course. They define a *design phase* to be a spell of data collection with a stable frame, sample, and recruitment protocol and *phase capacity* as the point during a design phase at which the additional responses cease influencing key statistics. The idea is that instead of terminating data collection or transitioning to a new design phase at some arbitrary

threshold, such as a target response rate, one should monitor the accumulating data and stop when phase capacity has been reached. As Wagner and Raghunathan (2010) point out, however, Groves and Heeringa (2006) offer no specific, calculable rule to test for phase capacity. The concept is only illustrated visually in Figure 2 of their paper, in which they plot the trend of a key, nonresponse-adjusted estimate over the data collection period and comment on how the estimate stabilizes well before the design phase concludes. The general aim of this dissertation is to fill this research gap by developing and evaluating a series of methods to formally test for phase capacity.

As an aside, it should be acknowledged that the survey methodology literature abounds with strategies and considerations for allocating resources when following up with nonrespondents (Hansen and Hurwitz, 1946; Filion, 1976; Deming, 1953; El-Bawdry, 1956; Elliott et al., 2000). These typically involve targeting a subset(s) of the remaining nonrespondents with the goal of maximizing precision, minimizing costs, and/or minimizing nonresponse error. One can think of the strategies discussed herein as a way to determine whether it is time to intervene with one of those alternative strategies (i.e., change design phases). Again, the fundamental goal of testing for phase capacity is to detect estimate stability within a fixed data collection protocol. This is not to say the nonresponse-adjusted estimate is free of nonresponse error; we are saying that its immobility following the most recent wave(s) is evidence that its magnitude is no longer changing. Once phase capacity has been reached, it

seems likely future follow-up attempts will be equally inefficacious, and therefore inefficient.

Another critical point worth emphasizing is that the phase capacity tests previously appearing in the literature are often referred to as "stopping rules". This label carries with it the connotation that the nonrespondent follow-up campaign should be discontinued altogether once phase capacity has been reached. This is not precisely the case. As stated previously, phase capacity marks the point at which a new design phase is warranted. Stopping the nonrespondent follow-up campaign is one form of a design phase change, the one exclusively considered in this dissertation, but alternative interventions include switching modes (de Leeuw, 2005) or increasing the incentive offered to the remaining nonrespondents (McPhee and Hastedt, 2012).

1.2 Illustrating Phase Capacity in the Federal Employee Viewpoint Survey

To further elucidate the concept of phase capacity and introduce the real-world survey data set on which the proposed tests will be evaluated, we next discuss the Federal Employee Viewpoint Survey (FEVS). The FEVS, formerly known as the Federal Human Capital Survey (FHCS), was first launched in 2002 by the U.S. Office of Personnel Management (OPM). Initially administered biennially, the Web-based survey is now conducted yearly on a sample of full- or part-time, permanently employed civilian personnel of the U.S. federal government. The core survey instrument consists of 84 work environment questions followed by 14 demographic

questions. Most questions are attitudinal, capturing answers in the form of a five-point Likert scale ranging from Very Satisfied to Very Dissatisfied. Tests of statistical significance are typically performed after collapsing these categories into the dichotomy of a positive/non-positive response. Responses for which a "Do Not Know" or "No Basis to Judge" option is provided are treated as if the positive/non-positive indicator was missing. The key estimate from each item thus reduces to the proportion (or percentage) of employees who react positively to the statement posed. The typical terminology used to describe this statistic is the "percent positive" for a particular survey item. Although this dichotomization ostensibly foregoes some information, Jacoby and Matell (1971) argue that it does not cause any significant decrement in reliability or validity.

Of the myriad uses of the survey's data, one highly visible application is various "Best Places to Work" rankings. OPM publishes a series of rankings as do a few other entities keenly interested in the data. The underlying ranking calculations are not uniform, but all involve grouping thematically-linked subsets of the 84 attitudinal items and amalgamating the percent positive estimates of the items therein. For instance, the OPM formula is the simple average of the percent positive estimates. Example themes are job satisfaction and talent management. Table 1.1 below lists the seven FEVS items comprising OPM's Job Satisfaction index.

Table 1.1: Federal Employee Viewpoint Survey Items Comprising the U.S. Office of Personnel Management's Job Satisfaction Index.

Item	Wording
4	My work gives me a feeling of personal accomplishment.
5	I like the kind of work I do.
13	The work I do is important.
63	How satisfied are you with your involvement in decisions that affect your work?
67	How satisfied are you with your opportunity to get a better job in your organization?
69	Considering everything, how satisfied are you with your job?
70	Considering everything, how satisfied are you with your pay?

The sample frame for the FEVS is derived from a personnel database maintained by OPM. In FEVS 2011, a total of 560,084 individuals from 83 agencies were sampled as part of a single-stage stratified design, where strata were defined by the cross-classification of agency-subelement and one of three supervisory categories: non-supervisors, supervisors, and executives. Agency-subelement is the first organizational component below the agency level. For instance, whereas the U.S. Department of Homeland Security is considered an agency, two of its agency-subelements are the Transportation Security Administration and the U.S. Secret Service. The stratification scheme ensures adequate numbers of supervisors and executives appear in the sample, as they constitute a domain of analytic interest. Base weights equaling the reciprocal of an employee's selection probability are assigned to all sampled individuals to account for the variable sampling rates across strata.

The overall FEVS 2011 field period ran from March 29 to June 1, but the 83 participating agencies had staggered survey start and close dates. The agencies' field

period lengths varied to some degree, but the median duration was six weeks. The data collection protocol fits well into the paradigm of a stable recruitment process with multiple waves of nonrespondent follow-up. On the survey start date, an initial email invitation containing the website URL and log-in credentials was sent to sampled employees. Upon completing the survey, each employee's unique identification number and response vector were time stamped and appended real-time to a database stored on the site's server. Weekly reminders were sent to nonrespondents. Hence, one straightforward demarcation of a data collection wave is the set of responses obtained between any two weekly email invitations. Table 1.2 shows the wave-specific respondent counts and corresponding relative percent increase for three example agencies that will be analyzed throughout this dissertation. It is plain to see how the relative increases quickly diminish after the first few waves. At the conclusion of the last respective wave undertaken, these three particular agencies had achieved roughly 50% response rates, very near the governmentwide average.

Table 1.2: FEVS 2011 Achieved Responses by Data Collection Wave (a Calendar Week) for Three Example Agencies Analyzed in this Dissertation.

	Agency 1		**Agency 2**		**Agency 3**	
Wave	Respondents	Percent Increase	Respondents	Percent Increase	Respondents	Percent Increase
1	2,175	--	240	--	2,178	--
2	1,568	72.1%	139	36.7%	1,516	69.6%
3	1,117	29.8%	49	11.4%	1,304	35.3%
4	865	17.8%	39	8.4%	959	19.2%
5	557	9.7%	31	6.2%	613	10.3%
6	594	9.5%	30	5.7%	510	7.8%
7	532	7.7%	22	4.0%	439	6.2%
8	592	8.0%	22	3.8%	381	5.1%
9	105	1.3%	--	--	408	5.2%
10	--	--	--	--	379	4.6%
	8,105		572		8,687	

The FEVS sample frame contains a plethora of auxiliary variables known for both respondents and nonrespondents, a subset of which is utilized in a three-step weighting process to compensate for unit nonresponse (Kalton and Flores-Cervantes, 2003). In the first step, base weights are computed as the inverse of each sampled individual's selection probability. In the second step, base weights of nonrespondents are proportionally allocated to respondents within classes formed by the cross-classification of agency and demographics such as minority status, gender, tenure with the federal government, and full- or part-time work status. In the last step, weights are raked such that they aggregate to certain known frame totals for the agency as a whole.

The survey reminder schedule is generally fixed for each agency prior to the start of the survey, yet it can be argued that phase capacity occurs before the final reminder email is sent. Since data is electronically recorded real-time and all weighting adjustments can be made after merging this response indicator back onto the sample frame, a series of nonresponse-adjusted point estimates can be charted across time as additional waves of data are incorporated.

Figure 1.1 illustrates this type of plot for an example agency based on item 4, which asks employees their level of agreement with the statement "My work gives me a feeling of personal accomplishment." One can observe how the estimate increases over the course of data collection, even after adjusting for unit nonresponse. By about wave 6, however, the estimate has more or less stabilized. Consequently, this is a pattern observed for many FEVS items: estimates derived from earlier respondents tend to be lower than estimates generated using the ultimate set of respondents (Sigman et al., 2012).

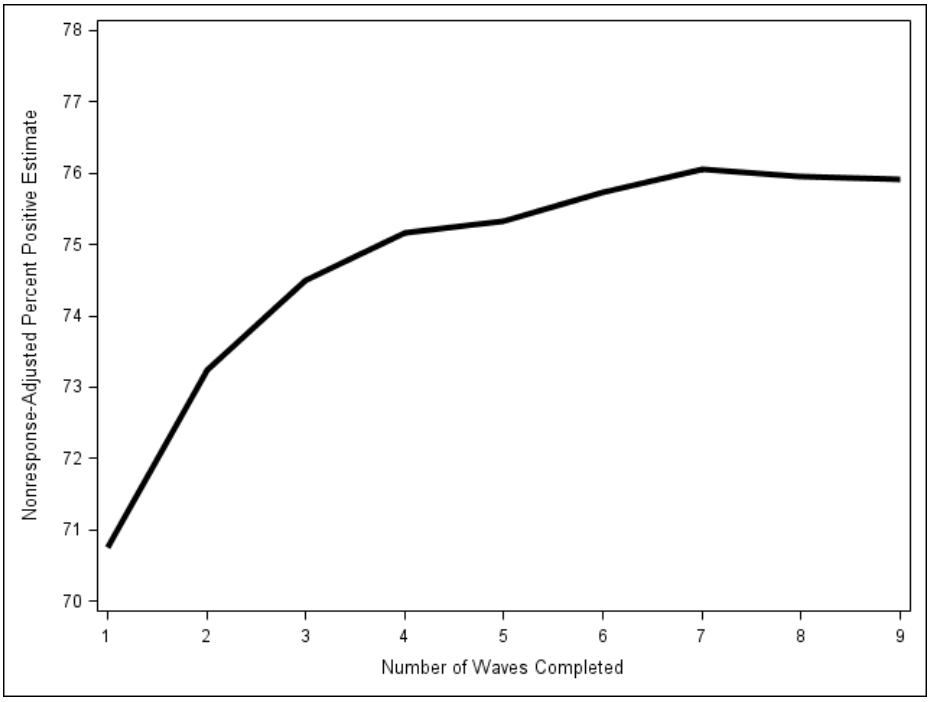

Figure 1.1: Plot of the Nonresponse-Adjusted Percent Positive Statistic for FEVS Item 4 Using Cumulative Data as of the Given Wave of Nonrespondent Follow-Up.

In general, the tendency for nonresponse-adjusted estimates to bounce around more in the earlier waves than latter waves is not unique to FEVS (*cf*. Figure 3 in Wagner (2010) and Figure 3 in Peytchev et al. (2009)). The hope is that a test for phase capacity detects estimate stability at the earliest possible point of stability. Before delving into the specifics of these proposed methods, some background is given in the next section regarding the traditional perspectives of nonresponse and the fundamental assumptions behind techniques to compensate for it. Chapter 2 posits modifications to these perspectives by factoring in a temporal dimension to the response process. Specifically, it provides a framework for phase capacity considerations by extending certain familiar nonresponse-related formulas from the

literature to account not only for the dichotomy of a response or nonresponse, but the polytomy of responding during a specific wave or not responding at all.

1.3 Traditional Nonresponse Perspectives and Terminology

The typical survey's data collection campaign commences by selecting a random sample of size n from a sample frame constructed to represent all N units in a finite population U. It has long been known from survey sampling theory that a randomly selected sample, even of moderate size, can be used to form unbiased (or approximately unbiased) estimates of the attributes of the target population. Specifically, Horvitz and Thompson (1952) proved that, so long as each unit is assigned a fixed, non-zero probability of selection, which we can denote π_i, unbiased estimation can be achieved by assigning each sampled unit a weight that is the inverse of this probability, or $w_i = 1 / \pi_i$. This weight has many names, including the *base weight*, *sampling weight*, or *design weight*, and can be interpreted as the number of population units represented by the sampled unit. The conundrum introduced by nonresponse is that, because only a portion of the sample is observed, the unbiasedness properties demonstrated in Horvitz and Thompson (1952) are no longer guaranteed to hold. Analyzing only the observed portion without making any statistical adjustments may introduce *nonresponse error* (Groves, 1989), or a deviation from the quantity that would be computed from the full sample.

As discussed in Chapter 1 of Groves and Couper (1998), the magnitude of nonresponse error in the sample set S depends on both the statistic at hand and the

degree of dissimilarity between S_1, the set of r observed cases and S_0, the set of m missing cases ($r + m = n$ and $S_1 \cup S_0 = S$). For example, suppose that the quantity of interest is a finite population total $Y = \sum_{i \in U} y_i$ of a particular variable taking on strictly positive values. An unbiased estimate of this quantity could be obtained from the sample by $\hat{Y}_n = \sum_{i \in S} w_i y_i$. The estimator utilizing only the observed portion of the sample, $\hat{Y}_r = \sum_{i \in S_1} w_i y_i$, is certain to underestimate Y since $\hat{Y}_r < \left(\hat{Y}_n = \hat{Y}_r + \hat{Y}_m\right)$, where $\hat{Y}_m = \sum_{i \in S_0} w_i y_i$ represents the base-weighted total of the m missing cases. In contrast, suppose that the quantity of interest is a finite population mean $\bar{y} = \frac{1}{N} \sum_{i \in U} y_i$, for which an approximately unbiased estimate from the full sample can be computed by finding $\hat{\bar{y}}_n = \frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i}$. In the presence of nonresponse, if we let $\hat{\bar{y}}_r = \frac{\sum_{i \in S_1} w_i y_i}{\sum_{i \in S_1} w_i}$ denote the base-weighted mean of the r observed cases and $\hat{\bar{y}}_m = \frac{\sum_{i \in S_0} w_i y_i}{\sum_{i \in S_0} w_i}$ the like for the m missing cases, the nonresponse error is

$$NRerror(\hat{\bar{y}}_r) = \hat{\bar{y}}_r - \hat{\bar{y}}_n$$

$$= \hat{\bar{y}}_r - \frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i}$$

$$= \hat{\bar{y}}_r - \frac{\sum_{i \in S_1} w_i y_i + \sum_{i \in S_0} w_i y_i}{\sum_{i \in S} w_i}$$

$$= \left(1 - \frac{\sum_{i \in S_1} w_i}{\sum_{i \in S} w_i}\right) \hat{\bar{y}}_r - \left(\frac{\sum_{i \in S_0} w_i}{\sum_{i \in S} w_i}\right) \hat{\bar{y}}_m$$

$$= \left(\frac{\sum_{i \in S_0} w_i}{\sum_{i \in S} w_i}\right) (\hat{\bar{y}}_r - \hat{\bar{y}}_m) \qquad (1.1)$$

In words, nonresponse error is the product of the base-weighted nonresponse rate and the difference in base-weighted means between the observed and missing cases. In contrast to the negative nonresponse error for $\hat{Y}_r$ when $y_i > 0$ for all $i \in U$, the quantity in equation 1.1 can be either positive or negative. Specifically, if $\hat{\bar{y}}_r > \hat{\bar{y}}_m$, the quantity is positive, but if $\hat{\bar{y}}_r < \hat{\bar{y}}_m$, the quantity is negative. Another important takeaway is that a larger portion of missing data does not necessarily increase the magnitude of nonresponse error, a point that has been demonstrated empirically in the survey methodology literature (Merkle and Edelman, 2002; Groves and Peytcheva,

2008). The basic notion is that if $\hat{\bar{y}}_r \approx \hat{\bar{y}}_m$, a base-weighted nonresponse rate of 80% is no more detrimental than a rate of 20%.

Figure 1.2 is an analogy provided to help visualize the fundamental concept of nonresponse error for a sample mean. Imagine the outer rectangle represents a three-dimensional water tank (a cube) of which we have a two-dimensional view, and that this tank has been partitioned by a separator running perpendicularly to the bottom of the tank, rendering two subdivisions of water. The water level of the left-hand subdivision represents the base-weighted respondent mean, while the water level of the right-hand subdivision represents the like for nonrespondents. Nonresponse error is the distance between water level of the left-hand subdivision and the resting water level that would be observed if the partition were removed and the two subdivisions were permitted to commingle. This resting water level is represented by the horizontal dashed line in Figure 1.2. The relative portion of the tank's length to the left of the separator represents the base-weighted response rate. Regardless of where it falls, if the "water levels" of both the left- and right-hand side are similar, nonresponse error will be minimal, at least with respect to the sample mean.

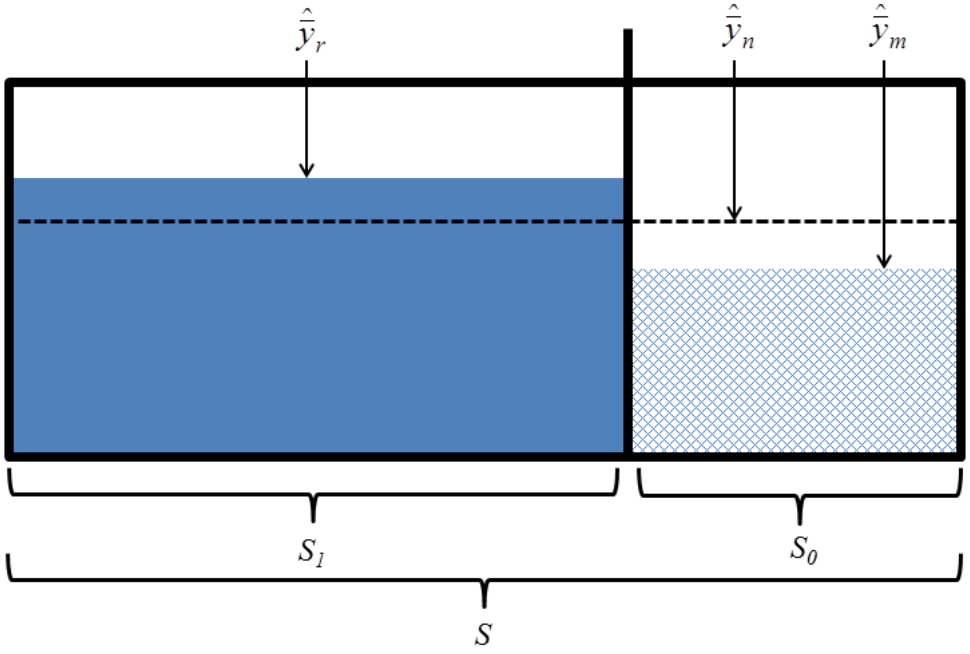

Figure 1.2: Visualization of Nonresponse Error for a Sample Mean Using the Analogy of a Partitioned Water Tank.

Nonresponse error can be partitioned further to account for two or more causes of nonresponse. For example, a common differentiation of nonresponse is the portion attributable to noncontact versus refusal to participate given contact (e.g., Lynn et al., 2002). Let us suppose that the set of m nonrespondents in S_0 is comprised of S_{0A}, the set of m_{nc} units unable to be located, and S_{0B}, the set of m_{ref} units who were located but refused to participate in the survey ($r + m_{nc} + m_{ref} = n$ and $S_1 \cup S_{0A} \cup S_{0B} = S$). If we let $\hat{\bar{y}}_{nc}$ denote the base-weighted mean of the m_{nc} units and $\hat{\bar{y}}_{ref}$ denote the base-weighted mean of the m_{ref} units, starting with the result in equation 1.1, we have

$$NRerror(\hat{\bar{y}}_r) = \left(\frac{\sum_{i \in S_0} w_i}{\sum_{i \in S} w_i} \right) (\hat{\bar{y}}_r - \hat{\bar{y}}_m)$$

$$= \left(\frac{\sum_{i \in S_0} w_i}{\sum_{i \in S} w_i} \right) \left\{ \hat{\bar{y}}_r - \frac{\sum_{i \in S_{0A}} w_i y_i + \sum_{i \in S_{0B}} w_i y_i}{\sum_{i \in S_0} w_i} \right\}$$

$$= \left(\frac{\sum_{i \in S_0} w_i}{\sum_{i \in S} w_i} \right) \left\{ \hat{\bar{y}}_r - \frac{\left(\sum_{i \in S_{0A}} w_i \right) \hat{\bar{y}}_{nc} + \left(\sum_{i \in S_{0B}} w_i \right) \hat{\bar{y}}_{ref}}{\sum_{i \in S_0} w_i} \right\}$$

$$= \left(\frac{\sum_{i \in S_0} w_i}{\sum_{i \in S} w_i} \right) \left\{ \left(\frac{\sum_{i \in S_{0A}} w_i}{\sum_{i \in S_0} w_i} \right) \left(\hat{\bar{y}}_r - \hat{\bar{y}}_{nc} \right) + \left(\frac{\sum_{i \in S_{0B}} w_i}{\sum_{i \in S_0} w_i} \right) \left(\hat{\bar{y}}_r - \hat{\bar{y}}_{ref} \right) \right\}$$

$$= \frac{\sum_{i \in S_{0A}} w_i}{\sum_{i \in S} w_i} (\hat{\bar{y}}_r - \hat{\bar{y}}_{nc}) + \frac{\sum_{i \in S_{0B}} w_i}{\sum_{i \in S} w_i} (\hat{\bar{y}}_r - \hat{\bar{y}}_{ref}) \qquad (1.2)$$

Further decompositions of *m* are possible, but the augmentation of the nonresponse error formula abides by the same basic pattern: a new term is added representing the product of the respective base-weighted prevalence in *S* and the distance between this group's base-weighted sample mean relative to the base-weighted sample mean of the *r* responding cases.

Another important classification of nonresponse is the distinction between *unit nonresponse*, referring to situations in which the sampled unit fails to respond to

the survey request (i.e., answers no questions), and *item nonresponse*, referring to situations in which some, but not all, survey items are answered. These two situations are contrasted in Figure 1.3 for a hypothetical survey with four outcome variables. Ideally, a set of auxiliary variables **X** are known for both observed and missing cases and can be utilized in statistical adjustments to eliminate any error attributable to nonresponse.

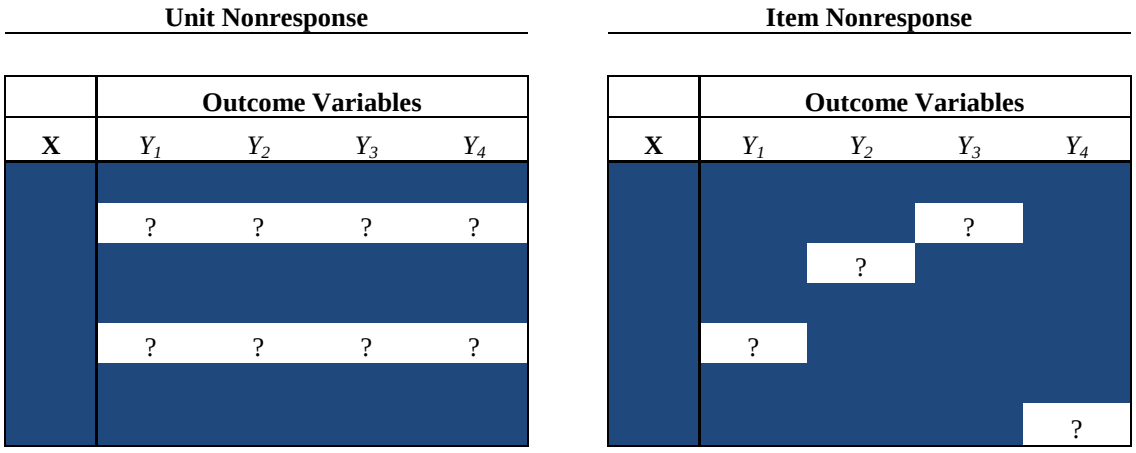

Figure 1.3: Illustration of Unit Nonresponse vs. Item Nonresponse.

The typical remedy for unit nonresponse is to conduct weighting adjustments (Kalton and Flores-Cervantes, 2003) that transfer the base weights of missing cases to the observed cases such that the newly calculated weights (of only the observed cases) better reflect the original sample or population. On the other hand, the typical remedy for item nonresponse is to exploit the relationship between **X** and the vector of outcome variables to form a model which is then used to impute, or fill in, plausible values of the outcome variables for missing cases (Brick and Kalton, 1996).

These are termed "typical" remedies because there can be some overlap. For example, imputation can be employed to combat unit nonresponse. Weighting adjustments are less commonly used to compensate for item nonresponse, but they are feasible. The cumbersome practicality is that separate sets of weights may be needed for separate analyses, particularly in the face of an arbitrary nonresponse pattern such as the one depicted by the right-hand image in Figure 1.3.

The appropriateness of any particular nonresponse-adjustment method depends on the underlying assumption of what Little and Rubin (2002) term the *missingness mechanism*. The three fundamental mechanisms they delineate are governed by the distribution of the sampled units' *propensity* to respond to the given survey request. The terminology and application are most often credited to ideas in Rosenbaum and Rubin (1983), although it can be argued that the concept traces back as far as Hartley (1946) and Politz and Simmons (1949). Denoted ϕ_i, the response propensity is defined as the probability of data being observed (or 1 minus the probability of being missing). The first assumption is that data are *missing completely at random* (MCAR), which means that the propensities are independent of both the auxiliary variables, **X**, and the outcome variable, *y*. If we let R_i denote the response indicator for the i^{th} sample unit, meaning $R_i = 1$ if the unit responds and $R_i = 0$ otherwise, this is to say $\Pr(R_i = 1 \mid \mathbf{X_i}, y_i) = \Pr(R_i = 1) = \phi_i = \phi$ for all i. This is a strong assumption, essentially positing that the observed cases are a completely random subset of the cases originally sampled. The second assumption is that data are *missing at random* (MAR), which means the propensities may vary based on the

auxiliary variables, but not on the outcome variables. Mathematically, this means $\Pr(R_i = 1 | \mathbf{X_i}, y_i) = \Pr(R_i = 1 | \mathbf{X_i}) = \phi_i$. This is the assumption implied for many of the weighting and imputation techniques utilized in practice. Conditional on a common vector of auxiliary variables $\mathbf{X_i}$, data are assumed MCAR. The first two assumptions are sometimes collectively referred to as *ignorable missingness mechanisms*. The third assumption is the most perilous, data that are *not missing at random* (NMAR), implying that the propensities depend on the outcome variable beyond what can be explained by the auxiliary variables, or that $\Pr(R_i = 1 | \mathbf{X_i}, y_i) \neq \Pr(R_i = 1 | \mathbf{X_i})$. In contrast to the first two, this is referred to as a *non-ignorable missingness mechanism*.

Given a fixed and known (but not necessarily equal) propensity of responding for all units in the population, Bethlehem (1988) showed that, over repeated samples of the same size from a population of *N* units, the *nonresponse bias* utilizing $\hat{\bar{y}}_r$, the base-weighted estimate of the sample mean for only the observed portion of the data, is approximately equal to

$$NRbias(\hat{\bar{y}}_r) \approx \frac{1}{N\bar{\phi}} \sum_{i=1}^{N} (\phi_i - \bar{\phi})(y_i - \bar{y}) \qquad (1.3)$$

where $\bar{\phi} = \frac{1}{N} \sum_{i=1}^{N} \phi_i$ symbolizes the average response propensity for all *N* units. That is, the bias is proportional to the population covariance of the propensities and the survey variable. Brick and Jones (2008) derive bias expressions similarly in spirit for other estimators.

It will prove useful at this point to formally derive the Bethlehem (1988) result reported in equation 1.3 because certain intermediate results will be referenced later as part of the theoretical developments presented in Chapter 2. Let $I_i = 1$ if the i^{th} unit from universe U is selected into the sample set S and 0 otherwise, and let $R_i = 1$ if the i^{th} unit is responds to the survey given $I_i = 1$ and $R_i = 0$ otherwise. We can think of the nonresponse bias in $\hat{\bar{y}}_r$ as the expected value of the nonresponse error in $\hat{\bar{y}}_r$, where the expectation is over both the sampling mechanism, $E_S\{\bullet\}$, and missingness mechanism, $E_M\{\bullet\}$. Supposing a sample size large enough such that for two generic survey estimates $\hat{\theta}_1$ and $\hat{\theta}_2$, $E\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right) \approx \frac{E(\hat{\theta}_1)}{E(\hat{\theta}_2)}$ for both the sampling and missingness mechanisms,

$$NRbias(\hat{\bar{y}}_r) = E_S E_M\left\{NRerror(\hat{\bar{y}}_r)\right\}$$

$$= E_S E_M\left\{\left(\frac{\sum_{i\in S_0} w_i}{\sum_{i\in S} w_i}\right)(\hat{\bar{y}}_r - \hat{\bar{y}}_m)\right\}$$

$$= E_S E_M\left\{\left(\frac{\sum_{i\in S_0} w_i}{\sum_{i\in S} w_i}\right)(\hat{\bar{y}}_r - \hat{\bar{y}}_m)\middle| S\right\}$$

$$= E_S E_M\left\{\left(\frac{\sum_{i\in S}(1-R_i)w_i}{\sum_{i\in S} w_i}\right)\left(\frac{\sum_{i\in S} R_i w_i y_i}{\sum_{i\in S} R_i w_i} - \frac{\sum_{i\in S}(1-R_i)w_i y_i}{\sum_{i\in S}(1-R_i)w_i}\right)\middle| S\right\}$$

$$\approx E_S\left\{\left(\frac{\sum_{i\in S}(1-\phi_i)w_i}{\sum_{i\in S}w_i}\right)\left(\frac{\sum_{i\in S}\phi_i w_i y_i}{\sum_{i\in S}\phi_i w_i}-\frac{\sum_{i\in S}(1-\phi_i)w_i y_i}{\sum_{i\in S}(1-\phi_i)w_i}\right)\right\}$$

$$\approx\left\{\frac{N(1-\bar{\phi})}{N}\left(\frac{\sum_{i\in U}\phi_i y_i}{\sum_{i\in U}\phi_i}-\frac{\sum_{i\in U}(1-\phi_i)y_i}{\sum_{i\in U}(1-\phi_i)}\right)\right\}$$

Pausing for a moment, note that the expression $\frac{\sum_{i\in U}\phi_i y_i}{\sum_{i\in U}\phi_i}$ is the approximate expected value of $\hat{\bar{y}}_r$ over the sampling and the nonresponse mechanisms, a result we will use in derivations appearing in Chapter 2.

Continuing on with the derivation of the Bethlehem (1988) formula, the two terms in the right-hand parentheses are factored as follows:

$$=(1-\bar{\phi})\left(\frac{\sum_{i\in U}\phi_i y_i}{N\bar{\phi}}-\frac{Y-\sum_{i\in U}\phi_i y_i}{N(1-\bar{\phi})}\right)$$

$$=\frac{(1-\bar{\phi})}{N}\left(\frac{\sum_{i\in U}\phi_i(1-\bar{\phi})y_i-\bar{\phi}\left(Y-\sum_{i\in U}\phi_i y_i\right)}{\bar{\phi}(1-\bar{\phi})}\right)$$

$$=\frac{1}{N}\left(\frac{\sum_{i\in U}\phi_i y_i-\bar{\phi}\sum_{i\in U}\phi_i y_i-\bar{\phi}Y+\bar{\phi}\sum_{i\in U}\phi_i y_i}{\bar{\phi}}\right)$$

$$= \frac{1}{N}\left(\frac{\sum_{i \in U} \phi_i y_i - \bar{\phi} Y}{\bar{\phi}} \right)$$

$$= \frac{1}{N\bar{\phi}} \sum_{i \in U} (\phi_i - \bar{\phi})(y_i - \bar{y})$$

This expression can be related to the three missingness mechanisms defined by Little and Rubin (2002). The MCAR assumption implies $\phi_i = \bar{\phi}$ for all units in the population, which forces the summation term (and thus the overall bias term) to be 0. The MAR assumption allows the ϕ_i's to vary across $\mathbf{X_i}$, but not within. The objective of nonresponse-adjustment techniques making the MAR assumption is to partition the sample based on $\mathbf{X_i}$, such that within these groupings there is very little variation in the ϕ_i's (i.e., data are MCAR). Finally, the NMAR assumption implies that conditioning on the vector of auxiliary variables does not suitably explain all variation in the ϕ_i's, and that a residual covariance component exists. To see this, consider the alternative expression given on p. 220 of Brick and Kalton (1996). Supposing that the population can be partitioned into *C* classes, they used an analysis of covariance decomposition to re-express the Bethlehem (1988) formula as follows:

$$NRbias(\hat{\bar{y}}_r) = \frac{1}{N\bar{\phi}} \left\{ \sum_{c=1}^{C} \sum_{i=1}^{N_c} (\phi_{ci} - \bar{\phi}_c)(y_{ci} - \bar{y}_c) + \sum_{c=1}^{C} N_c (\bar{\phi}_c - \bar{\phi})(\bar{y}_c - \bar{y}) \right\} \qquad (1.4)$$

where N_c is the number of population units in the class c, and $\bar{\phi}_c$ and $\bar{y}_c$ represent the mean response propensity and outcome variable in class c, respectively. The proof is as follows:

$$\frac{1}{N\bar{\phi}}\sum_{i\in U}(\phi_i-\bar{\phi})(y_i-\bar{y})=\frac{1}{N\bar{\phi}}\sum_{c=1}^{C}\sum_{i=1}^{N_c}(\phi_{ci}-\bar{\phi})(y_{ci}-\bar{y})$$

$$=\frac{1}{N\bar{\phi}}\sum_{c=1}^{C}\sum_{i=1}^{N_c}(\phi_{ci}-\bar{\phi}_c+\bar{\phi}_c-\bar{\phi})(y_{ci}-\bar{y}_c+\bar{y}_c-\bar{y})$$

$$=\frac{1}{N\bar{\phi}}\left\{\sum_{c=1}^{C}\sum_{i=1}^{N_c}(\phi_{ci}-\bar{\phi}_c)(y_{ci}-\bar{y}_c)+\sum_{c=1}^{C}\sum_{i=1}^{N_c}(\phi_{ci}-\bar{\phi}_c)(\bar{y}_c-\bar{y})+\right.$$
$$\left.\sum_{c=1}^{C}\sum_{i=1}^{N_c}(\bar{\phi}_c-\bar{\phi})(y_{ci}-\bar{y}_c)+\sum_{c=1}^{C}\sum_{i=1}^{N_c}(\bar{\phi}_c-\bar{\phi})(\bar{y}_c-\bar{y})\right\}$$

$$=\frac{1}{N\bar{\phi}}\left\{\sum_{c=1}^{C}\sum_{i=1}^{N_c}(\phi_{ci}-\bar{\phi}_c)(y_{ci}-\bar{y}_c)+0+0+\sum_{c=1}^{C}N_C(\bar{\phi}_c-\bar{\phi})(\bar{y}_c-\bar{y})\right\}$$

$$=\frac{1}{N\bar{\phi}}\left\{\sum_{c=1}^{C}\sum_{i=1}^{N_c}(\phi_{ci}-\bar{\phi}_c)(y_{ci}-\bar{y}_c)+\sum_{c=1}^{C}N_C(\bar{\phi}_c-\bar{\phi})(\bar{y}_c-\bar{y})\right\}$$

Consider the popular weighting class adjustment strategy that partitions the sample into C classes and transfers the base weights of nonrespondents to respondents within each. As Brick and Kalton (1996) note, the bias of the weighting class estimator of a sample mean is approximately $\frac{1}{N\bar{\phi}}\left\{\sum_{c=1}^{C}\sum_{i=1}^{N_c}(\phi_{ci}-\bar{\phi}_c)(y_{ci}-\bar{y}_c)\right\}$, which implies that the net effect of this adjustment strategy is to eliminate the second term in equation 1.4. This result lends credence to the recommendation in the literature that the ideally efficacious classification scheme is one substantively differentiating both the probabilities of responding *and* the outcome variable (e.g., p. 63 of Kalton,

1983, Little and Vartivarian, 2005), because if either the $\bar{\phi}_c$'s or the $\bar{y}_c$'s hardly differ amongst the *C* classes, the term eliminated will already be close to zero and the weighting adjustment estimator will hardly differ from the unadjusted sample mean.

The weighting class adjustment strategy assumes data are MAR, where **X** can be thought of as a set of class membership indicator variables. Within a class, it is assumed $\phi_{ci} = \bar{\phi}_c$, or that the propensities are constant. This is generally an untestable assumption given the propensities are rarely known, but there are a variety of methods proposed to obtain $\hat{\phi}_i$'s, or sample-based estimates of the propensities (see Chapter 13 of Valliant et al., 2013). One intuitive approach is to group the sample into *C* classes based on the ordered magnitude of the $\hat{\phi}_i$'s, a technique referred to by Little (1986) as *response propensity stratification*. Typically, *C* is of moderate size (*C* = 5 is common); Eltinge and Yanseneh (1997) and D'Agostino (1998) suggest a few diagnostics for assessing whether the propensity strata structure is suitable. Regardless of the manner in which classes are formed, however, unless the true propensities of cases within a class are approximately equivalent, the missingness mechanism has not been correctly specified and the first summation term in 1.4 may not be zero. For the term to be nonzero, however, there must be a systematic relationship between the deviation of ϕ_{ci} about $\bar{\phi}_c$ and y_{ci} about $\bar{y}_c$.

The notion of response propensities and the bias formula given by equation 1.3 are products of the *stochastic perspective of nonresponse*, which is arguably more realistic and widely adopted than the *deterministic perspective of nonresponse* that

stipulates the sample frame of *N* units consists of *R* units that always respond and *M* units that never respond (Lessler and Kalsbeek, 1992). A nonresponse bias formula with respect to the sampling process can also be derived from the deterministic perspective, however. The proof follows immediately from simply treating the two sets of *R* and *M* units as two domains in the population. Specifically, Valliant et al. (2013) report this quantity (equation 13.1) to be

$$NRbias(\hat{\bar{y}}_r) = \left(\frac{M}{N}\right)(\bar{y}_R - \bar{y}_M) \tag{1.5}$$

where $\bar{y}_R$ represents the population mean of the units that always respond and $\bar{y}_M$ represents the population mean of the units that never respond. Note the resemblance between equations 1.5 and 1.1. Despite sharing a similar structure, the one here is expressed in terms of finite population quantities and the one presented previously in terms of sample-based estimates. Equation 1.1 is an estimate of the quantity in equation 1.5. Interestingly, Groves and Couper (1998, p. 12) assert that the difference in expected respondent mean biases between the two perspectives is minor, even though their expressions look quite different. The difference is more pronounced with respect to the variance of $\hat{\bar{y}}_r$.

1.4 Dissertation Outline

The purpose of this section is to provide a brief overview of the structure of this dissertation. The second chapter details extensions of the concepts and theory

from Section 1.3 to outline a general framework within which the missing data problem inherent to the phase capacity problem can be more directly understood. The subsequent three chapters consist of three distinct methodological studies. Each begins with a brief background section reviewing the literature and framing the problem, and follows with a description of the new method(s) proposed. Each involves a simulation study and application using FEVS 2011 data with the broad objective of comparing and contrasting the properties of the proposed method(s) with their competitors.

Chapter 3 critiques a retrospective phase capacity test recommended by Rao, Glickman, and Glynn (2008) that makes wave-specific adjustments for nonresponse via multiple imputation (Rubin, 1987). The proposed adaptation operates similarly in spirit, but applies to settings in which weighting adjustments are the nonresponse compensation method chosen.

A limitation of the ideas discussed in Chapter 3 is that they are univariate in nature. The tests aim to detect phase capacity with respect to a single estimate (i.e., a sample mean) for a single variable. In practice, however, a typical survey produces a diverse battery of estimates. It is not immediately obvious how to proceed if the test is conducted on multiple estimates with conflicting results. Rather than designating a single estimate as "most important" and basing all decisions thereupon, a multivariate test consolidating several estimates' findings into a single yes-or-no answer would be more useful. Chapter 4 proposes two techniques with that goal in mind.

Acknowledging the retrospective nature of the methods discussed in Rao, Glickman, and Glynn (2008), Wagner and Raghunathan (2010) proposed a prospective test of phase capacity. Given the set of nonrespondents who will respond in the future wave, they derived a formula for the expected variability of the nonresponse-adjusted sample mean of a continuous variable. The notion is that the current nonrespondent follow-up protocol can change once the expected variability is sufficiently small, a quantity that, albeit arbitrary, can be pre-specified by the practitioner. Chapter 5 notes certain limitations of their technique and proposes a more general approach with a broadened applicability.

Chapter 2: Alternative Nonresponse Perspectives to Frame the Phase Capacity Problem

2.1 Introduction

The purpose of this chapter is to discuss modifications to the traditional nonresponse perspectives introduced in the opening chapter that conform more closely to the issues intrinsic to the phase capacity problem. We consider the deterministic perspective in Section 2.2; the stochastic perspective is considered in Section 2.3. The forthcoming theory is provided not only to help frame the phase capacity problem, but also to proffer considerations as to how a sample mean might change over the course of a data collection period. Observing changes in the sample mean is an indication that there are changes to the underlying MCAR, MAR, or NMAR assumption(s). In other words, there is a temporal dimension to the three established missing data classifications. Be advised that the discussions in Sections 2.2 and 2.3 suppose nonresponse adjustments have not been undertaken. Section 2.4, however, briefly touches on considerations for the case of one particular nonresponse adjustment approach, the weighting class adjustment technique.

2.2 An Alternative Paradigm from the Deterministic Perspective

A straightforward extension of the deterministic perspective for a survey collecting data with a constant protocol over K waves is to conceptualize the N population units as falling within one of $K+1$ mutually exclusive and exhaustive domains: K of size $N_1, N_2, \ldots, N_K$ containing units that, if sampled, will always participate in the survey during the k^{th} wave, and a domain of size M containing units

that will never respond. Without loss of generality, let us assume a simple random sample of size n has been drawn from this population. We would anticipate the wave-specific respondent counts $r_1, r_2, \ldots, r_K$ and the count of nonrespondents m ($r_1 + r_2 + \ldots + r_K + m = n$) to fall in proportion to the respective domain's prevalence in the population—that is, $E(r_k) = n^*(N_k/N)$ for $k = 1, \ldots, K$ and $E(m) = n^*(M/N)$. Acknowledging the empirical finding that returns diminish with each subsequent follow-up, we might assume that the N_k's decrease for larger values of k, which, at least in a simple random sample design, would lead us to anticipate that the r_k's will decrease as well on the average. Provided $r_k > 1$ for all K waves, we can express the ultimate respondent sample mean as $\hat{\bar{y}}_r = \sum_{k=1}^{K} \frac{r_k}{r} \hat{\bar{y}}_{r_k}$, where $r = \sum_{k=1}^{K} r_k$ and $\hat{\bar{y}}_{r_k}$ represents the sample mean of the r_k sample units responding during wave k, specifically.

Following the same strategy used to arrive at equation 1.2, we can conceive of $\hat{\bar{y}}_1^k = \frac{\sum_{j=1}^{k} r_j \hat{\bar{y}}_{r_j}}{\sum_{j=1}^{k} r_j}$, the respondent mean using data from waves 1 to k inclusive ($k < K$) (i.e., calculated using data from the $r_1, r_2, \ldots, r_k$ responses thus far obtained) as susceptible to nonresponse error due to the fact that there have been m nonrespondents drawn into the sample with mean $\hat{\bar{y}}_m$ that will never respond and $\sum_{k^*=k+1}^{K} r_{k^*}$ cases that have yet to respond:

$$NRerror(\hat{\bar{y}}_1^k) = \hat{\bar{y}}_1^k - \hat{\bar{y}}_n = \frac{m}{n}(\hat{\bar{y}}_1^k - \hat{\bar{y}}_m) + \sum_{k^*=k+1}^{K} \frac{r_{k^*}}{n}(\hat{\bar{y}}_1^k - \hat{\bar{y}}_{r_{k^*}}) \qquad (2.1)$$

Using the partitioned water tank analogy first introduced in Figure 1.2, Figure 2.1 serves as a visual aid for the story told by equation 2.1, exploiting an example scenario in which $K = 4$ and only the first wave of data has been collected ($k = 1$). As before, imagine the outer rectangle represents a three-dimensional water tank (a cube) of which we have a two-dimensional view, and that this tank has been partitioned by four removable separators. The separators labeled 1, 2, and 3, represent wave thresholds rendering four subdivisions whose widths signify the relative proportions of r_1, r_2, r_3, and r_4. The four wave-specific sample means can be conceptualized as the "water level" of these subdivisions. The rightmost water level represents $\hat{\bar{y}}_m$. At $k = 1$, the nonresponse error is the vertical distance between the water level of the leftmost subdivision, $\hat{\bar{y}}_1^1$, and the horizontal dashed line representing $\hat{\bar{y}}_n$, the full sample mean that would be realized if all four partitions were removed.

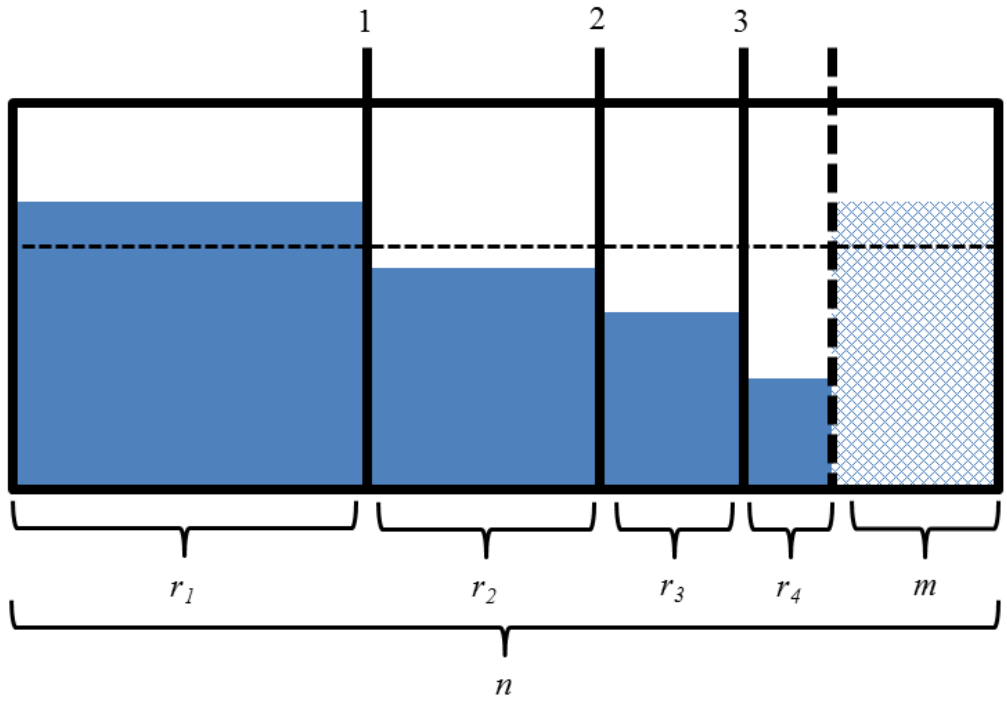

Figure 2.1: Visualization of Nonresponse Error over the Course of a Four-Wave Data Collection Period Using the Analogy of a Partitioned Water Tank.

An interesting facet of this particular nonresponse scenario is that the magnitude of nonresponse error is time-dependent. Specifically, there is minimal nonresponse error after the first wave, yet the magnitude of nonresponse error increases over the subsequent waves. This is because each new wave of data collected actually pulls the observed sample mean further away from $\hat{\bar{y}}_m$. Granted, this is but one example of the distribution of wave-specific respondent counts and sample means, and to further complicate matters, the situation is variable and domain-specific, but it is a simple and effective model for conceptualizing the progression of nonresponse error over the course of data collection.

We can consider $\hat{\bar{y}}_1^1$ an estimate of $\bar{y}_1^1$, the population mean of the domain consisting of N_1 cases, and $\hat{\bar{y}}_1^2$ an estimate of $\bar{y}_1^2$, the population mean of the domain consisting of $N_1 + N_2$ cases, and so on. The objective of the phase capacity test is to use the sample data to assess H_0: $\delta_{k-1}^k = \bar{y}_1^{k-1} - \bar{y}_1^k = 0$ versus H_1: $\delta_{k-1}^k = \bar{y}_1^{k-1} - \bar{y}_1^k \neq 0$, although the hypotheses can be written in terms of other population parameters as well, and non-zero differences for that matter. Note how we can also express the difference as $\delta_{k-1}^k = \left(\bar{y}_1^{k-1} - \bar{y}_n\right) - \left(\bar{y}_1^k - \bar{y}_n\right)$, which illuminates the parallel interpretation that this is an investigation into whether there was no significant change in the expected value of nonresponse error (i.e., nonresponse bias). Whichever the interpretation, if the null hypothesis cannot be rejected, there is evidence that phase capacity has occurred.

Ignoring nonresponse adjustments and focusing on sample-based estimates of this key quantity, an enlightening algebraic manipulation is as follows:

$$\hat{\delta}_{k-1}^k = NRerror(\hat{\bar{y}}_1^{k-1}) - NRerror(\hat{\bar{y}}_1^k)$$

$$= \frac{m}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_m) + \sum_{k^*=k}^{K} \frac{r_{k^*}}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_{r_{k^*}}) - \frac{m}{n}(\hat{\bar{y}}_1^k - \hat{\bar{y}}_m) - \sum_{k^*=k+1}^{K} \frac{r_{k^*}}{n}(\hat{\bar{y}}_1^k - \hat{\bar{y}}_{r_{k^*}})$$

$$= \frac{m}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_m - \hat{\bar{y}}_1^k + \hat{\bar{y}}_m) + \frac{r_{k^*}}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_{r_{k^*}}) + \sum_{k^*=k+1}^{K} \left(\frac{r_{k^*}}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_{r_{k^*}} - \hat{\bar{y}}_1^k + \hat{\bar{y}}_{r_{k^*}}) \right)$$

$$= \frac{m}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^k) + \frac{r_k}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_k) + \sum_{k^*=k+1}^{K} \left(\frac{r_{k^*}}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^k) \right)$$

$$= \left(\frac{m + \sum_{k^*=k+1}^{K} r_{k^*}}{n} \right) (\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^{k}) + \frac{r_k}{n} (\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_k) \qquad (2.2)$$

which shows how the change in nonresponse error is equal to the sum of (1) the product of the portion of sample cases yet to be observed at the conclusion of wave *k* and the observed change in the cumulative sample mean after wave *k* and (2) the product of the portion of sample responding during wave *k*, specifically, and the difference between those respondents' sample mean and the cumulative sample mean as of the previous wave.

Although we will not do so presently, this framework and the formulas given by equations 2.1 and 2.2 could be fleshed out to include terms representing additional causes of nonresponse. Moreover, in the presence of unequal probabilities of selection, base weights could easily be incorporated into the sample mean calculations discussed above and base-weighted versions of the terms such as r_k, m, n and fractions thereof could be introduced.

2.3 An Alternative Paradigm from the Stochastic Perspective

We next discuss amendments to the stochastic perspective of nonresponse to better frame the phase capacity problem. The fundamental change is that the we must broaden the notion of a single response propensity ϕ_i for the i^{th} unit to instead be a *K*-dimensional vector of wave-specific propensities, $\boldsymbol{\varphi}_i = [\phi_{1i}, \phi_{2i}, ..., \phi_{Ki}]$, where each

entry represents the unit's propensity to respond during the k^{th} wave, specifically. That is, we assume that the response process for the i^{th} sample unit follows a multinomial distribution with $K + 1$ events: responding during one of the K waves or not responding at all. Because the events are disjoint, we can treat the event of responding by the conclusion of a particular wave as the sum of the entries in $\boldsymbol{\varphi}_i$ from the first position up through the entry indexing that wave.

As was noted in Section 1.3, a noteworthy preliminary finding of Bethlehem's (1988) derivation is that, given a response propensity ϕ_i, the expectation of the sample mean of the responding units can be shown to equal

$$E(\hat{\bar{y}}_r) = \frac{\sum_{i \in U} \phi_i y_i}{\sum_{i \in U} \phi_i} \tag{2.3}$$

which is a weighted mean (over all units in the population U) in which the response propensity serves as the weight. This holds true regardless of the sampling mechanism, which was shown to cancel out during the derivation.

Using this result, we can reason that the expectation of the sample mean for units responding during the first wave would equal $E(\hat{\bar{y}}_1^1) = \frac{\sum_{i \in U} \phi_{1i} y_i}{\sum_{i \in U} \phi_{1i}}$ and that the expectation of the sample mean for units responding during either the first or second

waves would be $E(\hat{\bar{y}}_1^2) = \frac{\sum_{i \in U} (\phi_{1i} + \phi_{2i}) y_i}{\sum_{i \in U} (\phi_{1i} + \phi_{2i})}$, and so on. Therefore, we can express the

expectation of the difference between two adjacent-wave means as

$$E(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^{k}) = \frac{\sum_{i \in U} (\phi_{1i} + \ldots + \phi_{(k-1)i}) y_i}{\sum_{i \in U} (\phi_{1i} + \ldots + \phi_{(k-1)i})} - \frac{\sum_{i \in U} (\phi_{1i} + \ldots + \phi_{ki}) y_i}{\sum_{i \in U} (\phi_{1i} + \ldots + \phi_{ki})} \tag{2.4}$$

If we define $T_1 = \sum_{i \in U} (\phi_{1i} + \ldots + \phi_{(k-1)i}) y_i$ and $T_2 = \sum_{i \in U} (\phi_{1i} + \ldots + \phi_{(k-1)i})$, this difference can be

re-expressed as

$$E(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^{k}) = \frac{T_1}{T_2} - \frac{T_1 + \sum_{i \in U} \phi_{ki} y_i}{T_2 + \sum_{i \in U} \phi_{ki}} \tag{2.5}$$

The first major takeaway message from equation 2.5 is that the only occasion

in which the difference is exactly zero is when $\frac{\sum_{i \in U} \phi_{ki} y_i}{\sum_{i \in U} \phi_{ki}} = \frac{T_1}{T_2}$, but the difference is

anticipated to be close to zero whenever $\frac{\sum_{i \in U} \phi_{ki} y_i}{\sum_{i \in U} \phi_{ki}} << \frac{T_1}{T_2}$. In other words, barring any

nonresponse adjustments, the change with respect to nonresponse *bias* of the sample

mean will be zero only in the absence of a differential relationship between ϕ_{ki} and y_i relative to what has been observed over the previous wave(s). It seems safe to assume that the sums of wave-specific propensities in the population *U* will tend to decreases in magnitude as *k* increases. As such, we might also anticipate both $\sum_{i \in U} \phi_{ki} y_i$ and $\sum_{i \in U} \phi_{ki}$ to continually decrease, resulting in a progressively smaller change in the sample mean, which would help explain why estimates tend to shift less across later wave thresholds as compared to those in earlier waves. Of course, such a tendency may not always hold, particularly if there is a strong covariance between the ϕ_{ki}'s and y_i's.

2.4 Considerations When Nonresponse Adjustment Methods Are Utilized

The exposition presented thus far in the chapter has focused on the potential for nonresponse error in a sample mean assuming nothing has been done to compensate for it. In practice, weighting adjustments and/or imputation techniques are typically implemented with the aim of reducing this source of error. As can be inferred from Figure 1.1 (and other comparable figures noted from the literature), however, the nonresponse-adjusted sample mean estimates are not necessarily stable over the data collection period. This suggests the missingness mechanism has been misclassified in some way. Continuing with ideas posited with the help of the water tank analogy, let us next consider a few circumstances when this could occur even when a weighting class adjustment strategy is used.

For simplicity, assume one wave of data has been collected for a sample partitioned into $C = 2$ classes, and that an adjustment factor is applied to the wave 1 respondents within each class inflating their base weights such that they sum to N_c, the known population total for class c. The class-specific means at this point are represented by the two water levels in Figure 2.2a. The single separator marks the class threshold, and the relative areas to either side of the separator represent the relative sizes of N_1 and N_2. The full sample mean is represented by the dashed line, which can be interpreted as the resting water level if the lone separator were removed.

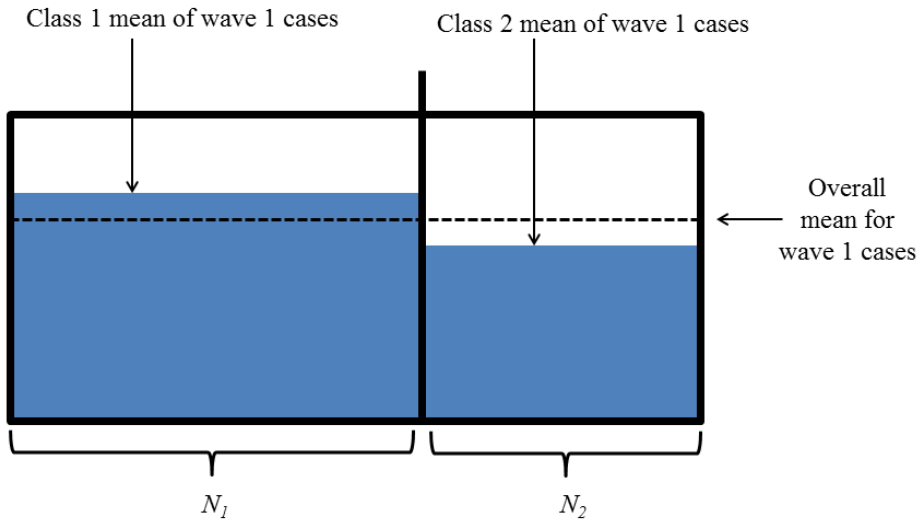

Figure 2.2a: Visualization of a Two-Class Weighting Adjustment Strategy Using the Analogy of a Partitioned Water Tank.

Recall that the assumption behind the weighting class adjustment strategy is that data are MCAR within each class, meaning the expected value (or water level, in this paradigm) of the given outcome variable for cases within a class should be the

same regardless of which sample units happened to respond in the first wave. To the extent this proves systematically incorrect, nonresponse error can result. Figure 2.2b visualizes the impact of incorporating a second wave of data in the presence of one such example. Relative to Figure 2.2a, the two classes have been additionally partitioned by a threshold representing the two waves of data. Notice how the class-specific means (i.e., water levels) for wave 2 respondents are larger in magnitude than the class-specific means of the wave 1 respondents. Within a class, the change after commingling the two waves' responses follows immediately from the discussion surrounding Figure 1.2, and has the same functional form as that of the deterministic perspective's nonresponse error formula given by equation 1.1. In terms of the notation utilized in Figure 1.2, for example, we can consider the wave 1 respondents as the region denoted by S_1, and the wave 2 respondents as the region denoted by S_0. As such, the class-specific change is simply the product of the weighted portion of wave 2 respondents and the difference between the weighted sample mean of the wave 1 respondents and wave 2 respondents, where the weights utilized are those calculated at the end of the second wave. The net change with respect to the overall sample is the sum of these class-specific changes weighted proportionally by N_c/N.

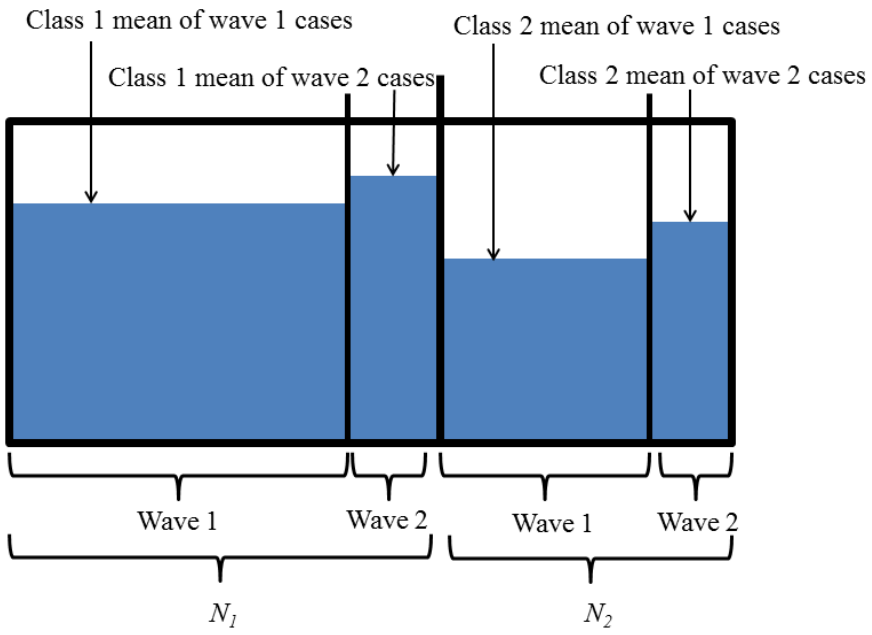

Figure 2.2b: Visualization of Wave-Specific Means for a Two-Class Weighting Adjustment Strategy Using the Analogy of a Partitioned Water Tank.

The scenario depicted by Figure 2.2b is one of a systematically incorrect assumption regarding the missingness mechanism. This is but one example of an infinite number of circumstances. As another example, there may be some classes in which wave 2 respondents' means are larger and others in which the wave 2 respondents' means are smaller. The relative impacts could be negated if the two are roughly proportional to one another. Also, these considerations are domain-specific. For instance, if some aggregation of classes in which the wave 2 respondents' means are larger constituted a domain of analytic interest, that domain mean would still be susceptible to nonresponse error. Another factor is the relative sizes of the N_c's For better or worse, it is entirely possible that what occurs in one class could dominate the overall picture if it comprises an outsized share of population.

Considerations of this simple model apply at later points in the data collection process, which brings into play yet another dimension: the relative size of the pending wave cohort to be introduced to the classes. More disparate wave-specific means have less of an impact when the weighted portion of cases introduced is small in comparison to the weighted portion of respondents from wave(s) already completed.

Regardless of the technique employed to compensate for nonresponse, the spirit of the phase capacity test is as follows: for a general population parameter θ estimated at the conclusions of two adjacent waves by $\hat{\theta}_1^{k-1}$ and $\hat{\theta}_1^{k}$, respectively, if $\hat{\delta}_{k-1}^{k} = \hat{\theta}_1^{k-1} - \hat{\theta}_1^{k}$ is significantly different from 0 (formal tests to assess this are discussed in the forthcoming chapters), the dynamics of the wave-specific nonresponse mechanism have not yet stabilized to a point where the marginal impact on the estimate is inconsequential, and so another wave of data collection is warranted.

Chapter 3: A Retrospective Test for Phase Capacity When Weighting for Nonresponse

3.1 Background

Rao, Glickman, and Glynn (RGG) (2008) was the first known attempt at quantifying estimate stability across waves of nonrespondent follow-up, although their motivation was a concurrently progressing literature on sequential decision rules in clinical trials (O'Quigley et al., 1990), not the concept of phase capacity as discussed in Groves and Heeringa (2006). RGG's research question was to determine when they could stop mailing replacement questionnaires to a sample of women recruited for a large pregnancy prevention study. Covariates collected during the recruitment stage served as the auxiliary variables $\mathbf{X}$ known for the entire sample as these women were followed over time. The estimate they considered was a sample mean, the proportion of women using birth control. Given the completion of wave k ($k \geq 2$), RGG questioned how much inferences would have changed had data collection stopped at wave $k - 1$. To help quantify the uncertainty surrounding that question, they derived three rules.

Rule 1 gauges whether units' response wave is associated with the outcome. Specifically, one uses the respondent data to fit a model relating covariates, wave of response, and interaction between the two to the outcome. One then fits a reduced model omitting the wave-related terms and forms a likelihood-ratio test—or an F for a linear regression when the outcome is continuous—to see if the reduced model holds. If so, phase capacity has been reached.

Rule 2 compares the change in the survey estimate itself by partitioning the respondent set into two mutually exclusive groups, those who responded during waves 1 through $k - 1$ and those who responded during wave k. A two-sample t test is conducted to determine whether the two cohorts yield significantly different mean outcomes. If not, there is evidence phase capacity has occurred. Rules 1 and 2 are intuitive but neither employs the known auxiliary variables to adjust for nonresponse. Moreover, the authors found Rule 2 to be prone to false discoveries in later waves due primarily to the continually decreasing respondent counts. RGG's third rule performed best in simulation and application.

RGG Rule 3 adjusts for nonresponse by multiply imputing (Rubin, 1987) the missing birth control usage indicator variable. In contrast to techniques that reweight respondent records to better reflect the target population, imputation methods attempt to fill in the unobserved values. A survey data set subject to missingness has an outcome vector $\mathbf{Y}$ that can be partitioned into two components $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_0)$, where $\mathbf{Y}_1$ is the observed component and $\mathbf{Y}_0$ the missing component. An imputation model exploits the relationship between $\mathbf{X}$ and $\mathbf{Y}_1$. The model can be either explicit (e.g., linear regression) or implicit (e.g., class-based, such as so-called *hot-deck* imputation). *Multiple imputation* (MI) is a technique whereby missing values are imputed M times ($M \geq 2$), thereby rendering M completed data sets. RGG (2008) use $M = 5$, a fairly common value in practice (e.g., Schenker et al., 2006). Rubin (1987)

advocates this technique over single imputation since an augmentation to the variance formula allows one to better reflect the missing data uncertainty.

Let $\hat{Q}_m$ denote the m^{th} completed data set estimate for any quantity Q. The MI estimate is the arithmetic mean of the *M* completed data set estimates, or $\hat{\bar{Q}}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m$. Let $\hat{U}_m$ denote the m^{th} completed data set estimated variance for $\hat{Q}_m$. The MI variance is the sum of (1) the average of the *M* completed data set variances $\hat{\bar{U}}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{U}_m$ plus (2) the between-imputation variance of the estimate $\hat{B}_M = \left(1+\frac{1}{M}\right)\sum_{m=1}^{M}\frac{\left(\hat{Q}_m - \hat{\bar{Q}}_M\right)^2}{M-1}$. That is, the overall multiple imputation variance formula is $\hat{T}_M = \hat{\bar{U}}_M + \hat{B}_M$. The term $\left(1+\frac{1}{M}\right)$ represents a finite imputation correction factor, which converges to 1 as $M \to \infty$.

RGG Rule 3 proceeds as follows. First, one imputes the current nonrespondents using data available through wave *k*. Then responses obtained during wave *k*, specifically, are deleted and imputation is performed using a model fit using data through wave $k-1$. The result is $2M$ completed data sets. The two sets of multiply-imputed data are obviously dependent, since the underlying models are based on the shared fully observed data through wave $k-1$. To circumvent the calculation of covariances, RGG cleverly construct a sequence of *M* individual-level difference variables, $d_{mi} = y_{mi}^{k-1} - y_{mi}^{k}$, where the superscript denotes the maximum

wave's data used in the imputation model and the subscript denotes the m^{th} completed data set value (imputed or observed) for the i^{th} individual. A contrived data set is presented in the Appendix to provide a visualization of the process. For respondents up to and including wave $k-1$, $d_{mi}=0$, but question marks indicate values subject to variation over repeated implementations of the imputation procedure.

Phase capacity is declared whenever $\hat{\bar{d}}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{d}}_m$ is not significantly different from zero. The quantity $\hat{\bar{d}}_M$ is standardized by dividing through by the square root of its MI variance and referenced against a student *t* distribution with desired level of confidence. The MI variance is defined as the sum of the sample variance of the *M* point estimates of $\hat{\bar{d}}_m$ times the finite imputation correction factor and the average of the *M* values of $\text{var}(\hat{\bar{d}}_m)$. The former is the between-imputation variance component and the latter is the within-imputation variance component. Depending upon the degree of overlap, the overall MI variance computed in this manner should be much smaller than a method assuming independence of the two sets of multiply-imputed data (i.e., ignoring what would certainly be a positive covariance).

3.2 New Methods

One potential downside to RGG's phase capacity test is that, for the imputation process to be truly effective, predictive covariates are needed. Not all surveys have that luxury. For example, there may be little known about unresolved sampled telephone numbers in a random-digit-dialing (RDD) survey. In these and

numerous other settings, respondent records might be reweighted to better represent the target population, perhaps by benchmarking to external control totals obtained from administrative records or a census. The purpose of this section is to introduce a proposed adaptation of the RGG's test amenable to reweighting the observed.

Suppose for the moment that we are still interested in determining whether $\hat{\bar{y}}_1^k$, the sample mean using data from waves 1 through wave *k*, is no different from $\hat{\bar{y}}_1^{k-1}$, the sample mean using data only through wave $k - 1$. Suppose further that the two sample means are weighted by w_1^k and w_1^{k-1}, the nonresponse-adjusted base weights computed to better represent the target population as of the conclusion of the two adjacent waves. For sample units that responded at or before wave $k - 1$, both weights would be positive. For sample units responding specifically during wave *k*, w_{1i}^k would be positive while $w_{1i}^{k-1} = 0$. For sample units that have yet to respond by wave *k*, both w_{1i}^k and w_{1i}^{k-1} would be 0.

As before, the objective is to standardize the difference between the two sample means, which requires an estimated variance of the difference. Fundamentals of Taylor series linearization can be employed after first observing how the difference can be expressed as a function of $p = 4$ estimated totals:

$$\hat{\delta}_{k-1}^k = \hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^k = \frac{\sum_{i=1}^n w_{1i}^{k-1} y_i}{\sum_{i=1}^n w_{1i}^{k-1}} - \frac{\sum_{i=1}^n w_{1i}^k y_i}{\sum_{i=1}^n w_{1i}^k} = \frac{\hat{Y}_1^{k-1}}{\hat{N}_1^{k-1}} - \frac{\hat{Y}_1^k}{\hat{N}_1^k} = \frac{\hat{T}_1}{\hat{T}_2} - \frac{\hat{T}_3}{\hat{T}_4} \qquad (3.1)$$

When written in this fashion, Wolter (2007, Section 6.5) demonstrates how a computational algorithm attributable to Woodruff (1971) can greatly simplify the Taylor series variance approximation process. Similarly to RGG's difference variable approach, the technique's appeal is that it bypasses the need to calculate $\binom{p}{2}$ covariances. The algorithm calls for one to create a primary sampling unit (PSU) level variate u_i equaling the sum of the function's partial derivatives multiplied by the corresponding estimated total. In the present case, $\text{var}(\hat{\delta}_{k-1}^{k}) \approx \text{var}\left(\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{\partial \delta_{k-1}^{k}}{\partial T_j}t_{ji}\right)$, where t_{ji} represents the PSU-level estimate of the j^{th} total in the function. Specifically, $t_{1i} = w_{1i}^{k-1}y_i$, $t_{2i} = w_{1i}^{k-1}$, $t_{2i} = w_{1i}^{k}$, and , $t_{2i} = w_{1i}^{k}$. After a little algebra, it can be shown

$$\sum_{j=1}^{p}\frac{\partial \delta_{k-1}^{k}}{\partial T_j}t_{ji} = u_i = \frac{1}{\hat{N}_1^{k-1}}w_{1i}^{k-1}y_i - \frac{\hat{Y}_1^{k-1}}{\left(\hat{N}_1^{k-1}\right)^2}w_{1i}^{k-1} - \frac{1}{\hat{N}_1^{k}}w_{1i}^{k}y_i + \frac{\hat{Y}_1^{k}}{\left(\hat{N}_1^{k}\right)^2}w_{1i}^{k} \qquad (3.2)$$

and so the estimated variance of the sum of the u_i's with respect to the sample design approximates $\text{var}(\hat{\delta}_{k-1}^{k})$. Table 3.1 provides a visualization of this technique using a simple, hypothetical survey data set where $k = 2$.

Table 3.1: Illustration of the Taylor Series Linearization Method to Approximate the Variance of the Difference of Two Adjacent Waves' Nonresponse-Adjusted Sample Means.

Observed Data					**Linearized Variate***
Sample Case ID	Wave	w_{1i}^1	w_{1i}^2	y_i	u_i
1	1	10.1	4	1.3	-0.0362
2	1	10.2	7	1.1	-0.0284
3	1	9.7	7	2.1	0.0213
4	1	10.6	5.4	1.8	0.0130
5	1	8.8	6.3	1.7	0.0030
6	1	10.6	6.2	2.0	0.0260
7	2	0	6.4	1.4	0.0300
8	2	0	5.7	1.8	-0.0113
9	2	0	5.3	1.6	0.0072
10	2	0	6.7	1.9	-0.0245

*Calculated as $u_i = \frac{1}{\hat{N}_1^1} w_{1i}^1 y_i - \frac{\hat{Y}_1^1}{\left(\hat{N}_1^1\right)^2} w_{1i}^1 - \frac{1}{\hat{N}_1^2} w_{1i}^2 y_i + \frac{\hat{Y}_1^2}{\left(\hat{N}_1^2\right)^2} w_{1i}^2$, where $\hat{N}_1^1 = 60$, and $\hat{Y}_1^1 = 99.96$, $\hat{N}_1^2 = 60$, and $\hat{Y}_1^2 = 100.86$.

Using figures in the table above, we find $\hat{\bar{y}}_1^1 = \frac{\sum_{i=1}^{6} w_{1i}^1 y_i}{\sum_{i=1}^{6} w_{1i}^1} = \frac{\hat{Y}_1^1}{\hat{N}_1^1} = \frac{99.96}{60} = 1.666$,

$\hat{\bar{y}}_1^2 = \frac{\sum_{i=1}^{10} w_{1i}^2 y_i}{\sum_{i=1}^{10} w_{1i}^2} = \frac{\hat{Y}_1^2}{\hat{N}_1^2} = \frac{100.86}{60} = 1.681$, and so $\hat{\delta}_1^2 = -0.015$. The estimate of $\text{var}(\hat{\delta}_1^2)$ is approximated by $\text{var}\left(\sum_{i=1}^{10} u_i\right)$ = 0.00567. The observed *t* statistic is then

$\frac{\hat{\delta}_1^2}{\sqrt{\text{var}(\hat{\delta}_1^2)}} = \frac{-0.015}{0.075302} = -0.199$, which is referenced against a student *t* distribution with *n* – 1 = 9 degrees of freedom to obtain a *p*-value under the two-tailed hypothesis test

$H_0: \delta_1^2 = \bar{y}_1^1 - \bar{y}_1^2 = 0$ versus H_1: $\delta_1^2 = \bar{y}_1^1 - \bar{y}_1^2 \neq 0$. As a general rule, the degrees of freedom would be calculated based on the tally from the wave *k* data set. In this hypothetical setting, it appears the nonresponse-adjusted sample mean did not change significantly between waves 1 and 2, implying phase capacity has occurred.

While the set-up thus far has pertained only to simple random sampling designs, complex survey features can be accommodated. For instance, many survey's sampling procedure involves hierarchical stages of clustering, often within strata. To simplify the variance approximation process, the "ultimate cluster" assumption (see p. 67 of Heeringa et al., 2010) is frequently adopted in which the u_i's are constructed as illustrated above at the PSU level and stratum-specific variances are estimated and summed across all strata. And although the present exposition focused only on the sample mean, the Woodruff (1971) technique is applicable to any difference that can be expressed as a differentiable function of unbiased totals, which covers a wide range of statistics. This is a notable advantage over RGG's version of the rule, whose difference variable approach was designed specifically to test for a difference of two sample means.

As an aside, there is an alternative computational algorithm practitioners may find easier to apply than the method outlined above, at least when the key estimate being monitored is a sample mean. Drawing upon concepts demonstrated in Example 5.13 of Heeringa et al. (2010), the first step is to stack the two fully observed data sets, one as of wave *k* and another as of wave $k - 1$, with a like-named weight variable

and PSU identifier. Note that even under a simple random sample design, one would treat the unique respondent identifier as the PSU (i.e., a cluster variable). The next step is to assign an indicator variable in this stacked data set taking on a value of 0 for cases from the wave k data set and a value of 1 for cases from the wave $k - 1$ data set. One then fits a linear regression model with an intercept and this indicator variable serving as the lone predictor variable on the outcome variable of interest. So long as the variance-covariance matrix of model parameters is estimated properly accounting for the clustering (and stratification, if applicable) (Fuller, 1975), it can be shown that the t statistic generated from the null hypothesis that the slope coefficient in this simple model is zero matches what was calculated above using the u_i's.

Another feasible method for approximating $\text{var}(\hat{\delta}^k_{k-1})$ is to employ a *replication* approach (Rust, 1985), one of a class of alternatives to Taylor series linearization. Replication techniques are particularly handy tools for simplifying variance calculations of estimates derived from complex sample designs. One example is *balanced repeated replication* (BRR) (Ch. 3 of Wolter, 2007), which was developed for the commonly encountered two-PSU-per-stratum design. One creates a series of R replicate weights by doubling the weights for one cluster's observations within a stratum while setting the other cluster's weights to zero. A Hadamard matrix from the field of experimental design is used to ensure *balance* between the number of PSUs maintained or dropped across the replicates. The point estimate's variance is approximated by a straightforward function of the full-sample point estimate and the like calculated using each of the R replicate weights. A nice feature of the technique,

as well as other replication techniques, is that there is generally a single variance formula, regardless of the underlying quantity being estimated. If we let $\hat{\theta}^r$ denote the r^{th} replicate weight estimate ($r = 1, \ldots, R$) for any quantity θ and denote the full-sample point estimate $\hat{\theta}$, the BRR variance is approximated by $\text{var}_{BRR}(\hat{\theta}) = \frac{1}{R}\sum_{R}\left(\hat{\theta}^r - \hat{\theta}\right)^2$.

BRR can be applied to the phase capacity problem by forming a set of R replicate weights for (1) respondents through wave $k-1$ and (2) respondents through wave k. In sum, $2R$ replicate weights are constructed. One then conducts the full nonresponse adjustment routine on all replicate weights independently. After finding both $\hat{\theta}_1^{(k-1)r}$ and $\hat{\theta}_1^{kr}$ using the two sets replicate weights, the $2R$ estimates are consolidated by forming $\hat{\theta}^r = \hat{\theta}_1^{(k-1)r} - \hat{\theta}_1^{kr}$. Ultimately, the average squared deviation of these R estimated differences from the full-sample difference $\hat{\theta} = \hat{\theta}_1^{k-1} - \hat{\theta}_1^{k}$ serves as the approximation of the variance of the two sample means' difference. Applications using other replication approaches, such as the *jackknife* (Ch. 4 of Wolter, 2007) or *bootstrap* (Efron and Tibshirani, 1993), could be conducted in a similar manner.

3.3 Simulation Study

In order to evaluate the performance and of their proposed rules, RGG (2008) conducted a simulation study based on four hypothetical relationships between when a sample unit responds, a continuous covariate (i.e., auxiliary variable), and a dichotomous outcome variable. The four conditions were based on the combination of (1) whether or not the wave of response was associated with the covariate and (2)

whether or not the outcome variable was associated with wave. The covariate was always assumed associated with the outcome; otherwise, the imputation model would have been futile. They evaluated their three rules on 1,000 data sets of size $n = 200$ and $n = 10{,}000$, respectively.

RGG first assigned a random normal deviate to be a covariate x_i known for all sample units. For the condition where wave was not associated with the outcome, $wave_i \sim Poisson(1)$. The other condition was $wave_i \sim Poisson(1)$ if $x_i < 0$ and $wave_i \sim Poisson(5)$ otherwise. Of course, a draw from the Poisson distribution could return 0, so each value was incremented by 1. Next, a 0/1 outcome variable y_i was randomly generated based on an assumed log-odds relationship between the covariate and wave-of-response variable. Table 3.2 summarizes the four conditions.

Table 3.2: Parameters of the Rao, Glickman, and Glynn (2008) Simulation Study.

	Outcome Not Associated with Wave	**Outcome Associated with Wave**
Wave of Response Not Associated with Covariate	$wave_i \sim Poisson(1)$ $P(y_i = 1 \mid x_i) = \frac{e^{1+2x_i}}{1+e^{1+2x_i}}$	$wave_i \sim Poisson(1)$ $P(y_i = 1 \mid x_i) = \frac{e^{1+2x_i+wave_i+2x_i*wave_i}}{1+e^{1+2x_i+wave_i+2x_i*wave_i}}$
Wave of Response Associated with Covariate	$wave_i \sim Poisson(1)$ if $x_i < 0$; $wave_i \sim Poisson(5)$ otherwise $P(y_i = 1 \mid x_i) = \frac{e^{1+2x_i}}{1+e^{1+2x_i}}$	$wave_i \sim Poisson(1)$ if $x_i < 0$; $wave_i \sim Poisson(5)$ otherwise $P(y_i = 1 \mid x_i) = \frac{e^{1+2x_i+wave_i+2x_i*wave_i}}{1+e^{1+2x_i+wave_i+2x_i*wave_i}}$

Because data were available for the full sample, the rules' performance could be evaluated with respect to various quantifications of the impact of stopping early. A tacit assumption is that unit nonresponse could be eliminated entirely given enough follow-up attempts, which is not necessarily realistic, but at least permits a gold standard against which estimates formulated from the abridged sample could be compared. Although the authors termed the discrepancy *bias*, it could perhaps be more appropriately labeled *nonresponse error* following the terminology of Groves (1989).

One reason the authors concluded superiority of Rule 3 was that it suggested phase capacity at (or very near) the second wave, the earliest possible stopping point, for all four conditions and with virtually no error relative to the full-sample estimate. They attributed this to the imputation process recapturing a large portion of the missing information. It could be argued, however, that their conclusion was a byproduct of the simulated relationships between the outcome and wave of response not being strong enough. For example, the authors state that the sample mean for the condition where wave of response is independent of the outcome was 0.65, whereas the sample mean was 0.69 for the condition where wave of response is associated with the outcome. It seems plausible a stronger relationship could have engendered more nonresponse error. Moreover, the bounded nature of y_i restricts the potential imparity of the sample means. Because of these limitations and a few others to be discussed shortly, certain modifications to their simulation study design were made for the present study.

The fundamental goal of the simulation study discussed herein was to foster as balanced a testing ground as possible for the two competing methods to compensate for unit nonresponse. The first step was to randomly generate a covariate dichotomizing the sample into two classes within which both a weighting adjustment and multiple imputation routine can be performed. Effectively, data were assumed MCAR within each class. For the weighting rule, a single adjustment factor proportional to the inverse of the class-specific response rate was used to inflate the weights of respondents to the initial sample total within that class. RGG's imputation-based rule was carried out in the form of the *approximate Bayesian Bootstrap* (ABB). Outlined by Rubin and Schenker (1986), the ABB is the appropriate procedure for multiple imputation in a hot-deck setting.

The two-step ABB proceeds as follows. If, within a class, there are r respondents and m non-respondents, each comprised of data vectors $\mathbf{Y}_1$ and $\mathbf{Y}_0$, respectively, the first step is to select a sample of size r with replacement from $\mathbf{Y}_1$. From this set, one selects m values with replacement and uses those to impute the vector of missing outcome variables, $\mathbf{Y}_0$. The process is repeated independently M times. It is vital to incorporate this extra variability via the two-step, with-replacement sample selection scheme because failing to do so ignores the uncertainty inherent when modeling the missing data mechanism—in Rubin's (1987) terminology, such an imputation procedure would be "improper." Even if one were to simply draw m values from $\mathbf{Y}_1$ independently M times and apply Rubin's

combination rules, the variance would still be underestimated. Interestingly, Rubin and Schenker (1986) prove that the expected value of the variance of a sample mean after implementing the ABB equals the sample mean variance approximated using only the observed portion of the data, $\mathbf{Y}_1$.

Note that, within a class, a constant weight adjustment will have no effect on the variance of a mean. Taken together with the last point of the previous paragraph, the two techniques should be completely balanced in terms of their expected pre-and post-adjustment precision on $\hat{\bar{y}}_1^{k-1}$ and $\hat{\bar{y}}_1^{k}$.

To partition the sample into two classes of roughly equal size, a random uniform variate x_i between 0 and 1 was generated. A sample case was assigned to the first class if this number was less than 0.5, and the second class otherwise. The two wave-of-response conditions were defined similarly in spirit to those specified in RGG (2008), but were operationalized differently. The notion was still to simulate two settings in which the wave of response either was or was not associated with the covariate, but instead of being governed by a Poisson distribution, an empirical FEVS 2011 distribution was utilized. Table 3.3 summarizes the specific wave distributions. For the condition where wave was not associated with x_i, a sample case was assigned a response wave randomly in proportion to the distribution of Agency 3 given in Table 1.2. For the condition where wave was associated with the covariate, if $x_i < 0.5$ the sample case tended to respond sooner than when $x_i \geq 0.5$. These were carefully designed such that the expectation of the marginal distribution matched that of the

alternative condition—for instance, .5*(34.5% + 15.6%) ≈ 25.1% and .5*(20.7% + 14.2%) ≈ 17.5%.

Table 3.3: Summary of the Two Wave-of-Response Distributions used for the Simulation Study Comparing RGG Rule 3 Phase Capacity Test to the Weighting Variant.

	Wave Not Associated with Covariate	**Wave Associated with Covariate**	
Wave	for any x_i	$x_i < 0.5$	$x_i \geq 0.5$
1	25.1%	34.5%	15.6%
2	17.5%	20.7%	14.2%
3	15.0%	11.5%	18.5%
4	11.0%	9.2%	12.9%
5	7.1%	4.6%	9.5%
6	5.9%	4.6%	7.1%
7	5.1%	3.7%	6.4%
8	4.4%	3.5%	5.3%
9	4.7%	3.9%	5.5%
10	4.4%	3.7%	5.0%
	100.0%	100.0%	100.0%

Another substantive change relative to the RGG (2008) design was that the outcome variable y_i was assigned as continuous rather than dichotomous. For the condition where the outcome was not associated with wave of response, $y_i = 10x_i + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$. When the outcome was associated with respondent wave, $y_i = 10x_i + wave_i + \varepsilon_i$. Thus, the wave-specific mean outcome increases linearly in expectation.

As with RGG (2008), the four conditions were simulated 1,000 times, but for sample sizes n = 500 and n = 5,000 instead of n = 200 and n = 10,000. The reason for increasing the lower-end sample size was to minimize the occurrence of a "skipped" wave, when cases in the simulated data set were assigned as responding in wave $k - 1$ and others assigned as responding in wave $k + 1$, but no cases were assigned as responding in wave k. When there are no respondents during wave k, it is impossible to apply the weighting rule as prescribed because w_{1i}^{k-1} and w_{1i}^{k} are identical for all i, which causes $\text{var}(\hat{\delta}_{k-1}^{k}) = 0$ and so the t-test for phase capacity is undefined. This situation is unique to the weighting rule, because the estimated MI variance of the sample mean of the M difference variables in RGG's method will generally be positive, unless there is full response, a perfectly predictive imputation model, or some other extraordinarily unusual situation. Decreasing the larger sample size from n = 10,000 to n = 5,000 was done primarily in the interest of managing simulation run time. Initial evaluations indicated there were hardly any noteworthy differences between the two values of n.

A practical issue when employing multiple imputation is deciding on the size of M. A common value used by many researchers (e.g., Schenker et al., 2006), including RGG (2008), is M = 5. Graham et al. (2007) argue that this number may be insufficient in certain circumstances. During preliminary analyses, M = 20 and M = 100 were evaluated, but results did not deviate markedly from M = 5, so this was not a parameter manipulated during the simulation. Another consideration was the variance approximation method for $\text{var}(\hat{\delta}_{k-1}^{k})$. Although the exposition in the previous

section focused predominantly on the Taylor series linearization approach, it was commented that one of a class of replication approaches discussed in Rust (1985) would be a viable alternative. To this end, a nonparametric bootstrap routine was investigated during certain initial analyses, with results not substantively differing from those obtained via Taylor series linearization. As such, the particular variance approximation method implemented was deemed immaterial for the purpose of this simulation study.

One additional simulation parameter we did find enlightening to manipulate, however, was the variance of the ε_i terms. In addition to $\varepsilon_i \sim N(0,1)$, we evaluated $\varepsilon_i \sim N(0,9)$. This enabled an assessment of the impact of a more variable underlying distribution of y_i and, thus, a more variable sample mean.

Tables 3.4a and 3.4b below summarize results from the simulation study. The former presents a summarization where $n = 500$ and the latter where $n = 5{,}000$. The metrics tabulated are similar to those appearing in Tables I and II of RGG (2008). The mean stop wave is a useful quantification of the length of data collection prior to declaring phase capacity. Its standard deviation should be unambiguous. The row labeled "Mean NR Error" houses the average distance between the nonresponse-adjusted, abridged data set mean and the full-sample mean over all 1,000 replications. For each simulated sample's stopping wave, a 95% confidence interval on the sample mean was constructed. The "95 Percent Coverage" line measures the percentage of abridged sample mean confidence intervals encompassing the full-sample mean.

One overarching finding is that when the outcome is not associated with wave, as simulated in Conditions 1 and 3, both the imputation and weighting versions of the test are quick to detect phase capacity. Indeed, it is a rare occasion when phase capacity is *not* detected at the second wave. Intuitively, the abridged data set introduces minimal nonresponse error and the full-sample mean is adequately covered by the confidence interval formed on the sample mean at the earlier point in the data collection process. These are promising results that hold for both $n = 500$ and $n = 5{,}000$.

Phase capacity is not declared as quickly for Conditions 2 and 4, those in which a sample unit's expected outcome increases linearly with response wave. Despite the tests often dictating data collection to proceed well beyond the second wave, when $n = 500$, the abridged data set sample means are subject to a nontrivial amount of nonresponse error and an unsatisfactory rate of confidence intervals that cover the full-sample mean. That said, there is a fair amount of variability in terms of the mean stopping wave in the $n = 500$ setting. Another finding is that the mean stopping wave for Condition 2 is somewhat less than Condition 4 over all conditions and phase capacity tends to be detected earlier when the ε_i terms are characterized by a more variable distribution.

A theme permeating the results from Conditions 2 and 4, at least for the case where $n = 500$, is that the weighting version of the phase capacity test tends to call for

more waves of nonresponse follow-up. For the simulation setting in which n = 5,000 summarized in Table 3.4b, the mean stopping point is almost always the tenth (and final) wave. One possible explanation for this difference observed across sample sizes is that a larger sample size results in more precision for the underlying estimates of $\text{var}(\hat{\bar{y}}_1^{k-1})$, $\text{var}(\hat{\bar{y}}_1^{k})$, and, therefore, $\text{var}(\hat{\delta}_{k-1}^{k}) = \text{var}(\hat{\bar{y}}_1^{k-1}) + \text{var}(\hat{\bar{y}}_1^{k}) - 2\text{cov}(\hat{\bar{y}}_1^{k-1}, \hat{\bar{y}}_1^{k})$. Considering these terms comprise the denominator of the quotient that is the phase capacity test, it follows that this renders one more likely to *fail* to reject the test. In other words, as the precision increases, the test becomes more sensitive to observed differences in the two nonresponse-adjusted estimates and dictates more waves of follow-up are needed. On the one hand, this could be perceived as an advantage, as there is seemingly less risk for residual nonresponse error. On the other hand, a lack of precision alone should not be the sole or primary determinant of phase capacity. It may be wise for practitioners to designate a minimum precision threshold that must be met prior to adhering to the conclusions of the tests discussed in this dissertation. It would be presumptuous to recommend any particular threshold(s), as that will depend on the analytic objectives of the survey effort and the estimator of interest, among other factors.

Table 3.4a: Simulation Study Results Comparing RGG Rule 3 with the Weighting Variant (n = 500).

		RGG Rule 3		Weighting	
Condition	**Measure**	$\varepsilon_i \sim N(0,1)$	$\varepsilon_i \sim N(0,9)$	$\varepsilon_i \sim N(0,1)$	$\varepsilon_i \sim N(0,9)$
1. Wave not associated with covariate; outcome not associated with wave	Mean Stop Wave	2.02	2.01	2.00	2.01
	Std. Dev. Stop Wave	0.13	0.10	0.03	0.12
	Mean NR Error	0.00	0.00	0.00	0.00
	95 Percent Coverage	99.80	98.90	100.00	99.80
2. Wave not associated with covariate; outcome associated with wave	Mean Stop Wave	4.36	2.22	7.90	4.54
	Std. Dev. Stop Wave	2.76	0.51	3.52	3.64
	Mean NR Error	-1.62	-2.28	-0.62	-1.60
	95 Percent Coverage	13.20	0.00	73.70	30.50
3. Wave associated with covariate; outcome not associated with wave	Mean Stop Wave	2.01	2.03	2.00	2.02
	Std. Dev. Stop Wave	0.12	0.16	0.00	0.13
	Mean NR Error	-0.01	-0.01	-0.01	-0.01
	95 Percent Coverage	99.70	98.20	100.00	99.80
4. Wave associated with covariate; outcome associated with wave	Mean Stop Wave	3.92	2.17	6.15	3.45
	Std. Dev. Stop Wave	2.51	0.51	3.99	2.93
	Mean NR Error	-1.74	-2.28	-1.13	-1.90
	95 Percent Coverage	8.60	0.00	51.80	16.20

Table 3.4b: Simulation Study Results Comparing RGG Rule 3 with the Weighting Variant (n = 5,000).

		RGG Rule 3		Weighting	
Condition	**Measure**	$\varepsilon_i \sim N(0,1)$	$\varepsilon_i \sim N(0,9)$	$\varepsilon_i \sim N(0,1)$	$\varepsilon_i \sim N(0,9)$
1. Wave not associated with covariate; outcome not associated with wave	Mean Stop Wave	2.01	2.01	2.00	2.02
	Std. Dev. Stop Wave	0.10	0.07	0.04	0.12
	Mean NR Error	0.00	0.00	0.00	0.00
	95 Percent Coverage	99.80	98.40	100.00	99.80
2. Wave not associated with covariate; outcome associated with wave	Mean Stop Wave	10.00	9.76	10.00	10.00
	Std. Dev. Stop Wave	0.00	1.37	0.00	0.00
	Mean NR Error	0.00	-0.07	0.00	0.00
	95 Percent Coverage	100.00	97.00	100.00	100.00
3. Wave associated with covariate; outcome not associated with wave	Mean Stop Wave	2.01	2.01	2.00	2.01
	Std. Dev. Stop Wave	0.12	0.12	0.03	0.09
	Mean NR Error	0.00	0.00	0.00	0.00
	95 Percent Coverage	99.90	98.70	100.00	100.00
4. Wave associated with covariate; outcome associated with wave	Mean Stop Wave	10.00	9.42	10.00	10.00
	Std. Dev. Stop Wave	0.00	2.07	0.00	0.00
	Mean NR Error	0.00	-0.17	0.00	0.00
	95 Percent Coverage	100.00	92.80	100.00	100.00

Also evident from contrasting the mean stopping waves for any given simulation setting is that the weighting version of the test typically calls for more waves of follow-up than RGG Rule 3. Because the expected values of $\hat{\bar{y}}_1^k$ and $\hat{\bar{y}}_1^{k-1}$ are the same for either version, the weighting version of the phase capacity test must produce a smaller value of $\text{var}(\hat{\delta}_{k-1}^k)$. This is confirmed by Figure 3.1, which, for each condition simulated, overlays the two average values of $\text{var}(\hat{\delta}_{k-1}^k)$ at each wave threshold over all 1,000 iterations of the simulation setting where n = 500 and $\varepsilon_i \sim N(0,1)$. One can observe how the variance is consistently smaller for the weighting rule until the two converge near the final wave threshold.

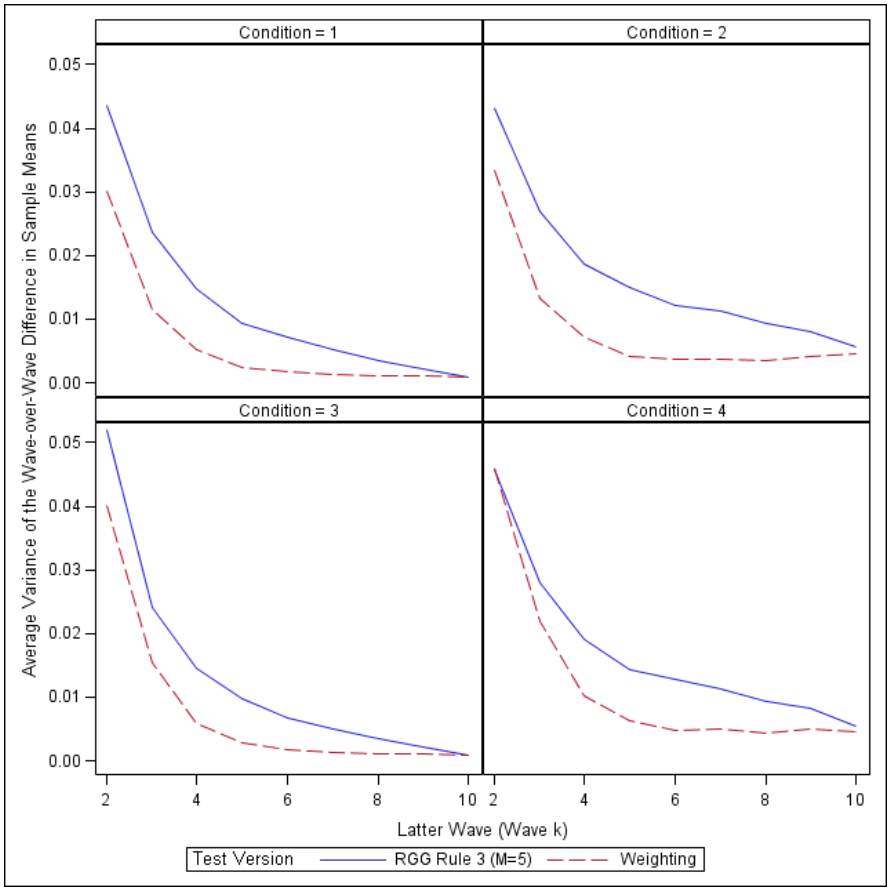

Figure 3.1: Average Approximated Variance of the Difference between Two Adjacent Wave Sample Means by Phase Capacity Test Method for the Simulation Study Setting where n = 500 and $\varepsilon_i \sim N(0,1)$.

Recall both tests' computational algorithms avoid the explicit calculation of $\text{cov}(\hat{\bar{y}}_1^{k-1}, \hat{\bar{y}}_1^{k})$ embedded within $\text{var}(\delta_{k-1}^{k}) = \text{var}(\hat{\bar{y}}_1^{k-1}) + \text{var}(\hat{\bar{y}}_1^{k}) - 2\text{cov}(\hat{\bar{y}}_1^{k-1}, \hat{\bar{y}}_1^{k})$. Bearing in mind the argument made previously regarding the equivalence of the ABB and a single weight inflation factor on the variance of a sample mean, any discrepancy in the overall variance must be attributable to the implicit calculation of $\text{cov}(\hat{\bar{y}}_1^{k-1}, \hat{\bar{y}}_1^{k})$. Clearly, the covariance from the weighting version of the phase capacity test is larger

in magnitude than the like calculated via the difference variable approach of RGG Rule 3. Although the direct derivation of $\text{cov}(\hat{\bar{y}}_1^{k-1}, \hat{\bar{y}}_1^{k})$ would be tedious in either version, an indirect derivation would be to solve for it using known inputs of $\text{var}(\hat{\delta}_{k-1}^{k})$, $\text{var}(\hat{\bar{y}}_1^{k-1})$, and $\text{var}(\hat{\bar{y}}_1^{k})$.

An informative quantification of its effect is $1 - \text{var}(\hat{\delta}_{k-1}^{k}) / (\text{var}(\hat{\bar{y}}_1^{k-1}) + \text{var}(\hat{\bar{y}}_1^{k}))$, which can be interpreted as the proportion of the variance reduced accounting for the covariance. Figure 3.2 below plots this quantity at each wave threshold for the same four conditions and simulation settings as in Figure 3.1. We can see that by about the sixth wave the covariance between the two adjacent nonresponse-adjusted sample means as estimated via the weighting version of the test is so strong that it renders $\text{var}(\hat{\delta}_{k-1}^{k})$ close to zero. The convergence is much more gradual under the RGG approach. Returning to an argument made previously, in the extreme case in which no new respondents are captured after a particular wave of follow-up, $\text{var}(\hat{\delta}_{k-1}^{k})$ would be zero for the weighting version of the test, meaning that incorporating the covariance would result in a 100% reduction in variance. But the same would not hold true for RGG Rule 3 unless the imputation model was perfectly predictive (i.e., all *M* difference variables were 0).

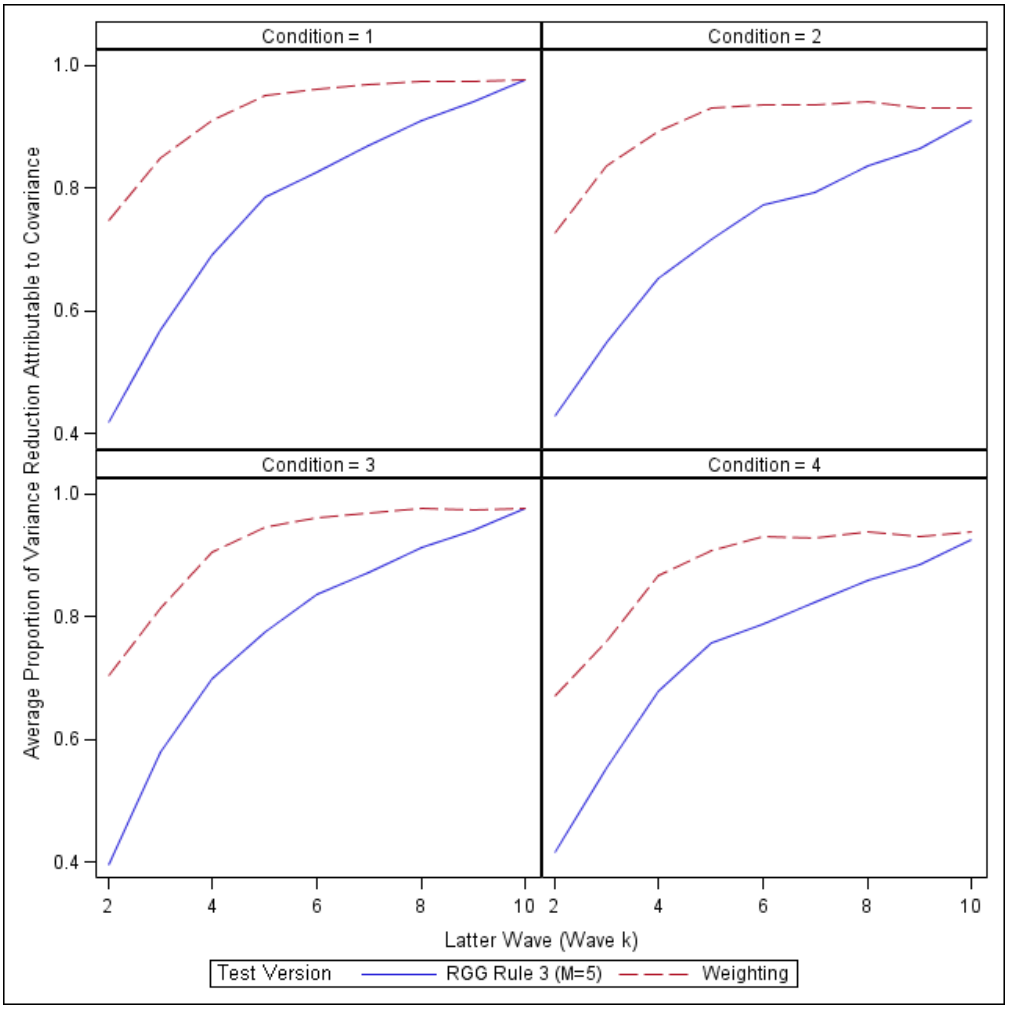

Figure 3.2: Average Proportion of the Approximated Variance of the Difference between Two Adjacent Wave Sample Means Reduced after Incorporating the Covariance by Phase Capacity Test Method for the Simulation Study Setting where $n = 500$ and $\varepsilon_i \sim N(0,1)$.

3.4 Application to the Federal Employee Viewpoint Survey

We next discuss an application of these methods using the three agencies participating in FEVS 2011 whose wave-specific respondent distributions were summarized in Table 1.2. As before, the estimates under investigation are sample

means—namely, the seven percent positive estimates for items constituting OPM's Job Satisfaction index listed in Table 1.1. Note that the interpretation of nonresponse error is subtly different in this application as compared to the simulation study. In the simulation study, the full-sample mean was known for all 1,000 replications of a given sample size and condition, and it was further assumed that unit nonresponse could be eliminated entirely with enough follow-up attempts. Here, we define nonresponse error as the difference between the estimate computed once phase capacity has been declared and the full-sample estimate computed after the agency's maximum wave undertaken during FEVS 2011.

As in the simulation study, the fundamental objective was to evaluate the performance of the two competing tests of phase capacity. To promote a balanced comparison, a shared set of auxiliary variables were used in both nonresponse adjustment procedures: agency-subelement; an indicator of whether the employee works at the agency headquarters or in a field office; gender; a minority/non-minority indicator variable; and supervisory status (non-supervisor, supervisor, and executive).

For the RGG Rule 3 version of the test, these variables served as main effects in a sequence of logistic regression models fitted to impute the missing data, independently fitted for each agency. For nonrespondents at the conclusion of any given wave, the seven positive/non-positive indicators for items comprising the Job Satisfaction index were multiply imputed $M = 5$ times using the %IMPUTE module within IVEware, a free, SAS-callable set of macros developed by researchers at the

Institute for Social Research at the University of Michigan. The macro implements the sequential regression multiple imputation (SRMI) algorithm detailed in Raghunathan et al. (2001).

The SRMI algorithm proceeds as follows. Let **X** denote the fully observed matrix of auxiliary (and possibly outcome) variables and let $y_1, y_2, \ldots, y_P$ represent the sequence of outcome variables subject to missingness ordered according to their item-specific nonresponse rates, smallest to largest. Data are assumed MAR, but the vector of outcome variables need not abide by a monotone pattern. The first step is to impute y_1 using **X**. For the present case where all outcome variables are dichotomous, a sequence of logistic regression models is utilized, but the appendix of Raghunathan et al. (2001) details other model forms that are available within IVEware for alternative variable scales. At this and each subsequent step, the regression model coefficients are independently perturbed prior to deriving each of the *M* imputed values to account for the imputation model uncertainty. Next, one imputes y_2 using **X** and y_1 (including both observed and imputed values), and then proceeds to y_3 using **X**, y_1, and y_2, and so on. In addition to cycling through all *P* variables susceptible to missingness, the algorithm cycles back through the sequence of *P* imputations a predetermined number of "rounds" (p. 87 of Raghunathan et al., 2001) and re-imputes the missing values prior to releasing each of the *M* completed data sets. This is done to build interdependence and foster stability with respect to the imputed values. The %IMPUTE module allows the user to specify this parameter in

advance. For the present analysis, the default setting of five rounds was deemed sufficient.

For the weighting version of the phase capacity test, base weights for the set of respondents at the end of any given wave were *raked* (Kalton and Flores-Cervantes, 2003) to marginal, agency-level totals aggregated from the sample frame. The totals were derived from the same set of categorical variables serving as main effects in the imputation models used in the RGG approach. The SAS macro developed by Izrael, Hoaglin, and Battaglia (2004) was used to carry out the raking process. As with the simulation, Taylor series linearization was utilized to approximate the variance of the adjacent-wave weighted mean difference.

Table 3.5 summarizes the results from the FEVS application. The wave at which phase capacity was declared is given as well as the nonresponse-adjusted estimate at that point and the nonresponse error relative to the nonresponse-adjusted estimate calculated using the ultimate set of respondents. Note that these estimates are not precisely the same when arrived at via multiple imputation versus weighting, but they are close. This is mentioned because the reader may observe that the item-specific sums of the "Estimate" and "Relative NR Error" columns are not always equivalent across the two methods. It is assumed, however, that as $M \rightarrow \infty$, the estimates derived using multiple imputation are asymptotically equivalent to those derived from raking, and so this moderate amount of random variation reflected by the finite M employed should not substantively alter any conclusions made.

In many respects, the conclusions to be gleaned from Table 3.5 coincide with the main takeaways from the simulation study. The weighting version of the test tends to dictate more wave of nonresponse follow-up are needed than does the multiple imputation version proffered by RGG, which surpasses the second wave only in a few instances. Due to the proclivity of the nonresponse-adjusted percent positive estimates to increase with each additional wave (*cf.*, Figure 1.1), it is of little surprise to observe that the nonresponse error is smaller for the weighting variant. The differences are relatively small, however. For example, the average difference in Agency 1's nonresponse error for the seven estimates analyzed is -1.4. This is the largest of such average differences for any of the three agencies examined. Still, 1 to 2 percentage points could make a difference when assessing whether a change relative to the previous years' survey results was statistically significant, a very popular technique human resources managers use to flag items deserving celebration or requiring intervention.

There is a strong negative relationship between wave of response and the absolute value of nonresponse error, which is to say that the nonresponse error tends to decrease with each additional wave. The Pearson correlation coefficient for this relationship is $\rho = -0.53$ ($p < .05$) for the weighting version of the phase capacity test. Calculating the comparable correlation coefficient for the RGG Rule 3 version of the test would not be very informative since there is scant variability in the stopping waves.

Table 3.5: Results from a Federal Employee Viewpoint Survey Application using Data from Three Agencies to Compare RGG Rule 3 with the Weighting Rule Variant.

	RGG MI (M = 5)			Weighting		
Item	**Stopping Wave**	**Estimate**	**Relative NR Error**	**Stopping Wave**	**Estimate**	**Relative NR Error**
Agency 1						
4	3	74.0	-2.0	5	75.3	-0.6
5	2	82.4	-1.7	2	82.6	-1.5
13	2	86.6	-2.2	5	88.6	-0.3
63	3	54.5	-1.7	5	55.7	-0.4
67	2	33.8	-3.3	4	35.8	-1.4
69	2	68.3	-2.9	5	70.8	-0.4
70	2	68.6	-1.6	2	69.1	-1.3
Agency 2						
4	2	79.0	-1.1	2	78.9	-0.5
5	2	84.2	-0.8	2	84.2	-1.2
13	2	86.3	-2.8	2	88.2	-0.9
63	2	62.8	-1.9	2	63.2	-1.4
67	2	40.1	-1.9	3	41.1	-1.4
69	2	73.6	-0.6	3	72.7	-1.1
70	2	63.1	3.0	2	62.2	1.0
Agency 3						
4	2	77.7	-1.7	4	79.1	-0.3
5	2	84.8	-1.4	4	86.2	-0.1
13	2	86.4	-1.3	2	86.9	-0.7
63	2	63.2	-1.5	2	63.4	-1.3
67	2	46.5	-1.8	2	46.3	-1.7
69	2	75.2	-1.8	3	75.7	-1.1
70	2	73.5	-0.4	2	73.8	0.0

Lastly, another result that parallels a finding from the simulation study is how phase capacity is concluded earlier for Agency 2, which is comprised of a notably smaller sample size (n = 1,057) than Agency 1 (n = 16,565) and Agency 3 (n =

17,177). There is no evidence that the upward mobility exhibited in the nonresponse-adjusted percent positive estimates is any less pronounced for Agency 2. As such, we suspect that the decreased precision attributable to the smaller sample size relative to the other two agencies is the most probable explanation.

As was commented previously, this type of analysis addresses only relative nonresponse error, considering the survey estimate using all waves is still subject to error. For an assessment of the more formal definition of nonresponse error, we can treat a portion of sample frame variables as if they were collected on the survey instrument. Two variables investigated for this purpose were employee grade and length of service (in years) with the federal government. Grade is a ranking of sorts for the given individual (and job) based on the traditional General Schedule that forms the basis of the majority of federal employees' salary, from which adjustments are applied depending on one's duty location and employment duration within the particular grade. Grade can take on values between 1 and 15, with larger values generally indicative of higher pay. More information on the General Schedule pay system can be found at http://www.opm.gov/policy-data-oversight/pay-leave/pay-systems/general-schedule/.

The raking macro and %IMPUTE macro within IVEware were implemented with the same set of auxiliary variables as before, with the exception that both pseudo-outcome variables were treated as continuous for a linear regression imputation model to be utilized. For both the RGG Rule 3 and weighting versions of

the test, phase capacity was almost always declared at the second wave for two sets of agency-specific estimates. Because the results were so similar, we have opted to present the results visually in lieu of a tabular summarization. Figure 3.3 displays the trend of nonresponse-adjusted estimates of the mean grade as found by raking the weights of respondents at each point in time. There is a separate trend line for each agency. Also appearing in the plot are three horizontal reference lines denoting the agency-specific, full-sample estimates. Because each agency sample was actually a census, these can be interpreted as true means for the three finite populations, but without loss of generality we still refer to them as full-sample estimates. Figure 3.4 is a plot similar in spirit for length of service.

With respect to grade, the tendency for the trend lines of Agency 1 and 2 to converge towards their respective horizontal reference lines suggests net nonresponse error generally decreases with each wave of data incorporated. For Agency 3, the nonresponse-adjusted mean grade consistently overestimates the full-sample mean. In all cases, however, the absolute value of deviations is relatively minor, even in what appear to be the worst of circumstances. For example, the discrepancy for Agency 2 at the conclusion of the first wave is approximately 0.4, which constitutes a relative absolute error of $0.4 / 12.81 = 3.1\%$.

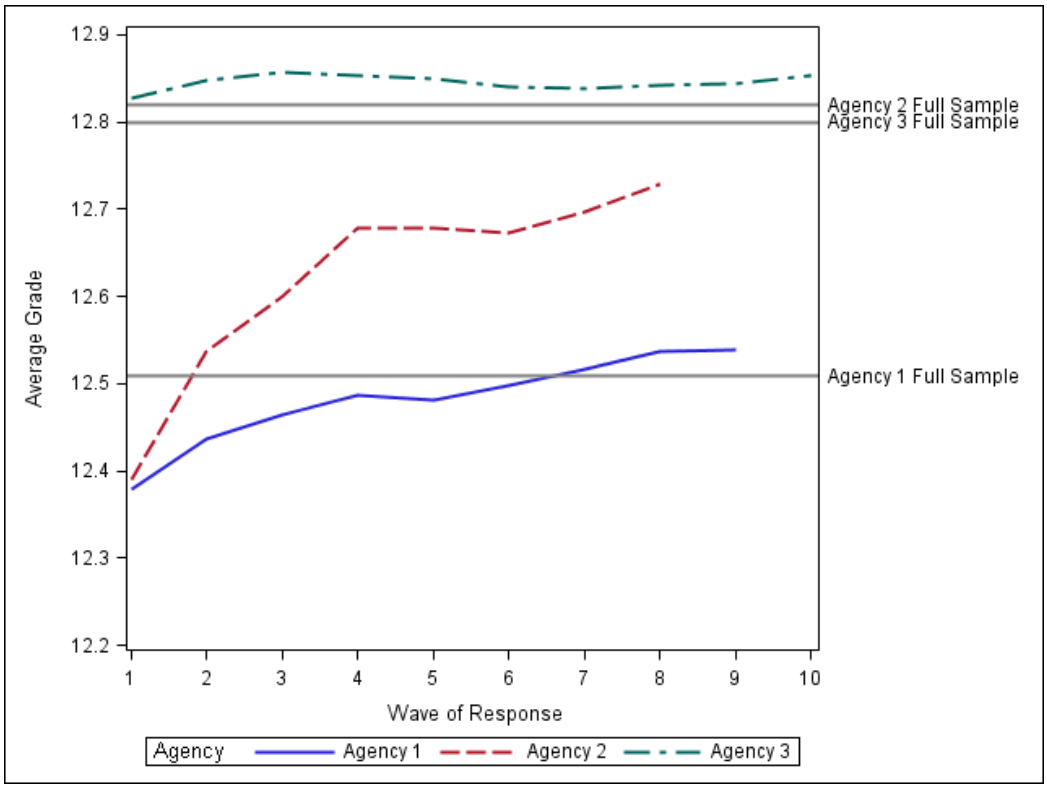

Figure 3.3: Trend of Nonresponse-Adjusted Estimates of Mean Pseudo-Outcome Variable Grade over the 2011 Federal Employee Viewpoint Survey Data Collection Period Overlaid with the Full-Sample Estimate.

Many of the same takeaway messages apply to the other pseudo-outcome variable plotted in Figure 3.4, the average number of years the individual has served as a federal employee at the time the FEVS 2011 was launched. Each agency's nonresponse-adjusted sample mean is typically nearer the full-sample estimate after the conclusion of later waves as compared to earlier waves. The reduced resolution relative to Figure 2.3 may lead one to initially infer absolute errors are smaller, but they are comparable if not greater than those of average grade. For instance, the deviation for Agency 1 at the conclusion of the first wave is approximately 1.2, which

corresponds to a relative absolute error of 1.2 / 18.1 = 6.6%. Still, one could argue that all agency-specific differences are inconsequential by about wave 3.

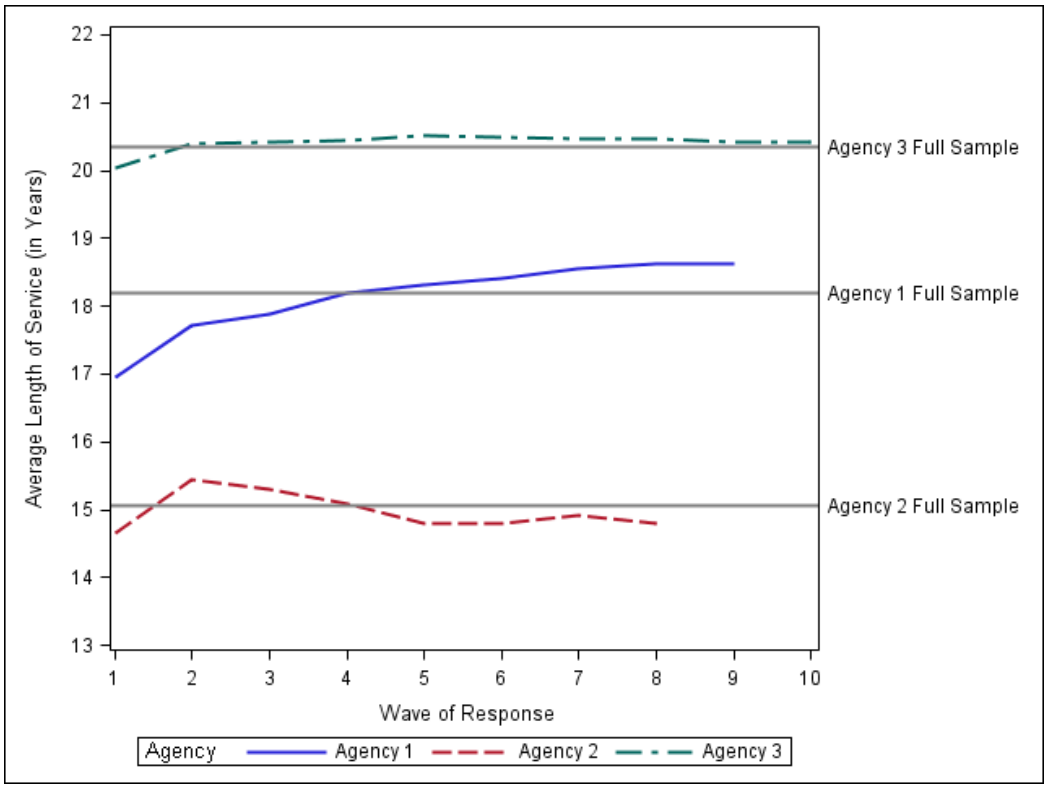

Figure 3.4: Trend of Nonresponse-Adjusted Estimates of Mean Pseudo-Outcome Variable Length of Service over the 2011 Federal Employee Viewpoint Survey Data Collection Period Overlaid with the Full-Sample Estimate.

3.5 Conclusion

The purpose of this chapter was to introduce and evaluate an adaptation of the "Rule 3" test for phase capacity proposed by Rao, Glickman, and Glynn (2008) amenable to scenarios in which weighting adjustments, as opposed to multiple imputation, are implemented to compensate for nonresponse. Although the discourse

centered on the sample mean, the weighting variant is more flexible because it can easily be altered to accommodate other estimators, whereas the *M* difference variable approach proposed by Rao, Glickman, and Glynn (2008) is geared specifically towards investigating a sample mean difference.

A simulation study was mounted to compare and contrast the two approaches. The design was based loosely on the simulation design utilized in Rao, Glickman, and Glynn (2008). Certain modifications were applied to foster a balanced testing ground for the two versions of the test. The results were enlightening. For any condition where the expected value of the outcome variable was stable, or unrelated to the wave in which a response was obtained, both versions were prone to detect phase capacity at the earliest possible point, the second wave. Varying the underlying sample size revealed the interesting finding that, all else equal, a larger sample size can prompt the test to be more sensitive to deviations and conclude more follow-ups are necessary. We surmised that this is a byproduct of increased precision relative to a smaller sample size. The impact of precision was also manifested by manipulating the residual term used to generate the raw data. All else equal, the less variable residual term tended to suggest more waves of follow-up were necessary.

Perhaps the most noteworthy discrepancy unveiled was that the variance of the difference of two adjacent-wave sample means was smaller in the weighting version. Figures 3.1 and 3.2 illustrated how the implicit incorporation of the covariance of the two means—owing to the fact that they are calculated from a shared

portion of the accumulating survey data set—into the approximation of the variance of the difference was much more dramatic for the weighting version. In response to the continually diminishing relative increase in observed data obtained with each new wave of data, the variance of the difference converges to zero much more rapidly. These findings were reaffirmed in the application using data from the 2011 Federal Employee Viewpoint Survey. The weighting version called for more waves and, because the nonresponse-adjusted estimates generally increase with each new wave of data, was less prone to (relative) nonresponse error.

We leave for further research the task of developing a more formal theoretical understanding as to why the covariance is not incorporated equivalently across both methods. A potential objective of such an avenue could be to determine ways to "retune" one method to behave more compatibly with the other. Further research could also explore the behavior of the weighting version of the phase capacity test when monitoring alternative estimators or employing alternative variance approximation methods. Although an admittedly cursory analysis indicated certain replication approaches mirrored the performance of the Taylor Series linearization method described by Woodruff (1971) and utilized herein, a more rigorous study investigating other estimators would be useful for ruling out potential anomalies.

Chapter 4: Multivariate Extensions of the Retrospective Phase Capacity Test When Weighting for Nonresponse

4.1 Background

A poignant limitation of the proposed method to test for phase capacity detailed in the previous chapter is that it is univariate in nature. It is designed to test $H_0: \delta_{k-1}^k = \theta_1^{k-1} - \theta_1^k = 0$ versus $H_1: \delta_{k-1}^k = \theta_1^{k-1} - \theta_1^k \neq 0$, by assessing whether $\hat{\delta}_{k-1}^k$, an estimate of this quantity using data from respondents through wave *k*, is significantly different from zero. Even though the simulation and FEVS application both dealt with sample means, the test can be adapted to other finite population quantities of interest.

In general, however, survey practitioners may not wish to concentrate solely on δ_{k-1}^k, but perhaps $\delta_{(k-1)d}^k$ for $d = 1, \ldots, D$ distinct differences. The subscript *d* could index multiple outcome variables, multiple domains of interest for a particular outcome variable, or even multiple estimators. Although one could conduct the test on each of the *D* differences independently, it is unclear how conflicting results would be coalesced. For instance, suppose the test was conducted on three separate outcome variables' sample mean changes. What is the decision on phase capacity when one variable shows a significant change after incorporating the most recent wave's data, but the other two variables do not? This study proposes two techniques to provide a single yes-or-no answer for these kinds of questions, and compares and contrasts them via simulation and an application using FEVS data.

The first technique, a direct multivariate extension of the *t* test discussed in Section 3.2, involves formulating a Wald chi-square test statistic that resembles a Mahalanobis distance metric. The second technique draws upon ideas from longitudinal data analysis (Singer and Willett, 2003) to test whether the trajectories of change for two or more estimates, measured in terms of their percent changes relative to the previous wave (to harmonize potential scale incongruities), differ substantively from a null trend. At present, consideration is given only to the weighting variant of the phase capacity test; multivariate extensions of the RGG (2008) version are left as an avenue for further research.

4.2 New Methods

The exposition of the first proposed multivariate extension requires us to define some vector and matrix notation. Let **D** represent the *D* x 1 vector of differences $\mathbf{D} = \begin{bmatrix} \hat{\delta}^{k}_{(k-1)1} \\ \hat{\delta}^{k}_{(k-1)2} \\ \vdots \\ \hat{\delta}^{k}_{(k-1)d} \end{bmatrix}$ and **S** denote its *D* x *D* variance-covariance matrix, or $\mathbf{S} = \begin{bmatrix} \text{var}(\hat{\delta}^{k}_{(k-1)1}) & \text{cov}(\hat{\delta}^{k}_{(k-1)1}, \hat{\delta}^{k}_{(k-1)2}) & \cdots & \text{cov}(\hat{\delta}^{k}_{(k-1)1}, \hat{\delta}^{k}_{(k-1)D}) \\ \text{cov}(\hat{\delta}^{k}_{(k-1)2}, \hat{\delta}^{k}_{(k-1)1}) & \text{var}(\hat{\delta}^{k}_{(k-1)2}) & \cdots & \text{cov}(\hat{\delta}^{k}_{(k-1)2}, \hat{\delta}^{k}_{(k-1)D}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{\delta}^{k}_{(k-1)D}, \hat{\delta}^{k}_{(k-1)1}) & \text{cov}(\hat{\delta}^{k}_{(k-1)D}, \hat{\delta}^{k}_{(k-1)2}) & \cdots & \text{var}(\hat{\delta}^{k}_{(k-1)D}) \end{bmatrix}$. In other words, **D** consists of the *D* estimate-specific differences between two respondent sets, one through wave *k* and another through wave $k - 1$, where each "theta" term embedded within the d^{th} difference $\hat{\delta}^{k}_{(k-1)d} = \hat{\theta}^{k-1}_{1d} - \hat{\theta}^{k}_{1d}$ is calculated using the pertinent sets of nonresponse-adjusted base weights, w^{k}_{1} and w^{k-1}_{1}. We can think of **D** as an estimate

of $\boldsymbol{\Delta}_{k-1}^{k}$, the *D* x 1 vector comprised of unknown terms $\delta_{(k-1)d}^{k} = \theta_{1d}^{k-1} - \theta_{1d}^{k}$. Furthermore, note that **S** is a symmetric matrix with the *D* difference-specific variances terms along the diagonal and the $\binom{D}{2}$ difference-to-difference covariance terms in the off-diagonal. We have already discussed two methods to estimate the variance terms, one using Taylor series linearization (TSL) and another using replication. Practitioners may find the replication approach more efficient in this multivariate context due to the potentially large number of terms in **S** and the straightforward manner in which these techniques can be used to populate its entries. For example, an efficient computational strategy is to construct a summary table where each row represents a replicate and a series of *D* columns stores the replicate-specific deviations from the full-sample difference. When oriented in this manner, variances are a simple function of the sum of these squared deviations, and covariances a function of the cross-products, where the particular formula to be used is contingent upon the underlying replication technique employed. The computational nuisance associated with applying the TSL method proposed by Woodruff (1971) and detailed in Section 3.2 is that, in addition to finding the variance of the sum of *D* distinct linear substitutes, one must also derive $\binom{D}{2}$ covariances.

The multivariate assessment of phase capacity hinges on the hypothesis test $H_0: \boldsymbol{\Delta}_{k-1}^{k} = \mathbf{0}$ versus $H_1: \boldsymbol{\Delta}_{k-1}^{k} \neq \mathbf{0}$, where **0** symbolizes a *D* x 1 vector of zeros, a null vector indicating none of the *D* differences are significant. If we fail to reject the null

hypothesis and conclude $\boldsymbol{\Delta}_{k-1}^{k}$ is not significantly different from $\mathbf{0}$, we declare phase capacity has occurred. The hypothesis test is carried out by calculating a Wald chi-square test statistic (p. 168 of Heeringa et al., 2010)

$$\chi_W^2 = \mathbf{D}^{\mathbf{T}}\mathbf{S}^{-\mathbf{1}}\mathbf{D} \tag{4.1}$$

which is a scalar distributed as a random chi-square variate with $D - 1$ degrees of freedom under the null hypothesis. Thus, the corresponding *p*-value for the observed test statistic can be ascertained using that reference distribution.

The second multivariate extension stems from concepts of longitudinal data analysis (Singer and Willett, 2003). The notion is to assess whether there is a non-zero trajectory of change across all *D* estimates; hence, we term this the *non-zero trajectory method*. The first step is to estimate the three most recent wave-over-wave relative percent changes in all *D* nonresponse-adjusted estimates, a measure chosen to ensure all estimate differences adhere to a common scale. One immediately evident aspect of this method is that it mandates a minimum of four waves of data. This is a notable drawback relative to the other retrospective methods discussed thus far, which only necessitated a minimum of two waves of data. Nonetheless, the approach is intuitive and straightforward to apply.

For sake of a numerical example, suppose a particular agency participating in FEVS sought to test whether $D = 3$ percent positive estimates, those based on items 4,

5, and 13, have stabilized after the three most recent waves of nonrespondent follow-up. The three estimates' trends and associated percent changes relative to the previous wave are summarized in Table 4.1.

Table 4.1: Example FEVS Trends for Three Items' Percent Positive Estimates across the Four Most Recent Waves.

Wave	Item 4	Item 4 Rel % Chg	Item 5	Item 5 Rel % Chg	Item 13	Item 13 Rel % Chg
k - 3	75.2%	--	83.6%	--	88.5%	--
k - 2	75.3%	0.2%	83.8%	0.2%	88.6%	0.1%
k - 1	75.7%	0.5%	83.9%	0.2%	88.6%	0.0%
k	76.1%	0.4%	84.2%	0.3%	88.7%	0.2%

The idea is to model Δ_d, the d^{th} estimate's relative percent change, as a function of the data collection wave. Specifically, in the presence of *D* distinct differences, if we let *w* represent the data collection wave, a predictor variable taking the form of an integer one unit apart (e.g., 0, 1, and 2), the following model is estimated:

$$\Delta_d = \beta_{01} + \beta_{02} + \ldots + \beta_{0D} + \beta_{11}w + \beta_{12}w + \ldots + \beta_{1D}w + \varepsilon_d \qquad (4.2)$$

Notice how the model specification in equation 4.2 allows each estimate's change to have its own unique intercept and slope. The $\beta_{0\bullet}$ terms represent estimate-specific intercepts, the $\beta_{1\bullet}$ terms represent estimate-specific slopes, and ε_d is an error term assumed to be normally distributed with some unknown but constant

variance σ_d^2. If phase capacity has truly been reached, we would anticipate all $2D$ estimated model parameters to be statistically indistinguishable from 0. As we will demonstrate shortly, from the theory of general linear models, an F test can be conducted to formally test veracity of this assertion.

Using the last three lines of Table 4.1, the estimated model will have $2D = 6$ terms, $D = 3$ estimate-specific intercepts and $D = 3$ estimate-specific slopes. The model parameters can be estimated using standard matrix theory of ordinary least squares regression after first creating the outcome vector $\mathbf{\Delta} = \begin{bmatrix} 0.2 \\ 0.5 \\ 0.4 \\ 0.2 \\ 0.2 \\ 0.3 \\ 0.1 \\ 0.0 \\ 0.2 \end{bmatrix}$ and design matrix $\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{bmatrix}$.

Figure 4.1 offers a visualization of the model using data from Table 4.1, which can be conceptualized as $D = 3$ simple linear regression models. Most points

lie above 0 on the *y*-axis scale, reflecting of an increasing trend in the estimate over the four most recent wave thresholds. Granted, the relative percent changes are modest.

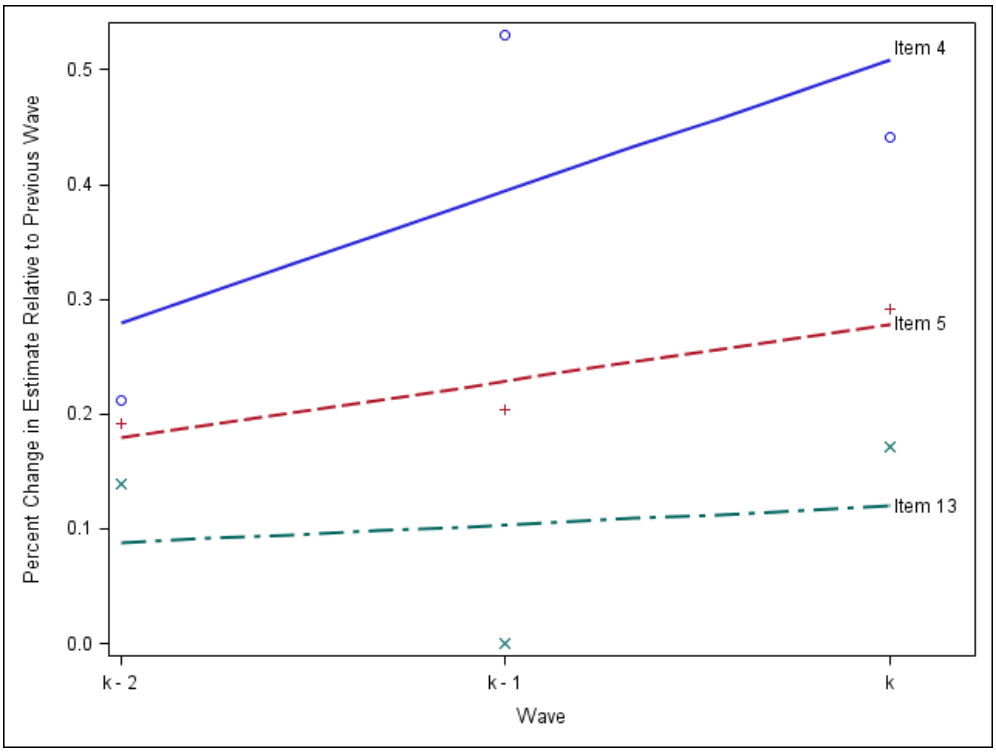

Figure 4.1: Visualization of the Non-Zero Trajectory Method for Testing Phase Capacity in a Multivariate Setting.

If we denote the estimated model parameters by the 2*D* x 1 vector $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\boldsymbol{\Delta}$ and the corresponding 2*D* x 2*D* covariance matrix $\text{cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}_d^2(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$, where $\hat{\sigma}_d^2$ is the estimated mean squared error of the model—an estimate of $\text{var}(\varepsilon_d)$—the multivariate assessment of phase capacity is contingent upon results of the hypothesis test $H_0: \boldsymbol{\beta} = \mathbf{0}$ versus $H_1: \boldsymbol{\beta} \neq \mathbf{0}$, where $\mathbf{0}$ is a 2*D* x 1 vector of zeros. Specifically, the test statistic is calculated as

$$F = \hat{\boldsymbol{\beta}}^{\mathbf{T}}\left(\text{cov}(\hat{\boldsymbol{\beta}})\right)^{-1}\hat{\boldsymbol{\beta}} \qquad (4.3)$$

which is the same as the overall *F* test statistic provided in the "Model" line of an analysis of variance (ANOVA) table. This test statistic can be referenced against an *F* distribution with 2*D* numerator degrees of freedom and *D* denominator degrees of freedom at the desired significance level. When the observed value of this test statistic fails to be statistically significant, there is evidence phase capacity has been reached.

In their purest forms, both methods implicitly treat each wave-over-wave estimate difference with equal importance. There may be occasions, however, when a practitioner wishes to assign differential degrees of importance. For instance, perhaps one of the *D* estimates is deemed "most important." The practitioner still seeks an overall test of phase capacity, but would like any conclusion(s) made more sensitive to changes in that estimate than the others. Given a set of user-defined relative weights, either method can easily be tailored.

To motivate a simple example using the data from Table 4.1, suppose one wanted changes in item 4 to be valued twice as much as changes in items 5 and 13. For the Wald chi-square approach, these relative weights could be incorporated by introducing a vector $\mathbf{C}^{\mathbf{T}}$ = [2 1 1] into the corresponding test statistic as follows: χ^2_W = $(\mathbf{C}^{\mathbf{T}}\mathbf{D})^{\mathbf{T}}(\mathbf{C}^{\mathbf{T}}\mathbf{S}\mathbf{C})^{-1}(\mathbf{C}^{\mathbf{T}}\mathbf{D})$. Similarly, for the second method, assuming an estimated

model parameter matrix $\hat{\boldsymbol{\beta}}^{\mathbf{T}} = \left[\hat{\beta}_{0(4)} \quad \hat{\beta}_{0(5)} \quad \hat{\beta}_{0(13)} \quad \hat{\beta}_{1(4)} \quad \hat{\beta}_{1(5)} \quad \hat{\beta}_{1(13)}\right]$, where the subscripts in parentheses reference the particular item, one could introduce $\mathbf{C}^{\mathbf{T}}$ = [2 1 1 2 1 1] and compute $F = (\mathbf{C}^{\mathbf{T}}\hat{\boldsymbol{\beta}})^{\mathbf{T}}\left(\mathbf{C}^{\mathbf{T}}\,\mathrm{cov}(\hat{\boldsymbol{\beta}})\mathbf{C}\right)^{-1}(\mathbf{C}^{\mathbf{T}}\hat{\boldsymbol{\beta}})$. Note that the reference distributions would not change under either version of the test.

4.3 Simulation Study

This section details a simulation study conducted to compare and contrast the performance of the two proposed multivariate extensions of the phase capacity test. Instead of generating data for outcome variables using one or more parametric distributions, the simulation study undertaken exploits observed data patterns from the FEVS 2011. Respondents from the same three agencies utilized in the first study were treated as three distinct, complete sample data sets and were independently partitioned into 10 distinct wave cohorts 1,000 times according to one of two conditions to be defined shortly. The "sample sizes" of these three agencies are the ultimate respondent counts reported in Table 1.2—namely, Agency 1 consisted of n = 8,105, Agency 2 of n = 572, and Agency 3 of n = 8,687, for a total sample size of 17,364. Percent positive estimates from the D = 7 items comprising the Job Satisfaction Index (see Table 1.1) were chosen as the set of items to simultaneously evaluate in this study. We chose this particular set of items because it constitutes one of the four Human Capital Assessment and Accountability Framework (HCAAF) indices established by the Chief Human Capital Officers Act of 2002. These indices delineate four distinct workplace dimensions along which the FEVS participating

agencies are ranked. The other three HCAAF indices will be analyzed as part of the FEVS application discussed in Section 4.4. As was defined previously, any given index is obtained by simply averaging the weighted percent positive estimates for all items therein. For completeness, Table 4.2 enumerates the FEVS 2011 item numbers and wording associated with each of the four indices.

Table 4.2: Items Comprising the U.S. Office of Personnel Management's Four Human Capital Assessment and Accountability Framework (HCAAF) Indices Derived from the Federal Employee Viewpoint Survey.

Job Satisfaction Index (JS)

Item	Wording
4	My work gives me a feeling of personal accomplishment.
5	I like the kind of work I do.
13	The work I do is important.
63	How satisfied are you with your involvement in decisions that affect your work?
67	How satisfied are you with your opportunity to get a better job in your organization?
69	Considering everything, how satisfied are you with your job?
70	Considering everything, how satisfied are you with your pay?

Leadership and Knowledge Management Index (LKM)

Item	Wording
10	My workload is reasonable.
35	Employees are protected from health and safety hazards on the job.
36	My organization has prepared employees for potential security threats.
51	I have trust and confidence in my supervisor.
52	Overall, how good a job do you feel is being done by your immediate supervisor/team leader?
53	In my organization, leaders generate high levels of motivation and commitment in the workforce.
55	Managers/supervisors/team leaders work well with employees of different backgrounds.
56	Managers communicate the goals and priorities of the organization.
57	Managers review and evaluate the organization's progress toward meeting its goals and objectives.
61	I have a high level of respect for my organization's senior leaders.
64	How satisfied are you with the information you receive from management on what's going on in your organization?
66	How satisfied are you with the policies and practices of your senior leaders?

Results-Oriented Performance Culture Index (ROPC)

Item	Wording
12	I know how my work relates to the agency's goals and priorities.
14	Physical conditions (for example, noise level, temperature, lighting, cleanliness in the workplace) allow employees to perform their jobs well.
15	My performance appraisal is a fair reflection of my performance.
20	The people I work with cooperate to get the job done.
22	Promotions in my work unit are based on merit.
23	In my work unit, steps are taken to deal with a poor performer who cannot or will not improve.
24	In my work unit, differences in performance are recognized in a meaningful way.
30	Employees have a feeling of personal empowerment with respect to work processes.
32	Creativity and innovation are rewarded.
33	Pay raises depend on how well employees perform their jobs.
42	My supervisor supports my need to balance work and other life issues.
44	Discussions with my supervisor/team leader about my performance are worthwhile.
65	How satisfied are you with the recognition you receive for doing a good job?

Talent Management Index (TM)

Item	Wording
1	I am given a real opportunity to improve my skills in my organization.
11	My talents are used well in the workplace.
18	My training needs are assessed.
21	My work unit is able to recruit people with the right skills.
29	The workforce has the job-relevant knowledge and skills necessary to accomplish organizational goals.
47	Supervisors/team leaders in my work unit support employee development.
68	How satisfied are you with the training you receive for your present job?

The outcome variables for all 17,364 distinct respondent records amongst the three agencies were fixed in all 1,000 simulations, but the order in which they were observed varied. For each simulation, a response wave between 1 and 10 was randomly assigned to each record based on one of two conditions crafted similarly in spirit to those used the first study—see Section 3.3 for a description. In Condition 1, an employee's response wave was generated independently from his or her outcome variables, whereas in Condition 2, response wave was simulated in such a way that

earlier respondents tended to be less positive. To maintain a realistic apportionment of the sample into 10 waves, the same distributions from Table 3.3 were employed. They are reproduced in Table 4.3 for ease of reference. Recall these percentages reflect the wave-specific distribution of FEVS 2011 respondents from Agency 3 (i.e., as originally reported in Table 1.2). As was commented in Section 3.3, the tacit assumption with this simulation study design is that nonresponse error can be extirpated altogether given enough waves of nonrespondent follow-up. Although this is not necessarily realistic, it enables a comprehensive comparison of the two methods' performance.

Table 4.3: Summary of the Two Wave-of-Response Distributions Used for the Simulation Study Comparing the Two Multivariate Extensions to the Phase Capacity Test When Weighting for Nonresponse.

	Condition 1: Wave Not Associated with Outcome Variables	**Condition 2: Wave Associated with Outcome Variables**	
Wave	All Respondents	Less Satisfied Respondents	More Satisfied Respondents
1	25.1%	34.5%	15.6%
2	17.5%	20.7%	14.2%
3	15.0%	11.5%	18.5%
4	11.0%	9.2%	12.9%
5	7.1%	4.6%	9.5%
6	5.9%	4.6%	7.1%
7	5.1%	3.7%	6.4%
8	4.4%	3.5%	5.3%
9	4.7%	3.9%	5.5%
10	4.4%	3.7%	5.0%
	100.0%	100.0%	100.0%

In Condition 1, each respondent was assigned as responding in wave 1 with probability 0.251, wave 2 with probability 0.175, and so on. For Condition 2, respondents were partitioned into two groups of roughly equal size based on an aggregate measure of their degree of satisfaction with the seven Job Satisfaction index items. Specifically, the Likert-scale responses for all seven items were converted to integers between 1 and 5 such that a 1 represented the most negative response (e.g., Very Dissatisfied) and a 5 represented the most positive response (e.g., Very Satisfied). The seven integers were then summed at the respondent level to create an aggregate measure of satisfaction ranging from a minimum of 7 (7 x 1) to a maximum of 35 (7 x 5). Two classes of respondents were then defined: (1) less satisfied respondents, or those respondents those whose aggregate measure fell below the median; and (2) more satisfied respondents, those whose aggregate measure fell above the median. An independently generated random uniform variate between 0 and 1 was first added to each aggregate measure to eliminate the possibility of ties and produce two groups of approximately equal size. Despite being a bit ad-hoc, we felt this classification scheme sufficiently met the principal objective to simulate a scenario in which the outcome variables were associated with the response wave. To provide a few numbers with respect to the specifications given in Table 4.3, the less satisfied respondents were assigned wave 1 with probability 0.345, and the more positive respondents were assigned wave 1 with probability 0.156. Furthermore, recall from the discussion in Section 3.3 that these percentages were designed such that the expected marginal percentage of wave 1 respondents in the whole of Condition 2 matches that of Condition 1, since $0.5*(34.5 + 15.6) \approx 25.1\%$.

The bifurcation of respondents based on the aggregate measure of satisfaction was not performed overall or by agency; rather, it was performed within one of 12 classes defined by the cross-classification of agency, minority status, and supervisory status (supervisor or non-supervisor). These 12 categorizations were also used as weighting classes for Conditions 1 and 2. To conduct a real-time nonresponse adjustment procedure, the sum of weights for respondents at the conclusion of each simulated wave from within the c^{th} class was calibrated such that it matched the known population total N_c. For the Wald chi-square method, these sets of weights were used as part of the TSL method by computing linear substitutes for wave-over-wave differences in a percent positive estimate following the general procedure outlined in Section 3.2. If we denote the d^{th} item's linear substitute for the i^{th} sample unit u_{di}, the corresponding diagonal term of $\mathbf{S}$, $\text{var}(\hat{\delta}^k_{(k-1)d})$, was estimated by finding $\text{var}(\sum_{i=1}^{n} u_{di}) = \text{var}(\hat{u}_d)$. The off-diagonal terms, or covariances between estimated differences d and d' ($d \neq d'$), were found by computing $\text{cov}(\hat{u}_d, \hat{u}_{d'})$.

Results from the simulation study are summarized in Table 4.4. Most of the quantities reported are the same as those appearing in Tables 3.4a and 3.4b. The measure labeled "Mean Stop Wave" represents the average data collection wave at which phase capacity was declared over all 1,000 iterations. The standard deviation of this average follows immediately thereafter. The measure labeled "Mean NR Error for Index" houses the average magnitude of nonresponse error in the Job Satisfaction index at the point phase capacity was determined, which, considering its expression

as the average of the seven underlying percent positive estimates, can also be interpreted as the mean nonresponse error amongst the seven items comprising the index. Below that is the root mean squared error (RMSE) of the index at the point of phase capacity, averaged over all 1,000 simulations, where the RMSE is defined as square root of the sum of the following two quantities: (1) the nonresponse error of the index squared, and (2) the approximated variance of the index, which was derived via Taylor series linearization as detailed in Lewis (2012). The final quantity reported is the percentage of 95% confidence intervals formed about the Job Satisfaction index at the point phase capacity was declared that encompassed the index as calculated from the full sample.

The first broad finding is that, under the first condition in which response wave is not associated with the outcome variables, both methods tend to detect phase capacity at their respective earliest possible points to do so: the second wave for the Wald chi-square method and the fourth for the non-zero trajectory method. For example, the mean stopping wave for Agency 1 was 2.05 for the former method and 4.16 for latter. There is scant differentiation amongst the three agencies investigated for any particular method, but the non-zero trajectory method appears to exhibit more variability in the mean stopping wave relative to the Wald chi-square method. Not surprisingly, there is very little nonresponse error in the Job Satisfaction index introduced by curtailing the data collection period in Condition 1. Additionally, confidence intervals formed around the index estimated once phase capacity was first

reached almost always cover the index value that would be obtained once all sample data is collected.

In Condition 2, the expected values of the seven percent positive estimates (and thus the index) were predisposed to increase with each subsequent wave of data incorporated. To the extent that the employees' varying degrees of satisfaction are not completely explained by the cross-classification of agency, minority status, and supervisory status, the three variables used in the weighting class adjustment procedure, we would anticipate some residual nonresponse error associated with stopping data collection early. Indeed, this is plainly observed in Table 4.3. Despite both methods generally calling for more than the absolute minimum number of waves, they often detect phase capacity prior to the tenth wave and, as such, are susceptible to nonresponse error and a decreased likelihood that the confidence interval formed about the index using the abridged data set contains the full-sample index.

Interestingly, at least for Condition 2, both methods proposed declare phase capacity earlier for Agency 2 than the other two agencies. Under the Wald chi-square approach, the mean stopping wave for Agency 2 is 2.13, in contrast to 6.84 and 6.12 for Agency 1 and 3, respectively. This is coupled with a much larger mean nonresponse error over the 1,000 simulations. At -5.76, the value observed for Agency 2 is roughly 3 times the like for Agency 1 (-1.55) and Agency 3 (-2.01). A similar story emerges comparing the 95% confidence interval coverage rates. A

possible explanation is that, at $n = 572$, the sample size for Agency 2 is much less than the sample sizes for the other two agencies, both of which exceed 8,000. Recall one of the conclusions made from the first study was that, all else equal, a smaller sample size led to the determination that fewer waves of follow-up were necessary. Although that finding pertained to the univariate version of the phase capacity test, this provides some evidence that the same may hold true for the multivariate version.

Table 4.4: Simulation Study Results Comparing the Two Multivariate Extensions to the Phase Capacity Test When Weighting for Nonresponse.

Method: Wald Chi-Square

Condition	Measure	Agency 1	Agency 2	Agency 3
1. Wave not associated with outcome variables	Mean Stop Wave	2.05	2.09	2.06
	Std. Dev. of Stop Wave	0.22	0.33	0.24
	Mean NR Error of Index	0.00	0.05	0.00
	Mean RMSE of Index	0.60	2.23	0.57
	95% CI Coverage Rate for Index	98.71	98.33	98.93
2. Wave associated with outcome variables	Mean Stop Wave	6.84	2.13	6.12
	Std. Dev. of Stop Wave	2.72	0.48	2.89
	Mean NR Error of Index	-1.55	-5.76	-2.01
	Mean RMSE of Index	1.71	6.08	2.14
	95% CI Coverage Rate for Index	64.98	8.64	56.13

Method: Non-Zero Trajectory

Condition	Measure	Agency 1	Agency 2	Agency 3
1. Wave not associated with outcome variables	Mean Stop Wave	4.17	4.17	4.16
	Std. Dev. of Stop Wave	0.46	0.45	0.44
	Mean NR Error of Index	0.00	0.02	-0.01
	Mean RMSE of Index	0.43	1.61	0.41
	95% CI Coverage Rate for Index	99.72	99.47	99.38
2. Wave associated with outcome variables	Mean Stop Wave	6.79	5.16	6.76
	Std. Dev. of Stop Wave	2.95	1.68	2.95
	Mean NR Error of Index	-1.22	-1.59	-1.15
	Mean RMSE of Index	1.38	2.23	1.31
	95% CI Coverage Rate for Index	47.64	74.54	46.89

4.4 Application to the Federal Employee Viewpoint Survey

In this section we turn our attention to an FEVS application. Rather than comparing and contrasting the two methods via a simulated data collection period, actual outcome variable patterns exhibited by the three agencies over the data collection period are utilized. And instead of analyzing only the seven items

comprising the Job Satisfaction index, we extend our investigation to include the other three HCAAF indices: (1) the twelve Leadership and Knowledge Management Index items; (2) the thirteen Results-Oriented Performance Culture Index items; and (3) the seven Talent Management Index items. With respect to nonresponse adjustments, the same wave-specific weights produced from the raking procedure described in Section 3.4 were employed. Recall that the procedure calibrates the weights of employees in the cumulating respondent sets such that they sum to known marginal agency totals of the first level of work unit below agency, an indicator of whether the employee works at headquarters or in a field office, a minority status indicator, gender, and supervisory status (non-supervisor, supervisor, or executive). Other than these itemized differences, the application of the two methods was identical to that from the previous section.

From the previous study we observed how the raking procedure described above rendered wave-specific weights and corresponding estimates that did not completely eliminate nonresponse error, evident from the fact that a discernible upward trend could be noted when plotting the percent positive estimates as a function of response wave. Although there are generally fewer and fewer new responses obtained in each subsequent wave, the latter respondents are disproportionally more positive, causing the nonresponse-adjusted percent positive estimates to increase over the course of data collection. Figure 4.2 confirms that the same holds true for the index estimates, which is not surprising considering they are merely averages of the percent positive estimates. It illustrates how all four

nonresponse-adjusted HCAAF indices increase with each new wave of data collected for Agency 1. Though not shown here, a comparable conclusion can be gleaned from plots for the other two agencies.

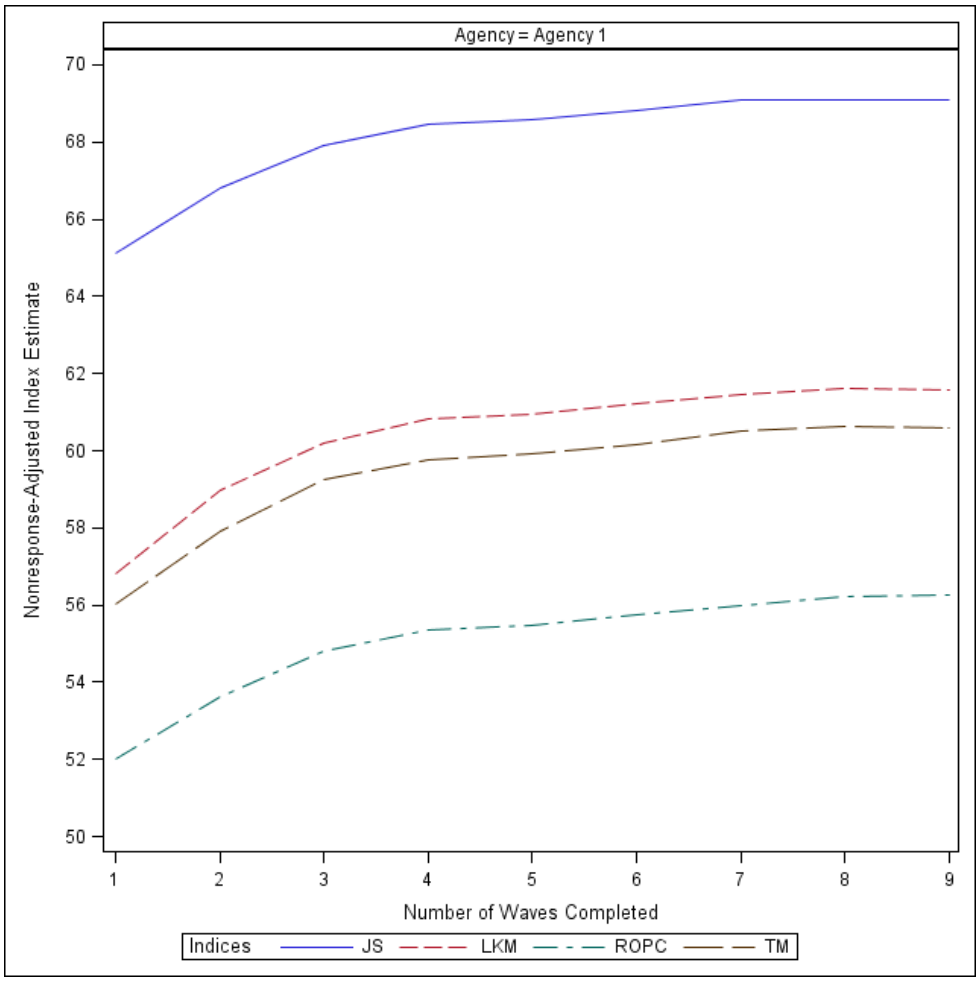

Figure 4.2: Plot of the Nonresponse-Adjusted Indices for Agency 1 Using Cumulative Data as of the Given Wave of Nonrespondent Follow-Up.

Table 4.5 summarizes results from the FEVS application. The column labeled "Stopping Wave" reports the wave at which phase capacity was declared, and is flanked by the corresponding nonresponse-adjusted estimate of the given index and

residual nonresponse error, where applicable. We say "where applicable" because phase capacity was not always declared prior to the final wave of data collection, such as the case for three of the HCAAF indices for Agency 1 under the non-zero trajectory method. As a result, there was no nonresponse error for these three index estimates. Of course, without acquiring the missing attitudinal data from the ultimate nonrespondents, those who never responded by the agency's final data collection wave, we can consider this only as relative nonresponse error, not absolute nonresponse error.

A ubiquitous finding is that the Wald chi-square method tends to declare phase capacity much sooner than the non-zero trajectory method. Indeed, there are no instances among the 12 indices tracked where the non-zero trajectory method calls for fewer waves of nonrespondent follow-up than the Wald chi-square method. This is certainly influenced by the fact that the former requires a minimum of four waves as opposed to two like the latter. That said, all else equal, the average stopping wave for the non-zero trajectory method (6.4) deviates further away from it minimum than the like for the Wald chi-square method (2.9). Given the tendency for the percent positive estimates underlying the estimate to increase with each new set of responses received, the more expeditious determination of phase capacity is coupled with a larger absolute magnitude of nonresponse error. For instance, we can note from Table 4.5 that the maximum absolute nonresponse error in the non-zero trajectory method is 0.5, whereas only two indices' nonresponse error measures fall below that threshold in the Wald chi-square method.

Table 4.5: Results from the FEVS Application Comparing the Two Multivariate Extensions to the Phase Capacity Test When Weighting for Nonresponse.

	Method: Wald Chi-Square			*Method: Non-Zero Trajectory*		
Index	**Stopping Wave**	**Estimate**	**NR Error**	**Stopping Wave**	**Estimate**	**NR Error**
Agency 1						
JS	4	68.5	-0.6	6	68.8	-0.2
LKM	3	60.2	-1.4	9	61.6	0.0
ROPC	2	53.6	-2.6	9	56.2	0.0
TM	5	59.9	-0.7	9	60.6	0.0
Agency 2						
JS	2	69.8	-1.0	5	71.0	0.1
LKM	2	72.8	-0.4	5	73.1	0.1
ROPC	4	66.3	0.1	5	66.4	0.2
TM	2	68.7	-1.3	5	70.0	0.1
Agency 3						
JS	3	73.1	-0.7	6	73.5	-0.3
LKM	2	70.5	-1.3	7	71.5	-0.2
ROPC	4	63.7	-0.6	5	63.8	-0.5
TM	2	69.4	-1.0	6	70.2	-0.2

4.5 Conclusion

This chapter proposed two multivariate extensions of the methods discussed in Chapter 3 for detecting phase capacity when weighting for nonresponse. Hence, one notable absence in the chapter is that we did not pursue any multivariate extensions of the method detailed in Rao, Glickman, and Glynn (2008), the competing test for phase capacity discussed in Chapter 3. Indeed, as previously noted, we leave this as an avenue for further research.

The stated objective at the outset was to develop and evaluate multivariate methods that consolidate the *D* wave-specific estimates and their associated measures of variability into a single yes-or-no answer as to whether phase capacity has occurred. The first method was to formulate a Wald chi-square test statistic in a straightforward multivariate extension of the *t* tests discussed in Chapter 3. The second method utilized concepts of longitudinal analysis (Singer and Willett, 2003) to assess whether the trajectories of change for the *D* estimates were jointly distinguishable from 0. If not, it would be indicative of a null trend suggesting that the estimates have stabilized.

The two methods were contrasted via simulation and application using data from FEVS 2011. Both the simulation and application revealed that, all else equal, the non-zero trajectory detection method tends to dictate more wave of nonrespondent follow-up are warranted, in large part because it requires a minimum of four waves of data, whereas the Wald chi-square method requires only two. Naturally, in settings where nonresponse error lingers even after weighting adjustments have been implemented, the non-zero trajectory method yields estimates with a smaller relative error. But the Wald chi-square method's proclivity for declaring phase capacity sooner proves efficient when there is no relationship between response wave and the outcome variables, as was the case in Condition 1 of the simulation study.

The four HCAAF indices published by OPM served as the sets of underlying estimates jointly tested for phase capacity, which is a bit limiting since focus was

restricted only to ratios and differences of ratios. Future research could investigate alternative estimators, such as differences in regression coefficients or quantiles. For more disparate sets of estimators, a replication approach is recommended when populating **S** as part of the Wald chi-square method. Presently, consideration was only given to the Taylor series linearization method for acquiring both the variance terms along diagonal of **S** and the covariance terms populating off-diagonal entries. The derivation of linear substitutes following the technique proposed by Woodruff (1971) was tractable in our setting, but the same may not be true for other estimator differences. Although we do not anticipate any situations that would lead to substantive differences between the Taylor series linearization approach and replication, except perhaps tracking quantiles and using the jackknife method of variance approximation (Kovar et al., 1988), additional simulations, applications to other surveys, or theoretical developments would be welcomed to help eradicate any such possibilities.

Chapter 5: Prospective Considerations of Phase Capacity

5.1 Background

A general criticism about the methods discussed in the first two studies is that they are retrospective in nature. Knowing the most recent wave's data did not significantly modify a key point estimate is useful information, but knowing so before conducting an inefficacious wave of data collection would be even more valuable. Acknowledging this, Wagner and Raghunathan (WR) (2010) proposed a "stop-and-impute" test that is prospective in nature. They focused on a continuous outcome variable of which the sample mean is of central interest and, as with RGG (2008), assume auxiliary variables are available on a sample frame such that an explicit regression imputation model can be fitted. An additional assumption they make is that one knows the current nonrespondents who will become respondents after the pending wave. Armed with this foresight, they derived a measure quantifying the variability in the difference between the two nonresponse-adjusted sample means calculated at the conclusions of waves k and $k + 1$. Essentially, they focus on quantifying $\text{var}(\hat{\bar{y}}_1^k - \hat{\bar{y}}_1^{k+1})$ as opposed to $\text{var}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^k)$, the focus of the methods previously considered.

Their derivation begins by conditioning on the observed data as of the conclusion of wave k, the parameters of an explicit imputation model, and the imputed values of nonrespondents. They reason that the anticipated difference in the two wave-specific sample means is zero and variability a function of how far the

wave $k + 1$ observed values fall from their respective expected values from the imputation model. The specific estimated variance term reported on p. 1016 is

$$\hat{\sigma}_{\varepsilon}^{2}\left(\frac{r_{k+1}}{n}\right)^{2}\left[1+\bar{\mathbf{x}}_{k+1}^{\mathbf{T}}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\bar{\mathbf{x}}_{k+1}+\frac{1}{r_{k+1}}\right] \tag{5.1}$$

where $\hat{\sigma}_{\varepsilon}^{2}$ is the mean squared error (MSE) of the linear regression imputation model fitted as of wave k relating fully-observed covariates to the outcome variable for respondents, n is the overall sample size, r_{k+1} represents the current nonrespondents who will become respondents after the next wave, $\bar{\mathbf{x}}_{k+1}$ denotes the mean covariate vector of these to-be respondents, and $(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$ is the variance-covariance matrix of the imputation model coefficients after having $\hat{\sigma}_{e}^{2}$ factored out front of the expression. The basic premise is that one can declare phase capacity once this variance measure or some function of it (e.g., a relative variance or coefficient of variation) is sufficiently small.

Though a promising improvement, the WR test faces criticism of its own. Its application solely to a continuous outcome variable is restrictive. As Heeringa et al. (2010, p. 149) comment, variables of this type are the exception rather than the rule in applied survey research. Categorical variables are far more prevalent. In fact, outcome variables in the FEVS are exclusively categorical. There is no guidance for imputation procedures other than those utilizing a linear regression model (e.g., hot-deck imputation). It is unclear how or if complex survey data features like weights,

stratification, and clustering would necessitate formulaic modifications. Other limitations are that a phase capacity test may be desired for statistics other than the sample mean, and that it seems unlikely one would have the prescience to know the exact set of sample cases that will respond in the pending data collection wave.

Additionally, the WR test assumes the imputed value for the i^{th} nonrespondent at wave $k + 1$ is the same as the imputed value for that nonrespondent at the conclusion of wave k. Instead, one could argue that the additional wave $k + 1$ responses would be used to refit the imputation model, causing its parameters and, thus, the distribution of plausible values drawn, to change somewhat relative to the imputation model used for the nonrespondents at wave k. Ignoring this additional uncertainty seems injudicious.

Finally, the variance term of the WR test is derived assuming single imputation, yet the simulation and application employ multiple imputation. It seems that modifications to the equation would be warranted when performing multiple imputation as opposed to single imputation.

Let us introduce a simple, fictitious data set to help illuminate these issues and lay the foundation for the new methods to be proposed. Table 5.1 portrays a survey of $n = 10$ sample units in which six responses have been recorded during wave 1, leaving four nonrespondents for which the key continuous survey variable y is to be singly imputed. This is accomplished by exploiting a continuous covariate x known

for the entire sample. The first step is to fit the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ using the wave 1 respondents. The estimated model is $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 5.6388 - 1.0816 * x_i$ with an estimated MSE of $\hat{\sigma}_\varepsilon^2 = 0.3384$. Following Example 4.4 in Little and Rubin (2002), the second step is to derive an imputed value for the i^{th} nonrespondent by finding the expected value of the outcome variable conditional on x_i, and then adding a random, normally distributed residual term in proportion to the square root of the model's estimated MSE. Specifically, imputed values are assigned as $y_i^* = 5.6388 - 1.0816 * x_i + z_i * \sqrt{0.3384}$, where z_i is a random normal variate.

Table 5.1: An Artificial Data Set to Facilitate the Discussion of Prospective Phase Capacity Considerations.

Sample Case ID	Wave	x_i	y_i	Completed Data Set as of Wave 1
1	1	1.1	4.5	4.5
2	1	1.7	3.8	3.8
3	1	2.4	2.8	2.8
4	1	2.8	3.1	3.1
5	1	3.1	1.9	1.9
6	1	4	1.4	1.4
7	2	1.3	?	3.6
8	2	1.9	?	3.4
9	2	2.7	?	2.7
10	2	3.6	?	1.5

At the conclusion of wave 1, the estimated sample mean using the completed data set is unbiased if the MAR assumption behind the imputation model holds. Yet suppose we know with certainty that a pending follow-up effort will produce data on

the four current nonrespondents, and we wanted to use all available information to quantify the uncertainty with respect to what value the nonresponse-adjusted sample mean will take on once that data is observed. In essence, using notation defined previously, the goal is to approximate $\text{var}(\hat{\delta}_1^2) = \text{var}(\hat{\bar{y}}_1^1 - \hat{\bar{y}}_1^2)$.

Generally speaking, however, $\text{var}(\hat{\delta}_k^{k+1})$ carries a subtly different interpretation in the prospective setting as compared to the retrospective setting considered in Chapters 3 and 4. Because we condition on the observed and imputed data at wave *k*, the wave *k* estimate is treated as fixed in the prospective setting, and so the only element of uncertainty is that attributable to plausible values of the future wave estimate. Hence, the variance we refer to in this chapter is not a measure of sampling error *per se*, as was the case in previous chapters. Rather, it is a quantification of the expected squared deviation of the nonresponse-adjusted point estimate after the wave $k + 1$ responses have been obtained relative to the current nonresponse-adjusted point estimate.

Returning to our artificial example from Table 5.1, the WR derivation asserts that $\hat{\delta}_1^2$ simplifies to the difference between the currently imputed and future observed values of the wave 2 respondents. In other words, if we denote an observed value y_i and an imputed value y_i^*, the observed values for wave 1 respondents fall out of the difference $\hat{\delta}_1^2 = \hat{\bar{y}}_1^1 - \hat{\bar{y}}_1^2 = \frac{\sum_{i=1}^{6} y_i + \sum_{i=7}^{10} y_i^*}{10} - \frac{\sum_{i=1}^{6} y_i + \sum_{i=7}^{10} y_i}{10} = \frac{\sum_{i=7}^{10} (y_i^* - y_i)}{10}$, and $\text{var}(\hat{\delta}_1^2)$ is found using equation 5.1.

We concur with Wagner and Raghunathan (2010) that the most logical estimate of the imputed-to-observed deviation of the outcome variable at the sample unit level is $\hat{\sigma}_{\varepsilon}^{2}$, the MSE of the imputation model fitted using the wave 1 respondents. But given all of the available information, we argue that the approximation of $\text{var}(\hat{\delta}_1^2)$ is more straightforward. In this particular instance, it is simply the number of new respondents, r_{k+1}, times the variance of their imputed values, $\hat{\sigma}_{e}^{2}$, divided by the denominator squared, $\left(\frac{1}{n}\right)^2$. Using data from Table 5.1, this would be calculated as $\text{var}(\hat{\delta}_1^2) = \left(\frac{1}{10}\right)^2 * 4 * 0.3384 = 0.0045$.

An alternative method to arrive at this quantity is to simulate a large number of hypothetical wave $k + 1$ data collection processes. While ostensibly unnecessary in the present context, the technique's appeal is that it generalizes to any point estimate and any imputation model. The idea is to impute the wave $k + 1$ values independently many times, say, $R = 1{,}000$, thereby creating a sequence of R hypothetical completed data sets. From each, one calculates a simulated wave $k + 1$ sample mean, which we can denote $\hat{\bar{y}}_1^{(k+1)r}$. If we define $\hat{\bar{y}}_1^{(k+1)\bullet} = \frac{1}{R}\sum_{r=1}^{R} \hat{\bar{y}}_1^{(k+1)r}$ to be the average of the R simulated sample means from wave $k + 1$, then the expected difference to be observed once wave $k + 1$ data is obtained is $E\left(\hat{\delta}_k^{k+1}\right) = \hat{\delta}_k^{(k+1)\bullet} = \hat{\bar{y}}_1^{k} - \hat{\bar{y}}_1^{(k+1)\bullet}$. Although in many instances it is reasonable to anticipate this value to be close to zero, it may not be *exactly* zero, because we condition on the specific set of imputed values drawn to

produce $\hat{\bar{y}}_1^k$, not the expected values of those sample-unit-specific distributions as do Wagner and Raghunathan (2010).

Inferences can be made by forming a prediction interval around the estimated difference. Because $\hat{\bar{y}}_1^k$ is fixed, variability reduces to only the component attributable to the $\hat{\bar{y}}_1^{(k+1)r}$'s. By defining $\text{var}\left(\hat{\delta}_k^{(k+1)\bullet}\right) = \text{var}\left(\hat{\bar{y}}_1^{(k+1)\bullet}\right) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\bar{y}}_1^{(k+1)r} - \hat{\bar{y}}_1^{(k+1)\bullet}\right)^2$, we can construct the interval by finding $\hat{\delta}_k^{(k+1)\bullet} \pm z_{1-\alpha} * \sqrt{\text{var}\left(\hat{\delta}_k^{(k+1)\bullet}\right)}$, where $z_{1-\alpha}$ is the $100(1-\alpha)^{\text{th}}$ percentile of the standard normal distribution.

Let us motivate an example of this simulation procedure using the data in Table 5.1. First, note that $\hat{\bar{y}}_1^1 = 2.87$. To simulate hypothetical values of $\hat{\bar{y}}_1^2$, 10,000 completed data sets were generated using the same explicit imputation model initially fitted using only the wave 1 respondents. The average of these simulated wave 2 means equals $\hat{\bar{y}}_1^{2\bullet} = \frac{1}{10000}\sum_{r=1}^{R=10000}\hat{\bar{y}}_1^{2r} = 2.977$, and so $\hat{\delta}_k^{(k+1)\bullet} = 2.87 - 2.977 = -0.107$. Figure 5.1 illustrates the distribution of the $\hat{\delta}_k^{(k+1)r}$'s. The variance of the simulated mean differences is $\text{var}(\hat{\delta}_1^{2\bullet}) = \frac{1}{10000-1}\sum_{r=1}^{R=1,000}\left(\hat{\bar{y}}_1^{2r} - \hat{\bar{y}}_1^{2\bullet}\right)^2 = 0.0046$, which we can confirm is approximately equivalent to the variance calculated by the closed-form version discussed above (0.0045). If a 95% prediction interval on the expected difference were desired, we have all the necessary inputs to calculate $-0.107 \pm 1.96 * \sqrt{0.0046} = (-0.2399, 0.0259)$. Note that we should not be surprised to find

the interval centered at zero because we have conditioned on the imputed values used to calculate $\hat{\bar{y}}_1^1$.

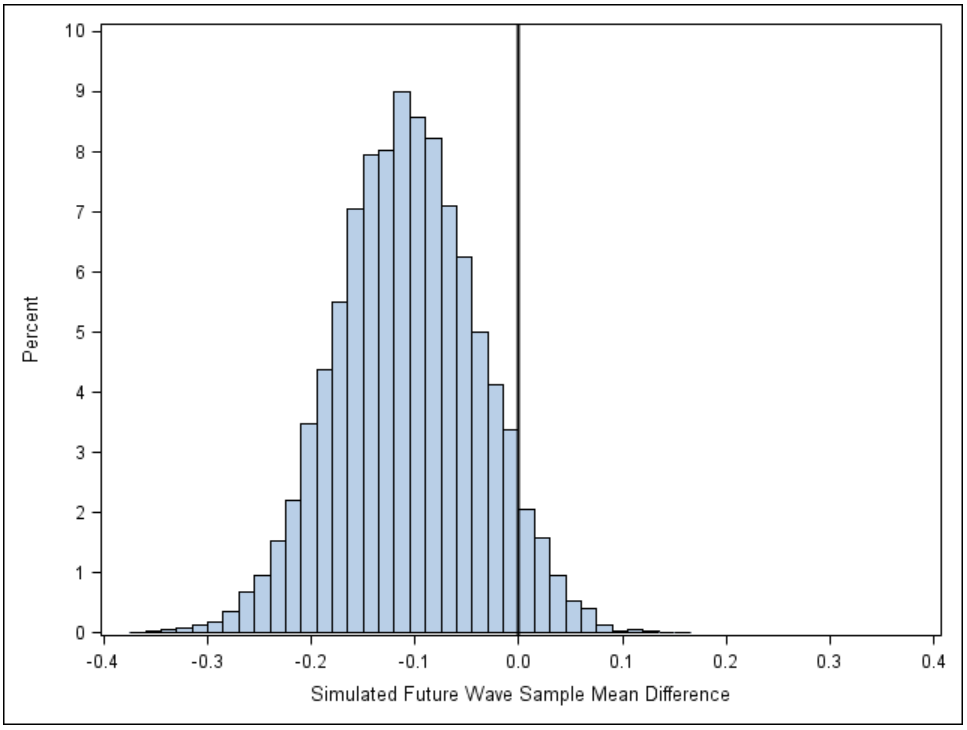

Figure 5.1: Distribution of Simulated Nonresponse-Adjusted Sample Mean Differences after a Second Wave of Data is Collected Using the Artificial Data in Table 5.1.

Alternatively, one can make inferences on the simulated distribution itself by, for example, assigning the 95% prediction interval boundaries using the $100(\alpha/2)^{th}$ and $100(1 - \alpha/2)^{th}$ percentiles of that distribution. In many scenarios, as evidenced by the one just presented, it would not be implausible to assume that the ultimate estimator difference is normally distributed, in which case far fewer than the R = 10,000 replications demonstrated here would be necessary to obtain satisfactorily

precise estimates of its mean and variance for use in the traditional prediction interval formulation.

To justify the supposition made previously that transitioning to a multiple imputation approach would necessitate some kind of formulaic modification, let us return to the same data set portrayed in Table 5.1 and consider two additional methods for simulating the R = 10,000 future wave completed data sets. In the first, suppose the same estimated parameters from the explicit regression model fitted to the observed data is used to multiply-impute the four nonrespondents' missing data M times. That is, each future wave mean, $\hat{\bar{y}}_1^{2r}$, is calculated by applying Rubin's straightforward combination rule to assimilate the 5 completed data set estimates. It turns out that the expectation of $\text{var}(\hat{\delta}_1^{2\bullet})$ for this case can be derived similarly as before, only with an additional $1/M$ term included. Specifically, it is the number of new respondents, r_{k+1}, times the variance of their imputed values, $\hat{\sigma}_e^2$, divided by both the n^2 and M.

Of course, this particular approach would not be "proper" in Rubin's (1987) terminology, since the imputation model parameters are assumed fixed, but it allows for a readily calculable variance to compare against single imputation. The proper approach is to incorporate the imputation model's uncertainty, something we would anticipate introduces more variability. Unfortunately, this makes the variance derivation intractable, yet we can explore the relative difference using the simulation approach. To illustrate, Figure 5.2 plots the distribution of the R = 10,000 simulated

sample mean differences for the original single imputation approach and the two multiple imputation approaches—improper and proper—both using $M = 5$. Notice how all distributions are centered at the same expected value, but the multiple imputation approaches' distributions are somewhat narrower. Recall that the variance of the single imputation approach was found to be 0.0046. The variance of the improper multiple imputation approach is approximately one-fifth of that, or $(1/5)*0.0046 \approx 0.000912$. The variance of the proper multiple imputation approach in this simple example is 0.0031, larger than its improper analog but still less than the single imputation approach. Again, the theoretical derivation of this result, in general, is not easily obtainable.

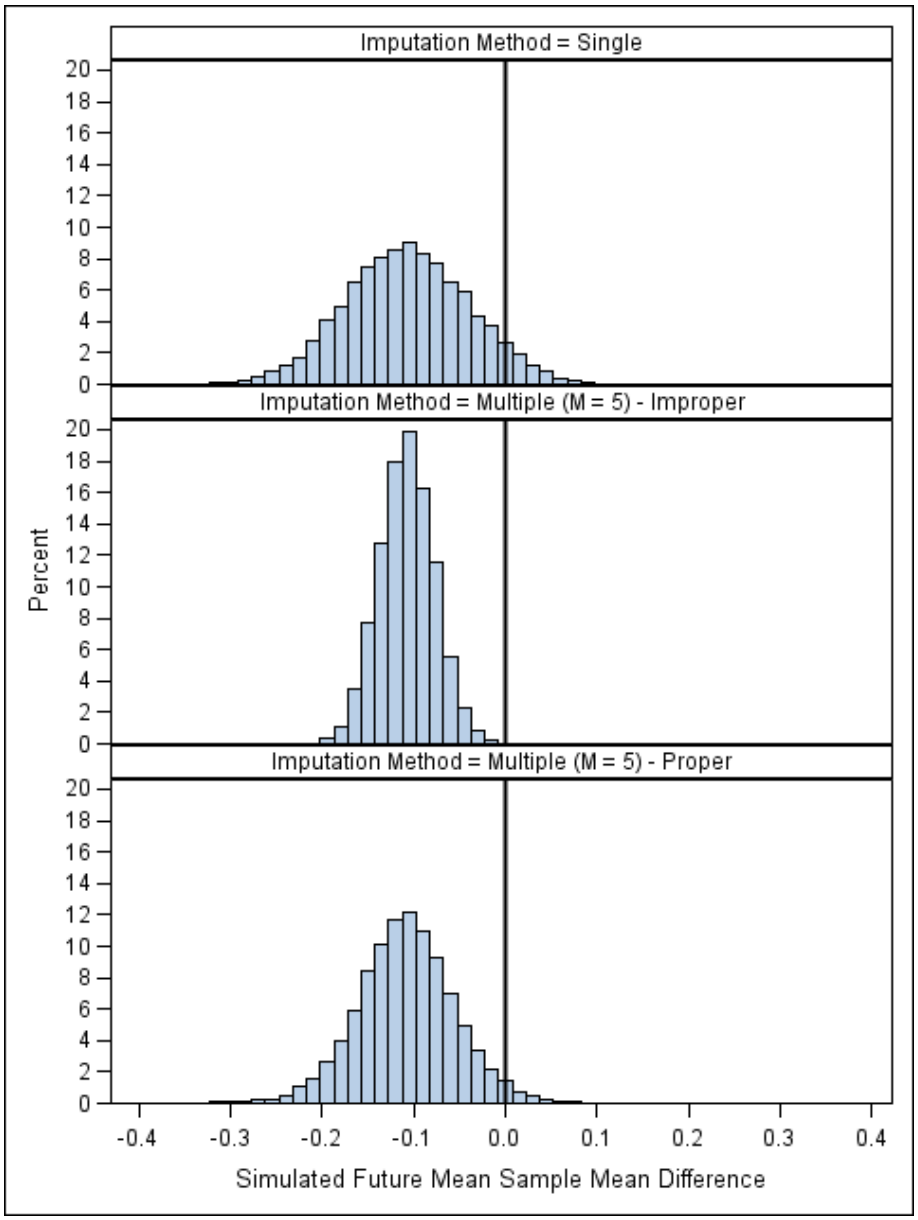

Figure 5.2: Comparative Distributions of Simulated Nonresponse-Adjusted Sample Mean Differences after a Second Wave of Data is Collected Using the Artificial Data in Table 5.1 – Single Imputation, Improper Multiple Imputation, and Proper Multiple Imputation.

The purpose of this introductory section of the chapter was to discuss the fundamental concepts behind prospective considerations of phase capacity, point out

some limitations of the method proposed by Wagner and Raghunathan (2010), and motivate a more general alternative approach in which the imminent wave's data collection process is repeatedly simulated. In the next section we more formally support the procedure with theory and discuss a few ways in which a practitioner could implement it. We also introduce an adaptation for surveys that conduct weighting adjustments for unit nonresponse. Section 5.3 contains results from a simulation study designed to evaluate the performance of the proposed procedure across a diverse set of circumstances, and Section 5.4 reports on an application using data from FEVS 2011.

5.2 New Methods

There are three potential sources of uncertainty inherent in $\text{var}\left(\hat{\delta}_k^{(k+1)\bullet}\right)$, which we will refer to in the most general sense as *events* and label E_1, E_2, and E_3. The first, E_1, is the component associated with which of the current nonrespondents will become respondents in the next wave. The second, E_2, reflects the specific values to be observed for these future respondents. And the third, E_3, symbolizes the resulting impact on the parameters of the imputation model utilized to fill in plausible values for those who have not yet responded at the conclusion of wave $k + 1$.

If we represent all available information (e.g., auxiliary variables, observed response patterns, imputation model parameters) at wave k by $\boldsymbol{\theta}_1^k$, the joint probability distribution we are seeking to make inferences on is $f(E_1, E_2, E_3 \mid \boldsymbol{\theta}_1^k)$. Unless certain dramatically simplifying assumptions are made, a tractable, closed-form expression is

difficult to produce. Note, however, that we can factor the joint probability distribution into a sequence of conditional distributions, since $f(E_1, E_2, E_3 \mid \boldsymbol{\theta}_1^k) = f(E_1 \mid \boldsymbol{\theta}_1^k) f(E_2 \mid \boldsymbol{\theta}_1^k, E_1) f(E_3 \mid \boldsymbol{\theta}_1^k, E_1, E_2)$. The conditional distributions may still prove intractable, but the alternative form intimates how we can pursue a Markov chain Monte Carlo (MCMC) computation approach to approximate the joint probability distribution, which is the spirit of the simulation approach advocated in the previous section. In general, the MCMC approach proceeds as follows:

1. Draw $E_1^* \sim f(E_1 \mid \boldsymbol{\theta}_1^k)$, or simulate who will respond during wave $k + 1$.
2. Using the result from the Step 1, draw $E_2^* \sim f(E_2 \mid \boldsymbol{\theta}_1^k, E_1^*)$, or generate a pseudo-observed value for each simulated respondent based on the same imputation model fitted using respondents as of wave *k*.
3. Using the result from the Step 2, draw $E_3^* \sim f(E_3 \mid \boldsymbol{\theta}_1^k, E_1^*, E_2^*)$. That is, treat the pseudo-observed values from the second step as observed for purposes of (re)fitting same imputation model used to form the completed data set at wave *k*. From this updated model, generate imputed values for those wave *k* nonrespondents not simulated as responding in Step 1.

At the conclusion of Step 3, one simulated wave $k + 1$ completed data set has been created. Therefore, we have one synthetic realization of $f(E_1, E_2, E_3 \mid \boldsymbol{\theta}_1^k)$ from which we can produce one simulated wave $k + 1$ estimate. The fixed value of the estimate at wave *k* is then subtracted to arrive at a simulated difference, or $\hat{\delta}_k^{(k+1)r}$. The

idea is to repeat this entire process independently $r = 1, \ldots, R$ times and base inferences on the resulting distribution.

The illustration provided in the previous section where we assumed all nonrespondents would respond at wave $k + 1$ can be classified as a special case in which there is no variability associated with E_1 or E_3. In effect, the problem simplifies to approximating the distribution $f(E_2 \mid \boldsymbol{\theta}_1^k)$, and so Steps 1 and 3 are unnecessary, as was implicit in the demonstration. A related special case worth mentioning is when one assumes a fixed wave $k + 1$ respondent set, but does not assume it will include all current nonrespondents. In other words, one is interested in quantifying the uncertainty with respect to how much a particular estimate will change given a predetermined set of sample units will respond. In that case, there is no variability associated with Step 1, but one could still repeatedly iterate between Steps 2 and 3 to approximate $f(E_2, E_3 \mid \boldsymbol{\theta}_1^k)$.

While there may be occasions when the goal is to quantify the uncertainty within specific scenarios such as the two just described, there are other plausible methods to simulate the wave $k + 1$ respondent set for each of the R replications. Considering imputation processes generally exploit auxiliary variables on the sample frame, a resourceful approach would be to draw upon some or all of these, perhaps alongside other paradata, to fit a discrete-time hazards model (Allison, 2010), where the ultimate objective is to affix an estimated probability that each current nonrespondent will respond in the pending wave—Wagner and Hubbard (2014)

discuss applications of these models in three example surveys. Given the estimated probability, a stochastic sampling procedure could be implemented in a straightforward manner. As another example, one could independently sample the lesser of the number of wave *k* respondents and the number of nonrespondents remaining. The reasoning behind this approach is the empirically well-documented tendency for the absolute count of respondents to decrease in each subsequent wave within the same design phase. Arguably the most appealing feature of this particular technique is its simplicity.

With minimal modification, the same three-step procedure can be used in settings where weight adjustment techniques are employed to compensate for nonresponse. One obvious difference is that $\hat{\bar{y}}_1^k$ is produced using w_1^k, the set of nonresponse-adjusted weights for the sample units that responded between waves 1 and *k*. There is nothing different about how the hypothetical wave $k + 1$ respondent sets are generated. To generate pseudo-observed values for these simulated respondents, however, one must fit and utilize some form of imputation model. At a minimum, or in the absence of predictive auxiliary variables, a single-class hot-deck routine could be implemented. The third step is to reweight the respondent set defined by the union of wave *k* respondents and the wave $k + 1$ simulated respondents. Using this new set of weights and the pseudo-observed values, one can then formulate an estimate of $\hat{\theta}_1^{(k+1)r}$ and, thus, $\hat{\delta}_k^{(k+1)r} = \hat{\theta}_1^k - \hat{\theta}_1^{(k+1)r}$. The entire process is repeated independently *R* times and inferences can be made using the resulting distribution of the $\hat{\delta}_k^{(k+1)r}$'s.

5.3 Simulation Study

In an effort to evaluate the performance of the proposed technique in a diverse range of scenarios, a simulation study was conducted systematically manipulating the following three factors: (1) the relationship between the outcome variable and response wave; (2) the nonresponse adjustment technique utilized; and (3) the procedure for simulating sets of future wave respondents. In total, the full factorial experimental design consisted of 2 x 3 x 2 = 12 distinct conditions. We first detail the source data and the specific sub-factors whose cross-classification defines the twelve conditions, and then define and report on the specific metrics tracked to assess performance.

As with the simulation study from Chapter 4, the ultimate sets of FEVS 2011 respondents from three example agencies were treated as three fully observed sample data sets. We supposed the key point estimates monitored for these three agencies were the seven percent positives estimates underlying the HCAAF Job Satisfaction index. The first factor manipulated was the method partitioning respondents into one of 10 possible wave cohorts. The same two allocations provided in Table 4.3 were used. In the first, a sample unit's response wave was assigned independently of anything else and in proportion to the FEVS 2011 empirical distribution of Agency 3 first reported in Table 1.2. This simulates a scenario where response timing is unrelated to the outcome variable. In contrast, the second allocation scheme followed the same algorithm described in Section 4.3 in which early respondents were

disproportionately predisposed to have more negative sentiments. While the details can be referenced from the discussion in Section 4.3, recall this was operationalized via a PPS sampling routine in which the measure of size was a respondent-level aggregate measure of positivity based on the seven items comprising the HCAAF Job Satisfaction index.

The second factor manipulated was the compensation technique used to handle unit nonresponse following each simulated wave, or the technique used to produce $\hat{\bar{y}}_1^k$ and $\hat{\bar{y}}_1^{(k+1)r}$ underlying $\hat{\delta}_k^{(k+1)r} = \hat{\bar{y}}_1^k - \hat{\bar{y}}_1^{(k+1)r}$. Two techniques were investigated: multiple imputation (MI) and weighting. The application of the two adjustment techniques was patterned after what was described in Section 3.3. To promote a balanced comparison, the same set of categorical auxiliary variables was exploited in an analogous manner for either technique. Specifically, four variables were used: gender, minority status, supervisory status, and a headquarters vs. field office duty station indicator. For the MI case, the positive/non-positive indicator variable for each of the seven Job Satisfaction index items was imputed $M = 5$ times using a logistic regression model with the aforementioned auxiliary variables serving as main effects. For the weighting case, base weights of respondents at the conclusion of the wave k were raked such that the weighted sum of each variable's categories matched the known population total.

The third factor manipulated was the method by which the wave $k + 1$ respondent sets were simulated. Three conditions were tested. Functioning as a

control of sorts, the first condition was that the future wave respondents were known with certainty. The second condition called for simulating the future wave respondent set by drawing a simple random sample of nonrespondents of size r_k, where r_k represents the number of wave k respondents. An exception was made whenever r_k exceeded the number of nonrespondents remaining after wave k; in that instance, all nonrespondents were simulated as responding in wave $k + 1$. The third condition was to derive a probability of responding in the pending wave by fitting a discrete-time hazards model (Allison, 2010). To be specific, if we collectively symbolize the set of four auxiliary variables identified above as $\mathbf{X_i}$ for the i^{th} sample unit, the following model was fitted:

$$\ln\left(\frac{\phi_{ki}}{1-\phi_{ki}}\right) = \alpha_0 + \alpha_1 k + \boldsymbol{\beta}\mathbf{X_i} \qquad (5.2)$$

where, following the notation from Chapter 2, ϕ_{ki} is the probability (i.e., response propensity) for the i^{th} individual responding during wave k, given the individual has not previously responded. To fit this model, following what is prescribed in Allison (2010), a person-period data set was constructed whereby each sample unit "at risk" of responding during a particular wave has one row of data. While the auxiliary variables comprising $\mathbf{X_i}$ were time-invariant for all records in the person-period data set associated with a particular individual, k was permitted to vary.

Note that the model parameter estimation process encounters a barrier when fitting the model specified in expression 5.2 at the conclusion of wave 1 because, at

that particular threshold, the design matrix columns corresponding to α_0 and $\alpha_1 k$ are both 1 for all rows. A simple work-around used when $k = 1$ was to drop the term $\alpha_1 k$ from the model, in effect reducing the discrete-time hazards model to a standard logistic regression model in which the outcome variable is an indicator of responding in the first wave.

After fitting the discrete-time hazards model, the estimated parameters were used to assign each nonrespondent as of wave k a probability of responding in the pending data collection wave. Denoting this probability $\hat{\phi}_{(k+1)i}$, a random uniform variate r_i between 0 and 1 was generated and the individual was simulated as responding if $r_i < \hat{\phi}_{(k+1)i}$ and not responding otherwise.

Table 5.2 summarizes the three factors and their associated sub-factors manipulated as part of the simulation study.

Table 5.2: Summary of Simulation Factors and Sub-Factors for the Study Evaluating the Newly Proposed Technique for Making Inferences on the Expected Deviation of a Nonresponse-Adjusted Point Estimate Following a Future Data Collection Wave.

Factor 1: Relationship between Response Wave and Outcome

Sub-Factor	Description
1	Wave independent of any outcome variables
2	Earlier respondents less positive than later respondents with respect to HCAAF Job Satisfaction index

Factor 2: Nonresponse Adjustment Technique

Sub-Factor	Description
1	Multiple imputation ($M = 5$)
2	Weighting via raking

Factor 3: Future Wave Respondent Simulation Technique

Sub-Factor	Description
1	Future wave respondents known exactly
2	Random sample of nonrespondents taken, with size equaling the lesser of the number of wave k respondents and the number of nonrespondents remaining
3	Stochastically based on probabilities generated from a discrete-time hazards model

For each of the twelve conditions defined by the cross-classification of the subfactors, $R = 200$ replications were conducted at each of the 9 unique wave thresholds for each of the three agencies' Job Satisfaction index items. In total, 3 agencies x 7 items x 9 wave thresholds = 189 comparisons were made for each of the 12 conditions. In each, $\hat{\delta}_k^{(k+1)\bullet}$ and $\text{var}\left(\hat{\delta}_k^{(k+1)\bullet}\right)$ were found and a corresponding 95% prediction interval was formulated by $\hat{\delta}_k^{(k+1)\bullet} \pm 1.96\sqrt{\text{var}\left(\hat{\delta}_k^{(k+1)\bullet}\right)}$. From there, we determined whether the actual nonresponse-adjusted point estimate calculated once the true set of wave $k + 1$ responses was obtained fell within its boundaries. This is the principal quantity of interest from our perspective, the prediction interval coverage rate with respect to the point estimate difference eventually observed.

Tables 5.3a and 5.3b report these coverage rates observed over 1,000 independent iterations of the twelve conditions. The six conditions reported in Table 5.3a correspond to the scenario in which wave was assigned independently of the outcome, while Table 5.3b reports on the six conditions in which early respondents are systematically more negative.

Unfortunately, the results are far from spectacular. The coverage rates in Table 5.3a show how, even for the condition where the expected value of the outcome variable does change during the data collection period, the prediction intervals contain the true difference between 75 – 80% of the time. There are hardly any noteworthy departures from this marginal rate for a particular agency or survey item. The only condition standing out is the weighting version during which future wave respondents were known with certainty. In all but one instance, its coverage rates exceeded 90%.

Table 5.3b reports coverage rates for the condition where early respondents are more negative in their attitudes. Results for this condition were even poorer than for the first, although more patterns emerge. One interesting finding is that the MI approach exhibits higher coverage rates than the weighting approach. Still low by most standards, the marginal coverage rate for the former is roughly 48%, whereas that figure is around 36% for the latter. Another noteworthy discrepancy is how the coverage rates for Agencies 1 and 3 are much lower than that for Agency 2. We

found this puzzling considering no such differences were found in the first condition; there, all three agencies' marginal coverage rates were very close to one another.

Table 5.3a: Prediction Interval Coverage Rates for the Simulation Study Condition in which Response Wave is Independent of the Outcome Variables.

	MI (M = 5)			*Weighting*		
Future Wave Respondent Simulation Technique	**Known**	**Random**	**DTH Model**	**Known**	**Random**	**DTH Model**
Agency 1						
Item						
4	77.5	78.1	71.1	94.8	75.2	66.7
5	77.2	76.4	74.2	94.8	80.7	71.1
13	77.2	79.4	74.4	91.9	77.8	68.1
63	74.7	80.8	75.3	92.2	78.5	70.0
67	77.8	77.8	72.2	92.2	76.7	69.6
69	78.6	83.1	71.9	94.8	79.3	74.1
70	74.7	80.0	75.3	94.4	77.8	72.6
Agency 2						
Item						
4	74.2	80.8	71.9	92.6	78.5	69.3
5	73.1	80.0	70.6	91.9	76.3	70.0
13	72.8	75.8	73.3	93.7	82.6	72.6
63	77.5	77.8	72.5	94.8	74.1	66.3
67	77.5	81.4	76.9	94.8	75.9	69.6
69	75.6	81.1	75.3	94.1	75.6	65.9
70	73.6	81.9	75.0	91.5	75.9	68.9
Agency 3						
Item						
4	74.4	77.5	75.6	92.2	73.7	65.6
5	72.8	74.2	73.9	90.7	77.4	71.1
13	76.1	74.7	73.6	89.3	75.9	62.6
63	74.2	75.3	73.1	92.2	76.7	68.1
67	76.4	82.2	73.6	93.7	75.9	62.6
69	74.7	77.8	75.6	93.7	78.5	67.8
70	74.4	80.3	73.9	92.2	75.9	68.9

Table 5.3b: Prediction Interval Coverage Rates for the Simulation Study Condition in which Response Wave is Associated with the Outcome Variables.

	MI (M = 5)			*Weighting*		
Future Wave Respondent Simulation Technique	**Known**	**Random**	**DTH Model**	**Known**	**Random**	**DTH Model**
Agency 1						
Item						
4	34.2	35.6	27.8	24.8	15.9	10.4
5	43.9	45.3	38.3	35.9	27.8	18.1
13	46.9	50.0	45.3	45.9	33.7	25.2
63	28.6	27.2	21.7	16.7	10.4	5.6
67	26.1	27.8	22.8	17.8	10.4	5.2
69	24.7	31.1	21.1	19.3	13.7	6.3
70	45.0	44.4	40.0	32.6	23.0	14.4
Agency 2						
Item						
4	73.6	72.5	66.9	84.4	59.6	51.1
5	72.5	76.1	68.9	87.0	65.2	54.8
13	73.9	76.4	68.1	91.1	73.0	62.6
63	65.0	65.6	66.9	72.6	49.6	43.3
67	63.9	69.2	64.7	69.6	48.5	40.4
69	69.7	71.1	66.9	81.1	56.7	47.4
70	77.2	75.0	65.3	84.4	62.6	55.9
Agency 3						
Item						
4	32.8	35.8	28.1	27.0	14.4	11.5
5	40.0	40.6	38.9	36.7	20.7	13.7
13	48.1	50.8	39.4	48.1	29.6	21.9
63	21.9	25.3	21.7	16.7	11.5	6.7
67	24.7	25.6	22.5	17.0	11.5	7.8
69	28.1	33.6	24.7	23.7	12.6	9.3
70	41.4	46.9	38.1	35.2	21.1	15.9

One performance dimension of interest masked by the presentations of Tables 5.3a and 5.3b is the trend along the progression of data collection wave thresholds. Figures 5.3a and 5.3b illustrate how the prediction interval widths decrease over time. In these figures, the length of the vertical bars represents the prediction interval widths constructed about $\hat{\delta}_k^{(k+1)\bullet}$, using Item 4 on the survey for the first iteration of Agency 3 as an example. Figure 5.3a reports on the condition where the response wave and the outcome are independent, and Figure 5.3b reports on the condition for which a relationship was embedded. A separate panel is given for each of the six permutations of a nonresponse adjustment approach and future wave respondent simulation technique. The overlaid 'X' symbolizes the actual difference observed once the true wave $k + 1$ respondent set was incorporated.

The proclivity of prediction interval widths to shrink over the simulated data collection period of a design phase is intuitive considering there are steadily diminishing counts of (actual and simulated) new respondents contributing to the change in the percent positive estimate. Also intuitive from Figure 5.3a, in particular, is how the intervals and observed differences gently oscillate at random about a null difference, owing to the fact that this figure report on the simulated condition where there was no change in the expected value of the outcome over time. On the other hand, for the condition where early respondents exude more negative attitudes, Figure 5.3b illustrates how the actual point estimate difference often falls outside the bounds of the prediction interval, particularly in the early waves. It is evident that the temporal patterns in the expected value of this particular variable are not fully

captured by the four covariates employed in the two nonresponse-adjustment procedures investigated.

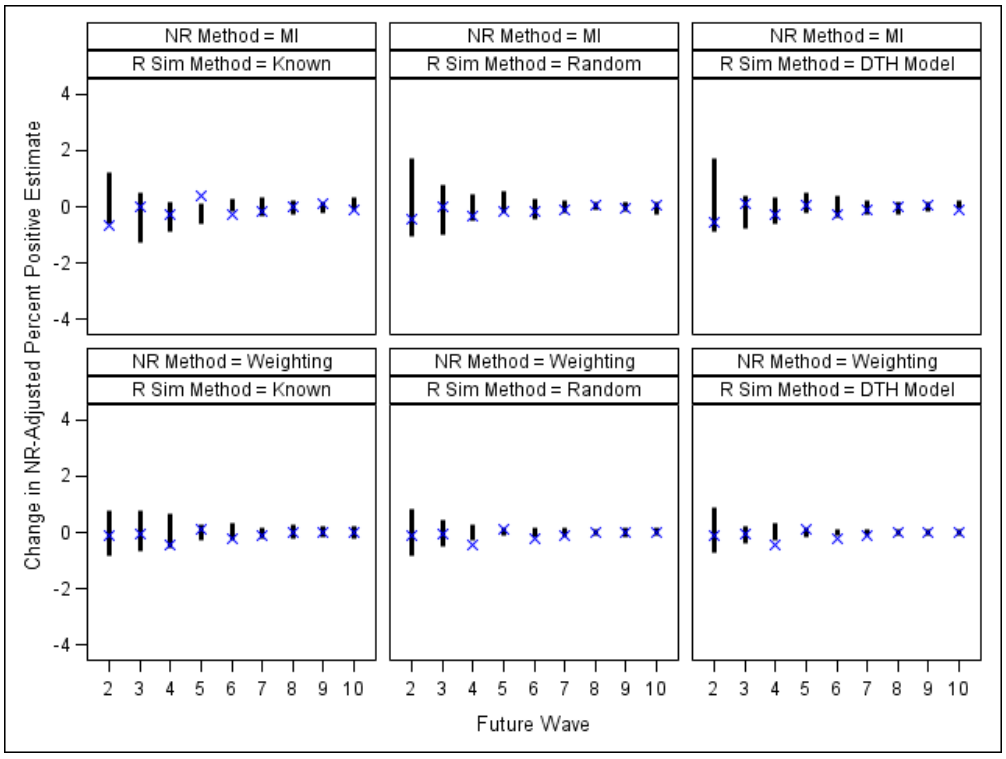

Figure 5.3a: Prediction Intervals Overlaid with Actual Nonresponse-Adjusted Sample Mean Differences Observed for the First Iteration of the Simulation Condition in which the Response Wave is Independent of the Outcome Variables – Using FEVS Item 4 for Agency 3 as an Example.

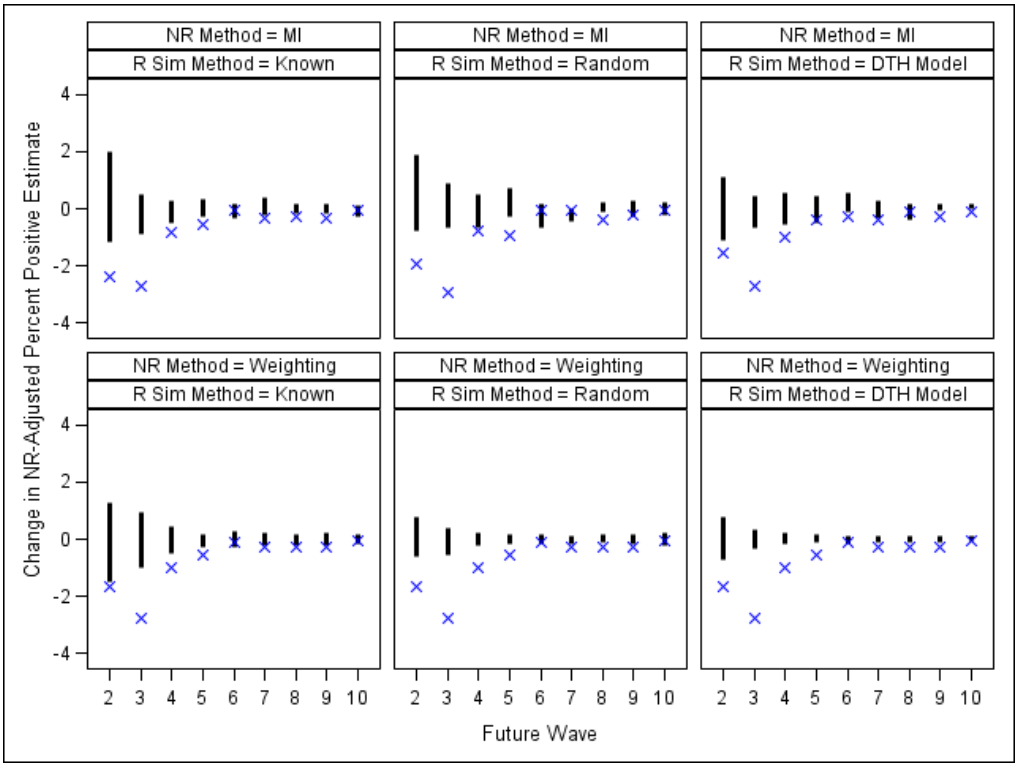

Figure 5.3b: Prediction Intervals Overlaid with Actual Nonresponse-Adjusted Sample Mean Differences Observed for the First Iteration of the Simulation Condition in which the Response Wave is Independent of the Outcome Variables – Using FEVS Item 4 for Agency 3 as an Example.

Figures 5.4a and 5.4b investigate wave-specific coverage trends from a somewhat broader perspective. Rather than focusing on one item, Figure 5.4a plots the average coverage rate trend for all seven Job Satisfaction index items for each of the six specific conditions in which the MI approach was used to produce a prediction interval. The agency-specific trends are broken out within each panel. There are no discernable trends for any future wave respondent simulation technique when the outcome is independent of when individuals respond, but for the condition where there is a relationship between those two factors, the coverage rates drop off after the

first wave threshold, but gradually climb across the remaining wave thresholds. For completeness, Figure 5.4b illustrates the comparable trends lines for the weighting version's six conditions, but the takeaway messages are generally the same.

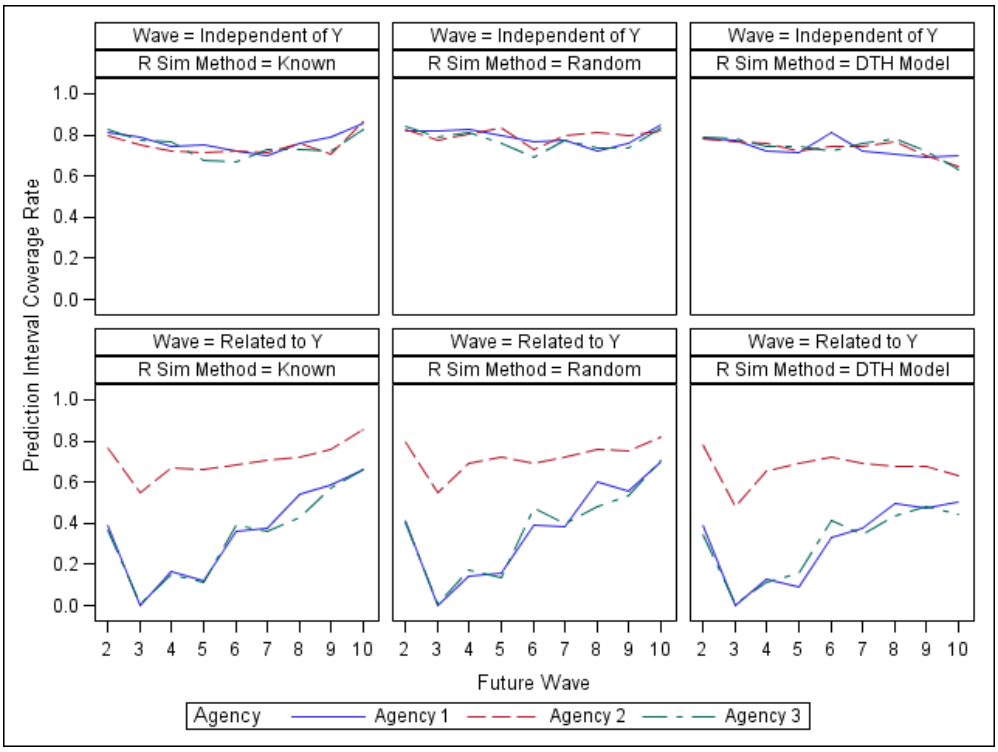

Figure 5.4a: Wave-Specific Prediction Interval Coverage Rates for the MI Method, Averaged over the Agency's Seven FEVS Items Investigated for all Six Sub-Conditions of the Simulation.

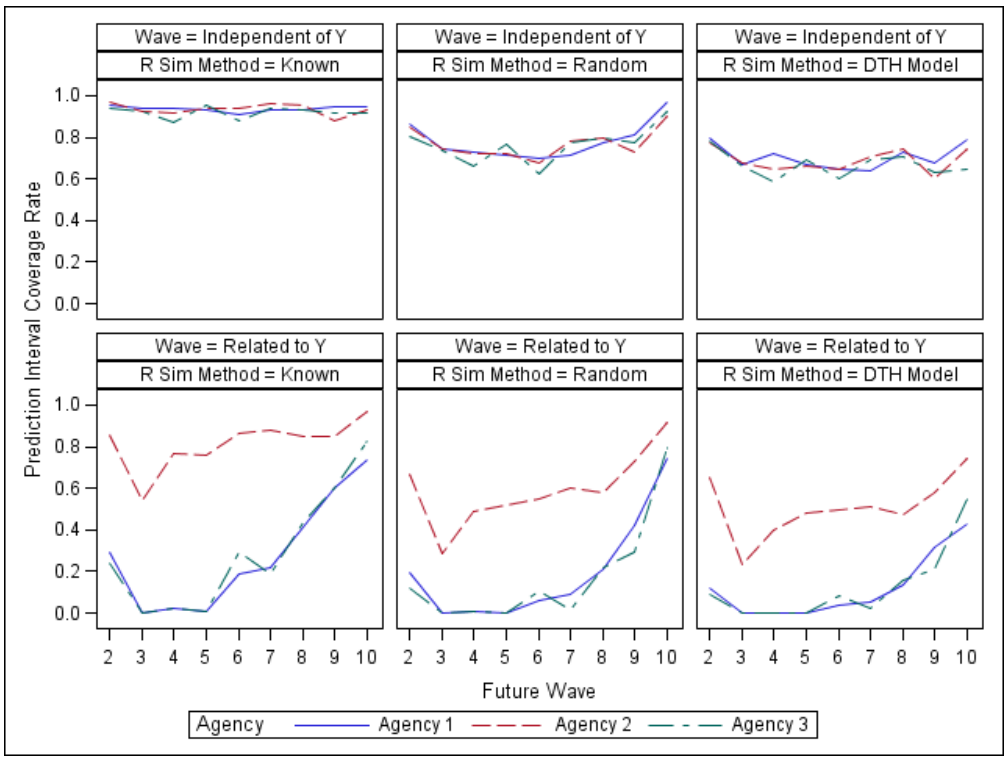

Figure 5.4b: Wave-Specific Prediction Interval Coverage Rates for the Weighting Method, Averaged over the Agency's Seven FEVS Items Investigated, for all Six Sub-Conditions of the Simulation.

5.4 Application to the Federal Employee Viewpoint Survey

In this section we discuss an application of these methods in a real-world survey. As in the simulation study, we use data from the 2011 FEVS and focus on the seven items comprising the HCAAF Job Satisfaction index; however, instead of randomly assigning each respondent's data collection wave, we use the actual response patterns observed in the survey's administration. Since Agency 1's data can be partitioned into 9 waves, Agency 2's data into 8 waves, and Agency 3's data into 10 waves, there were a total of 8 + 7 + 9 = 24 unique wave thresholds at which

differences in the seven nonresponse-adjusted percent positive estimates could be evaluated. For each comparison, the same three future wave respondent simulation techniques detailed in the previous section were investigated independently for each of the same two nonresponse adjustment procedures—namely, multiple imputation (M = 5) and weighting. These procedures were carried out in the same manner described for the simulation study, using the same set of auxiliary variables. In all, 24 x 7 = 168 distinct prediction intervals were formed for the anticipated point estimate difference to be observed in each of 3 x 2 = 6 unique combinations of (1) the future wave respondent simulation technique and (2) the nonresponse compensation procedure. As in the simulation, R = 200 iterations was deemed sufficient to approximate the $\text{var}\left(\hat{\delta}_k^{(k+1)\bullet}\right)$ term in the prediction -interval $\hat{\delta}_k^{(k+1)\bullet} \pm 1.96\sqrt{\text{var}\left(\hat{\delta}_k^{(k+1)\bullet}\right)}$.

Table 5.4 reports the agency- and item-specific prediction interval coverage rates in each setting, rates that are averaged across all of the given agency's wave thresholds. For example, a figure of 87.5 implies 87.5%, or 7 out of 8 of that item's prediction intervals encapsulated the difference ultimately observed. Because coverage rates vary so widely, it is difficult to make mention of any prevailing trends and patterns. There is some evidence that the techniques work better in Agency 2, which is the notably smaller (n = 1,057) than the other two agencies (n = 16,565 and n = 17,177, respectively).

Table 5.4: Agency- and Item-Specific Prediction Interval Coverage Rates across All Applicable Wave Thresholds.

Future Wave Respondent Simulation Technique	*MI (M = 5)*			*Weighting*		
	Known	**Random**	**DTH Model**	**Known**	**Random**	**DTH Model**
Agency 1						
Item						
4	37.5	75.0	37.5	50.0	75.0	37.5
5	75.0	37.5	87.5	100.0	100.0	100.0
13	62.5	75.0	62.5	50.0	62.5	50.0
63	50.0	50.0	37.5	62.5	87.5	37.5
67	25.0	62.5	50.0	12.5	37.5	12.5
69	50.0	50.0	62.5	50.0	62.5	50.0
70	62.5	25.0	62.5	75.0	87.5	62.5
Agency 2						
Item						
4	71.4	57.1	71.4	85.7	85.7	100.0
5	71.4	71.4	28.6	100.0	100.0	85.7
13	57.1	85.7	57.1	100.0	100.0	85.7
63	42.9	100.0	57.1	85.7	85.7	85.7
67	42.9	57.1	71.4	85.7	85.7	85.7
69	71.4	14.3	71.4	100.0	85.7	71.4
70	100.0	85.7	71.4	100.0	100.0	100.0
Agency 3						
Item						
4	44.4	77.8	66.7	44.4	66.7	44.4
5	66.7	88.9	22.2	66.7	66.7	44.4
13	66.7	77.8	66.7	33.3	44.4	44.4
63	44.4	77.8	66.7	100.0	100.0	88.9
67	55.6	66.7	66.7	77.8	55.6	66.7
69	88.9	66.7	55.6	66.7	88.9	100.0
70	88.9	66.7	55.6	77.8	77.8	66.7

One possible manifestation of Agency 2's smaller size impacting coverage rates is that it tends to produce a larger value of $\text{var}\left(\hat{\delta}_k^{(k+1)\bullet}\right)$. This is evident by visually

contrasting the length of the vertical bars in Figures 5.5a, 5.5b, and 5.5c against one another, as that length reflect the agencies' wave-specific prediction intervals for each of the six unique combinations of a particular nonresponse-adjustment procedure and future wave respondent simulation technique, using FEVS Item 4 as an example. Like the analogous plots provided in the previous section, the 'X' marks actual difference. Note that the *y*-axis minima and maxima for Agency 2 are slightly larger in magnitude than those for the Agencies 1 and 3. Even so, as judged by the vertical distance of the bars, the prediction intervals are still larger. At the same time, Agency 2's relative magnitudes of actual differences are no greater or less, on average, than the other two agencies.

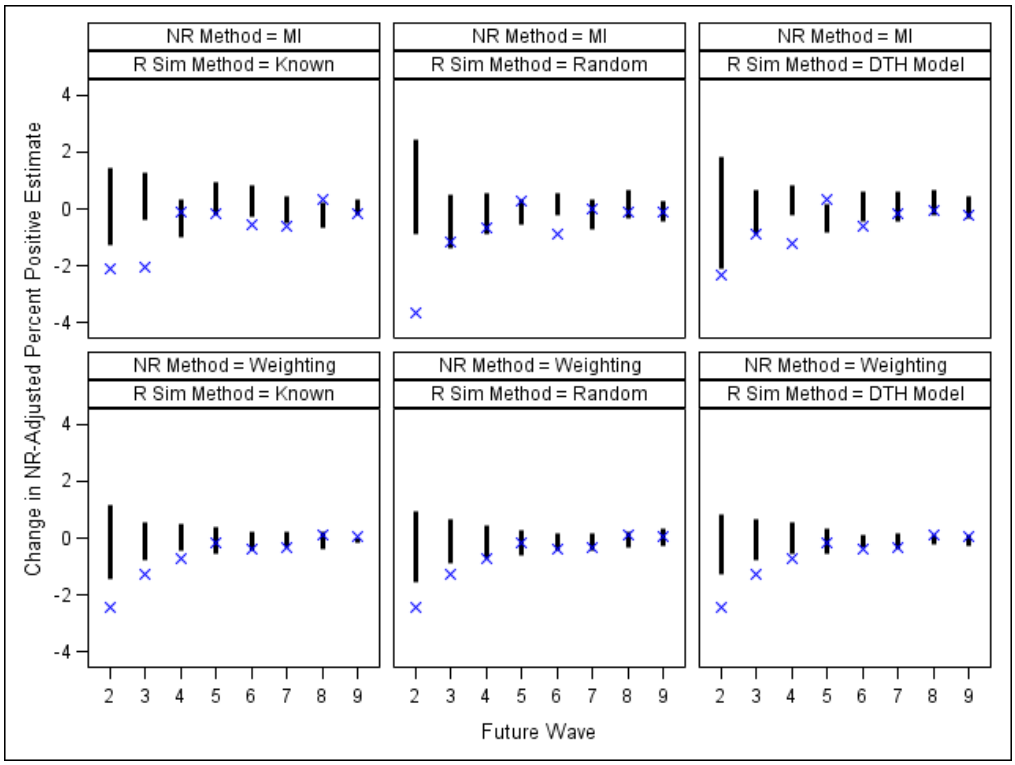

Figure 5.5a: Prediction Intervals Overlaid with Actual Nonresponse-Adjusted Sample Mean Differences Observed for the FEVS 2011 Application – Item 4 for Agency 1.

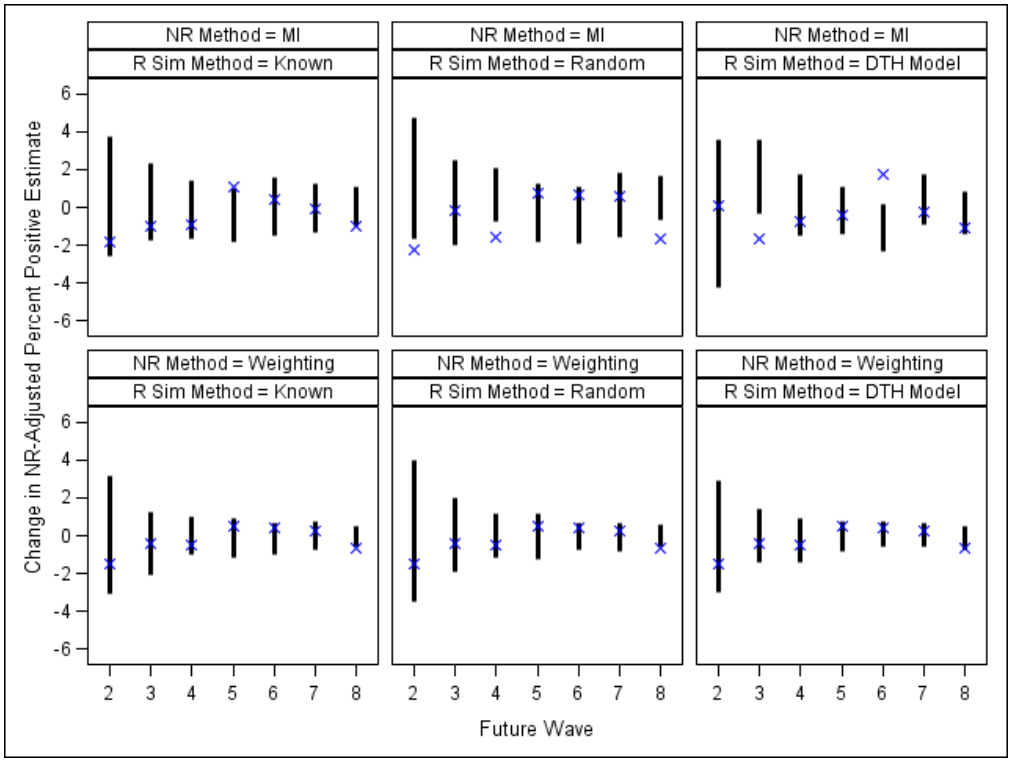

Figure 5.5b: Prediction Intervals Overlaid with Actual Nonresponse-Adjusted Sample Mean Differences Observed for the FEVS 2011 Application – Item 4 for Agency 2.

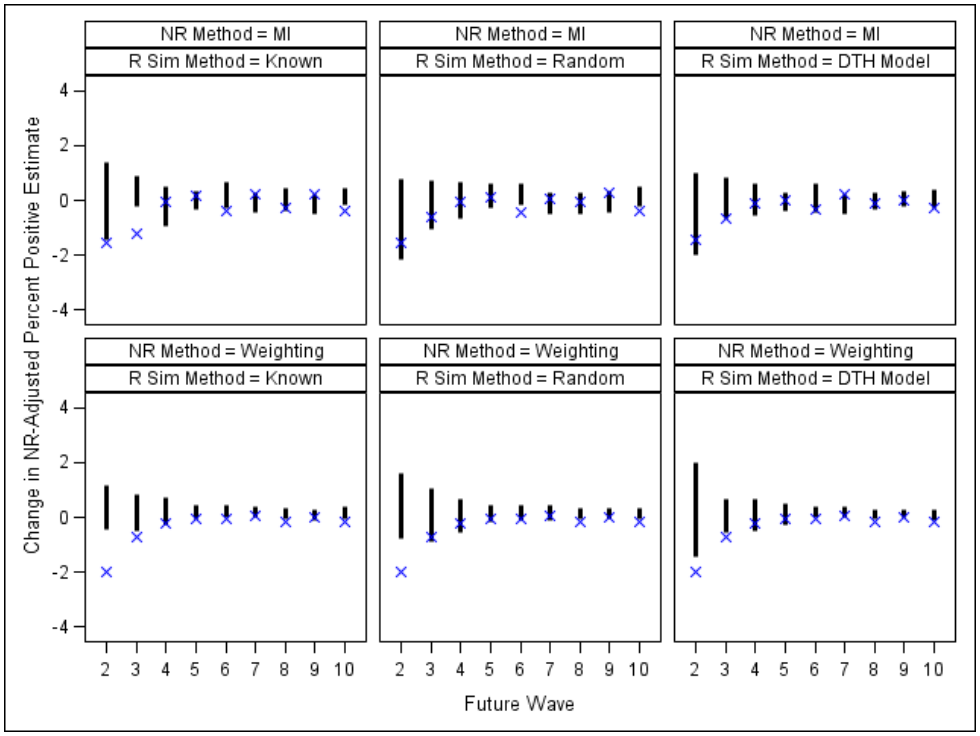

Figure 5.5c: Prediction Intervals Overlaid with Actual Nonresponse-Adjusted Sample Mean Differences Observed for the FEVS 2011 Application – Item 4 for Agency 3.

5.5 Conclusion

After pointing out some of the limiting factors of the prospective variance formula proffered by Wagner and Raghunathan (2010), we introduced a more general MCMC procedure that repeatedly simulates the pending wave data collection process in a sequence of three steps, sometimes fewer depending upon certain assumptions one might be comfortable making. The first step is to simulate which of the current nonrespondents will respond during wave $k + 1$. The second step is to use an imputation model to fill in a plausible value for those tapped to respond. For those not tapped to respond, the third step is to re-administer the nonresponse adjustment

process after updating the underlying model with the pseudo respondents and their plausible values. We discussed the mechanics if one were to use multiple imputation, but also an adaptation for when one were using a weighting approach instead. After completion of the third step, one simulated value of a nonresponse-adjusted point estimate to be observed following wave $k + 1$ can be formulated, and the original nonresponse-adjusted point estimate can be subtracted to get one plausible point estimate difference. The idea is to independently repeat this procedure R times and base inferences on the resulting distribution.

To assess the general performance of the approach, a simulation study was undertaken using the seven items comprising the 2011 FEVS HCAAF Job Satisfaction index. For manageability, we restricted focus to the same three anonymous agencies investigated as part of the first two studies. The simulation consisted of 12 unique conditions defined by the cross-classification of three methods for simulating the future respondent set, whether the multiple imputation or weighting adjustments were the nonresponse compensation technique of choice, and whether or not there was a relationship between the response wave and the outcome variable's expected value. In spite of the promise of the newly proposed method, results were lackluster. Naturally, we found this to be deflating. The performance metric we considered crucial for its endorsement was the coverage rate of prediction intervals constructed at wave thresholds. In only one setting were these rates reliably greater than 90%, that in which the weighting variant was used and the identities of future respondents were known with certainty (i.e., who they were, not what their responses

would be). An application using the actual response patterns observed in the 2011 FEVS did not fare much better. The one exception was Agency 2, whose prediction intervals widths were notably larger than those of Agencies 1 and 3. Improved coverage from a larger prediction interval, all else equal, is intuitive. After all, the wider the net cast, the more likely the true difference will be captured.

Despite the numerous factors systematically modified in the simulation study, there are still additional factors worthy of exploration in future simulation studies. One is the size of M for the MI version of the technique. Earlier, we remarked that the variability of the simulated mean differences shrunk in proportion to $1/M$, albeit for the "improper" example. It seems possible that the increased precision associated with a larger M, say, $M = 100$, could have an effect on results. Another potential factor is the ability of auxiliary variables to control the deviations in nonresponse-adjusted estimates over the data collection period. In this study, while there were advantages of using real survey outcomes and auxiliary variables, the fact of the matter is that the imputation and weighting approaches did little to curb the upward mobility exhibited by estimates using the accumulating data. Systematically varying whether or not that kind of movement can be accounted for by a nonresponse adjustment procedure would be interesting. That said, we found that coverage rates were only around 75 – 80% for the simulation condition where response wave was assigned independently of all else. In that particular case, the point estimates' movement was technically bridled, since there were only minor, random fluctuations. Another potentially fruitful extension would be to account for additional sources of

variability. For instance, in the discrete-time hazards modelling approach to simulate the future wave respondent set, we stochastically sampled from the nonrespondent pool using the same estimated future-wave response propensities in all *R* replications. A bootstrap step (Efron and Tibshirani, 1993) could have been embedded to reflect the uncertainty of the given estimated propensity.

Chapter 6: Discussion

6.1 Dissertation Summary

According to Biemer and Lyberg (2003), a tenet of overall survey quality is timeliness, and a key driver of a survey's timetable is the data collection period. Invariably, not all sample units respond in the first recruitment attempt, and a sequence of follow-ups in the form or reminder mailings, phone calls, or in-person visits typically ensues. Some survey sponsors sanction this process to continue indefinitely in pursuit of a target response rate or minimum respondent count, with the tacit assumption that the magnitude of nonresponse error decreases with each additional wave of data collected. In Chapter 2, we illustrated how that assumption can be false, both theoretically and via the analogy of a partitioned water tank, not to mention the widespread empirical findings in the nonresponse literature (Merkle and Edelman, 2002; Groves and Peytcheva, 2008) suggesting that the (non)response rate is only weakly associated with nonresponse error.

Bearing these issues in mind, Groves and Heeringa (2006) encourage practitioners to employ paradata and other real-time evaluations to inform when to cease data collection or, more generally, when to segue into a different design phase. They defined the notion of phase capacity being the point during a fixed design phase when estimates stabilized. A critical element absent in their exposition, however, is a well-defined, calculable rule practitioners can follow to determine whether phase capacity has occurred. The aim of this dissertation was to fill that void. Over the span of three methodological studies involving simulations and an application using

data from the 2011 FEVS, several specific tests for phase capacity were proposed and their performance assessed.

In the remainder of this section, we briefly recapitulate the essence and motivation behind the methods proposed in each study. In the section that follows, we take a step back and discuss the limitations of the research undertaken as part of this dissertation from a broader perspective, identifying several worthwhile avenues for further research.

To be fair, the origins of testing for phase capacity had appeared in the literature prior to this dissertation, albeit not exclusively motivated by the ideas conveyed in Groves and Heeringa (2006), and with ample room for improvement. To our knowledge, Rao, Glickman, and Glynn (2008) offered the first such contribution. While they evaluated several methods, they concluded that their third "stopping rule" performed best, yet a significant limitation is that it supposes auxiliary variables are employed to multiply-impute the missing data caused by unit nonresponse. Even when auxiliary variables are available, many surveys prefer to adjust the base weights of respondents to compensate for nonresponse. To that end, the first study reported in Chapter 3 proposed a variant operating similarly in spirit to Rao et al.'s third rule, but amenable to surveys that conduct weight adjustment methods in lieu of multiple imputation. Through several simulated data collection scenarios and an application using a real-world survey data set, the 2011 FEVS, the two tests' performance were compared. All else equal, we found the weighting variant to be more sensitive to

estimate changes and, thus, less likely to make the phase capacity declaration. By indirectly teasing apart all components of the estimated variance of the adjacent wave sample mean difference, we discovered that the covariance accounting for the shared data up through wave k -1 was handled differently in the multiple imputation version. In general, the reduction of the overall variance term after accounting for the covariance was much more drastic in the weighting version.

A limitation of the tests described in the first study is that they are univariate by design, meaning they focus on only one point estimate at a time. It is not immediately obvious how one would proceed if the test were conducted on two or more point estimates and contradictory conclusions resulted. The purpose of the second study was to adapt concepts of the weighting technique proposed in the first study into a multivariate technique permitting the practitioner to make a single yes-or-no determination of phase capacity for a battery of D point estimates simultaneously. Two methods were outlined for comparison. The first took the form of a Wald chi-square test statistic in a straightforward multivariate extension of the weighting variant discussed in the first study. The second was an adaptation of a method commonly used in longitudinal analysis to measure whether there a change has occurred over some timespan. We referred to this approach as the non-zero trajectory method. Both methods were able to detect phase capacity quickly and without any noteworthy residual nonresponse error when the expected value of the outcome was stable over the data collection period. All else equal, however, the non-zero trajectory method tends to determine phase capacity later than the Wald chi-square

approach, given it requires a minimum of four waves of data. When there is a trend in the expected value of an outcome variable over the data collection period, a trend that cannot be corrected for by some form of nonresponse adjustment procedure, the additional waves dictated by the non-zero trajectory method prove advantageous.

The methods proposed in the third and final study were not tests or rules *per se*; they were ways to quantify the uncertainty with respect to how much an estimate is expected to deviate from its current value once the pending wave of data is collection. This took the form of what we referred to as a prediction interval. It builds upon ideas of Wagner and Raghunathan (2010), who approached the task of detecting phase capacity from a prospective stance. Rather than determining whether the most recent wave of data collection substantively altered a key point estimate, they attempted to quantify the likelihood of phase capacity being concluded after a pending wave. Several limitations to their method were noted, and a more widely applicable, three-step MCMC simulation procedure was proposed. Regrettably, results were not great. The key quantity investigated in the simulation study and 2011 FEVS application was the coverage rate of the prediction intervals, meaning the portion of the time the actual nonresponse-adjusted point estimate was contained within the interval constructed. Over a variety of simulation conditions, even some in which the response timing was independent of everything else, coverage rates were generally well below any satisfactory level.

6.2 Limitations and Ideas for Further Research

For several reasons, some political in nature, the actual adoption of a phase capacity testing approach to guide the FEVS data collection process would face headwinds. At the forefront of resistance is the OPM survey administration team's dogma that each agency be treated equitably and abide by a common set of rules and restrictions. While the team tries to remain open to the each agency's unique needs and objectives for conducting the survey, to avoid any perception of favoritism and to facilitate the unwieldy process of emailing several million survey invitations and reminders during the field period, certain leniencies and flexibilities once offered had to be curtailed in recent years. For example, in the 2011 FEVS and administrations prior, agencies were given generous amounts of leeway with respect to the length and timing of their field period. As the survey's sample size continued to grow, however, accommodating these agency-specific requests became increasingly challenging. Consequently, beginning with the 2012 FEVS, the field period for all agencies was preset at six weeks, with each agency choosing from one of two possible start dates that are one week apart.

Confirmation of phase capacity for a portion of the participating agencies, if leading to an abridged data collection period, would introduce efficiencies for certain aspects of the survey cycle, such as the survey support center operation, which only provides assistance for individuals in those agencies for which the survey is still open. In all honesty, however, these efficiencies doubtfully constitute sufficient grounds to convince and secure buy-in from agency stakeholders. As evidenced by the trend in

Figure 1.1, a trend seen for most items and across almost all agencies, the tendency is for point estimates to increase over the course of data collection. As yet, this tendency cannot be extirpated by any kind of nonresponse adjustment procedure. The agency stakeholders alluded to are human resources managers tasked with more than just liaising with the OPM survey administration team on logistical aspects of the survey. They are charged with analyzing the results and developing action plans to drive organizational change and improve employee morale, with the end goals of boosting productivity and improving the overall quality of work output. More and more frequently, despite the FEVS not having been designed for this purpose, the success of these their efforts is measured by future years' FEVS estimates. From these stakeholders' perspective, the higher the point estimates, the better. Therefore, there will be opposition to any tactic, shortening the data collection period included, believed to result in lower point estimates, even if only by a statistically undetectable amount.

The methods presented over the course of this dissertation are not applicable to all types of surveys. Because nonresponse adjustments must be conducted in real-time, or at least periodically during the field period, surveys that do not collect data electronically in a more or less instantaneous manner might be precluded. For example, it may prove cumbersome testing for phase capacity in a survey for which the key point estimate is derived from survey staff categorizing an open-ended question or from a self-administered survey instrument that is not machine-readable.

Another working assumption thus far unexpressed is that the entire sample, or some germane subset (e.g., individuals within a particular agency), is "active," meaning all sample units are contacted for participation at the same time. This may be impractical for some surveys, such as an in-person household survey covering a vast geographical expanse with a sample listing taking weeks or months to exhaust. A related scenario is when a sample is partitioned into subsamples, perhaps for periodic release into the field. For example, Parsons et al. (2014) discuss how the National Health Interview Survey yearly sample is allocated into four marginally representative *panels* as "a contingency to handle potential budget cuts" (p. 16). Research investigating the feasibility of testing for phase capacity when the totality of sample units is not contemporaneously being contacted to participate could shed light on which situations permit direct application of these methods and which should be avoided.

There are also settings where the entire sample is active, but initial invitations and reminders do not occur at precisely the same time. A sensible adaptation to address this circumstance is to redefine a data collection wave using some alternative temporal demarcation. For example, while a wave of data was defined in this dissertation as the set of responses obtained between two reminders, one could instead define a wave as data collected between predetermined calendar days, days which need not necessarily be spaced equally apart.

In spite of our aversion to the phrase "stopping rule," arguing previously that determining phase capacity does not necessarily imply data collection should be terminated altogether, only that a new design phase is warranted, a major limitation of this dissertation is that the sole design phase transition examined is, in fact, terminating the nonrespondent follow-up process. More research is needed to understand how these techniques perform under alternative design phase changes, particularly switching data collection modes. One fitting data source for studies of this ilk would be the American Community Survey (ACS), which follows up with nonrespondents using a sequential mixed-mode design in the following order: self-administered Internet, self-administered mail, computer-assisted telephone interviewing (CATI), and computer-assisted personal interviewing (CAPI). The sequence is designed such that data are collected using the least expensive method first and followed by progressively more expensive modes. As described in Chapter 7 of Torrieri (2014), the yearly ACS sample is divided into independent monthly samples, and each is allotted three months for data collection efforts, one month for each of the Internet/mail, CATI, and CAPI stages. Figure 7-1 of Torrieri (2014) is a nice diagram illustrating the chronology and overlap of the stages with respect to the monthly samples. While the monthly allocation scheme assuredly facilitates logistical aspects of the data collection process, further research exploring an adaptive transitioning methodology founded on concepts of phase capacity testing might introduce additional cost-saving efficiencies.

Indeed, analyses on the data collection cost savings, if any, attributable to adopting a phase capacity testing strategy are urgently needed. A formidable hindrance to that occurring, however, is the proprietary nature of much of that data. This is especially true in the United States, where many of the nationally representative surveys disseminating data to the general public free of charge are sponsored by federal government agencies that typically award contracts to private research organizations to handle data collection for the survey. These private firms surely maintain and scrutinize detailed cost information from current and recently completed survey projects for budgeting purposes and to help arrive at the bidding price of a proposal. From these firms' perspective, however, there is concern disclosing this information could lead to it being used against them in some way, perhaps by a competitor.

Despite the paucity of detailed cost information in large-scale surveys and the fact that the incremental per-complete cost in an exclusively Web-based survey such as the FEVS is marginal, there are untapped avenues to indirectly measure cost savings. For example, one of the FEVS sample frame variables is the employee's annual salary. Considering that the survey takes approximately 20 minutes to complete, one could multiply the sampled employee's salary by [1/(2080 work hours in a year] x [1/(3 twenty-minute intervals in an hour)] = 1/6240 to get a crude measure of the opportunity cost associated with taking time away from one's official duties to fill out the questionnaire. Sample unit opportunity costs could be aggregated

in a variety of meaningful ways to provide insight as to whether the responsive design approach under consideration genuinely reaps cost savings.

With respect to prospective considerations of phase capacity, one anticipated extension is the desire to make inferences on the expected change in a point estimate following wave $k + 2$ or beyond. To tackle that particular problem, a more straightforward approach might be to draw upon time series analysis methods (Hamilton, 1994). For instance, econometricians make routine use of the economic indicators (e.g., unemployment rate, jobless claims) estimated from repeated survey efforts such as the Current Population Survey to generate forecasts of the future value of those estimates. It seems reasonable that those methods could be tailored to generate forecasts within the arena of a single survey effort's data collection period.

Another interesting application of these methods would be in a survey with two or more disparate stages of data collection, such as a survey where the respondent provides partial information that gets supplemented with information acquired from administrative records. This strategy is common in surveys charged with capturing highly technical or exceptionally detailed information the typical respondent is unable to readily recall with satisfactory precision. For example, the National Immunization Survey obtains general information about an age-eligible child from a telephone interview with the child's parent or guardian, but the detailed vaccination history is obtained in a subsequent data collection stage from the child's medical provider(s). As another example, the Residential Energy Consumption Survey commences with

an onsite interview during which the head of the household provides basic information about the housing structure and its gas, electricity, and heating and air conditioning equipment, but the critical measures of energy consumption and expenditures are obtained later upon following up with the energy supplier(s). One could certainly test for phase capacity in one or both stages. The potentially complicating factor, however, is that one is faced with unit nonresponse in the first stage, but what Brick and Kalton (1996) refer to as *partial nonresponse* in the second stage, a murky middle ground falling somewhere in between unit and item nonresponse. And although certain point estimates are formulated using data collected during the first stage, the preeminent estimates are those derived from the secondary data collection stage. Hence, assigning variable levels of tolerance, or detectability, across the two stages is a foreseeable goal. (Of course, this could also be a goal with respect to mixed-mode survey designs.) It may prove enlightening to delve deeper into the tradeoffs survey administrators must consider in this setting.

Appendix: Data Set Visualization of RGG Rule 3.

	Observed Data				Completed Data Sets Using Wave k - 1 Respondents			Completed Data Sets Using Wave k Respondents			Difference Variables		
Sample Case ID	**Wave**	w_i	$\mathbf{x_i}$	y_i	y_{1i}^{k-1}	…	y_{Mi}^{k-1}	y_{1i}^{k}	…	y_{Mi}^{k}	d_{1i}	…	d_{Mi}
1	1	5	10.1	1.3	1.3		1.3	1.3		1.3	0		0
2	1	5	11.4	1.1	1.1		1.1	1.1		1.1	0		0
3	1	5	8.2	2.1	2.1		2.1	2.1		2.1	0		0
4	1	5	7.7	1.8	1.8		1.8	1.8		1.8	0		0
5	1	5	7.9	1.7	1.7		1.7	1.7		1.7	0		0
6	1	5	9	2	2		2	2		2	0		0
7	2	5	11.5	1.4	?		?	1.4		1.4	?		?
8	2	5	11.1	1.8	?		?	1.8		1.8	?		?
9	2	5	8.8	1.6	?		?	1.6		1.6	?		?
10	2	5	9.1	1.9	?		?	1.9		1.9	?		?
11	?	5	9.5	?	?		?	?		?	?		?
12	?	5	9.2	?	?		?	?		?	?		?
13	?	5	9.4	?	?		?	?		?	?		?
14	?	5	7.1	?	?		?	?		?	?		?

Bibliography

Allison, P. (2010). *Survival Analysis Using SAS®: A Practical Guide. Second Edition*. Cary, NC: SAS Institute.

Atrostic, B., Bates, N., Burt, G., and Silberstein, A. (2001). "Nonresponse in US Government Household Surveys: Consistent Measures, Recent Trends, and New Insights," *Journal of Official Statistics*, **17**, pp. 209 – 226.

Bates, N., and Creighton, K. (2000). "The Last Five Percent: What Can We Learn from Difficult/Late Interviews?" *Proceedings of the Joint Statistical Meetings of the American Statistical Association*, pp. 120 – 125.

Bethlehem, J. (1988). "Reduction of Nonresponse Bias through Regression Estimation," *Journal of Official Statistics*, **4**, pp. 251 – 260.

Biemer, P., and Lyberg, L. (2003). *Introduction to Survey Quality*. Hoboken, NJ: Wiley.

Billiet, J., Philippens, M., Fitzgerald, R., and Stoop, I. (2007). "Estimation of Non-Response Bias in the European Social Survey: Using Information from Reluctant Respondents," *Journal of Official Statistics*, **23**, pp. 135 – 162.

Brick, M., and Kalton, G. (1996). "Handling Missing Data in Survey Research," *Statistical Methods in Medical Research*, **5**, pp. 215 – 238.

Brick, M., and Jones, M. (2008). "Propensity to Respond and Nonresponse Bias," *Metron-International Journal of Statistics*, **66**, pp. 51 – 73.

Couper, M. (1998). "Measuring Survey Quality in a CASIC Environment," *Proceedings of the Survey Research Methods Section of the American Statistical Association*.

Curtin, R., Presser, S., and Singer, E. (2000). "The Effect of Response Rate Changes on the Index of Consumer Sentiment," *Public Opinion Quarterly*, **64**, pp. 413 – 428.

Curtin, R., Presser, S., and Singer, E. (2005). "Changes in Telephone Survey Nonresponse Over the Past Quarter Century," *Public Opinion Quarterly*, **69**, pp. 87 – 98.

D'Agostino, R. (1998). "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group," *Statistics in Medicine*, **17**, pp. 2265 – 2281.

de Leeuw, E., and de Heer, W. (2002). "Trends in Household Survey Nonresponse: a Longitudinal and International Comparison," in *Survey Nonresponse,* eds. R. Groves, D. Dillman, J. Eltinge, and R. Little. New York, NY: Wiley.

de Leeuw, E. (2005). "To Mix or Not to Mix Data Collection Modes in Surveys," *Journal of Official Statistics,* **21**, pp. 233 – 255.

Deming, W. (1953). "On a Probability Mechanisms to Attain an Economic Balance Between the Resultant Error of Response and Bias of Nonresponse," *Journal of the American Statistical Association,* **48**, pp. 743 – 772.

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York, NY: Chapman and Hall.

El-Bawdry, M. (1956). "A Sampling Procedure for Mailed Questionnaires." *Journal of the American Statistical Association,* **51**, pp. 209 – 227.

Elliott, M., Little, R., and Lewitzky, S. (2000). "Subsampling Callbacks to Improve Survey Efficiency," *Journal of the American Statistical Association,* **95**, pp. 730 – 738.

Eltinge, J., and Yanseneh, I. (1997). "Diagnostics for Formation of Nonresponse Adjustment Cells, with an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey," *Survey Methodology*, **23**, pp. 33 – 40.

Filion, F. (1976). "Exploring and Correcting for Nonresponse Bias Using Follow-Ups of Nonrespondents," *Pacific Sociological Review*, **19**, pp. 401 – 408.

Fuller, W. (1975). "Regression Analysis for Sample Survey," *Sankhyā*, **37**, Series C, Pt. 3, pp. 117 – 132.

Graham J., Olchowski, A., and Gilreath, T. (2007). "How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory," *Prevention Science*, **8**, pp. 206 – 2010.

Groves, R. (1989). *Survey Errors and Survey Costs*. New York, NY: Wiley.

Groves, R., and Couper, M. (1998). *Nonresponse in Household Interview Surveys*. New York, NY: Wiley.

Groves, R. (2003). "Trends in Survey Costs and Key Research Needs in Survey Nonresponse," in Appendix B of Tourangeau, R. (2003). "Recurring Surveys: Issues and Opportunities," *A Report to the National Science Foundation Based on a*

Workshop Held on March 28 – 29, 2003. Retrieved November 12, 2002 at: http://www.nsf.gov/sbe/ses/mms/nsf04_211a.pdf

Groves, R. (2006). "Nonresponse Rates and Nonresponse Bias in household Surveys," *Public Opinion Quarterly*, **70**, pp. 646 – 675.

Groves, R., and Heeringa, S. (2006). "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs," *Journal of the Royal Statistics Society: Series A (Statistics in Society)*, **169**, pp. 439 – 457.

Groves, R., and Peytcheva, E. (2008). "The Impact of Nonresponse Rates on Nonresponse Bias: a Meta-Analysis," *Public Opinion Quarterly*, **72**, pp. 167 – 189.

Hamilton, J. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hansen, M., and Hurwitz, W. (1946). "The Problem of Nonresponse in Sample Surveys," *Journal of the American Statistical Association*, **41**, pp. 517 – 529.

Hartley, H. (1946). Discussion of "A Review of Recent Statistical Developments in Sampling and Sampling Surveys" by F. Yates, *Journal of the Royal Statistical Society: Series A*, **109**, pp. 37 – 38.

Heeringa, S., West, B., and Berglund, P. (2010). *Applied Survey Data Analysis*. Boca Raton, FL: Taylor & Francis.

Horvitz, D., and Thompson, D. (1952). "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association*, **47**, pp. 663 – 685.

Izrael, D., Hoaglin, D., and Battaglia, M. (2004). "To Rake or Not to Rake is Not the Question Anymore with the Enhanced Raking Macro," *Paper Presented at 29th SAS Users Group International (SUGI) Conference*. Montreal, Quebec, Canada. Retrieved February 2, 2013, from http://www2.sas.com/proceedings/sugi29/207-29.pdf

Jacoby, J., and Matell, M. (1971). "Three-Point Likert Scales are Good Enough," *Journal of Marketing Research*, 8, pp. 495-500.

Kalton, G. (1983). *Compensating for Missing Survey Data*. Survey Research Center, Institute for Social Research, University of Michigan: Ann Arbor, MI.

Kalton, G., and Flores-Cervantes, I. (2003). "Weighting Methods," *Journal of Official Statistics*, **19**, pp. 81 – 97.

Keeter, S., Kennedy, C., Dimock, M., Best, J., and Craighill, P. (2006). "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey," *Public Opinion Quarterly*, **70**, pp. 759 – 779.

Kish, L. (1965). *Survey Sampling*. New York, NY: Wiley.

Kovar, J., Rao, J.N.K., and Wu, C. (1988). "Bootstrap and Other Methods to Measure Errors in Survey Estimates," *Canadian Journal of Statistics*, **16**, pp. 25 – 45.

Kreuter, F., and Casas-Cordero, C. (2010). "Paradata," Working Paper 136. RatSWD Working Paper Series. Available online at http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_136.pdf

Kreuter, F. (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: Wiley.

Lessler, J., and Kalsbeek, W. (1992). *Nonsampling Error in Surveys*. New York, NY: Wiley.

Lin, I.-F., and Schaeffer, N. (1995). "Using Survey Participants to Estimate the Impact of Nonparticipation," *Public Opinion Quarterly*, **59**, pp. 236–258.

Little, R. (1986). "Survey Nonresponse Adjustments for Estimates of Means," *International Statistical Review*, **54**, pp. 139 – 157.

Little, R., and Rubin, D. (2002). *Statistical Analysis with Missing Data*. 2nd Edition. New York, NY: Wiley.

Little, R., and Vartivarian, S. (2005). "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology*, **31**, pp. 161–168.

Lewis, T. (2012). "Incorporating the Sampling Variability from an Employee Perception Survey into the Ranking Process of U.S. Government Agencies." *Proceedings of the Fourth International Conference on Establishment Surveys* (ICES-IV).

Lynn, P., Clarke, P., Martin, J., and Sturgis, P. (2002). "The Effects of Extended Interviewer Effects on Nonresponse Bias," in *Survey Nonresponse* by Groves, R., Dillman, D., Eltinge, J., and Little R. (eds). New York, NY: Wiley.

McPhee, C., and Hastedt, S. (2012). "More Money? The Impact of Larger Incentives on Response Rates in a Two-Phase Mail Survey," *Proceedings from the Federal Committee on Statistical Methodology (FCSM) Research Conference*.

Merkle, D. and Edelman, M. (2002). "Nonresponse in Exit Pools: A Comprehensive Analysis," in *Survey Nonresponse* by Groves, R., Dillman, D., Eltinge, J., and Little R. (eds). New York, NY: Wiley.

O'Quigley, J., Pepe, M. and Fisher, L. (1990). "Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer," *Biometrix*, **46**, pp. 33 – 48.

Parsons V., Moriarity, C., and Jonas, K. (2014). *Design and Estimation for the National Health Interview Survey, 2006–2015*. Vital Health Statistics, Series 2, Vol. 165. Hyattsville, MD: National Center for Health Statistics.

Politz, A., and Simmons, W. (1949). "An Attempt to Get the Not-at-Homes into the Sample without Callbacks," *Journal of the American Statistical Association*, **44**, pp. 9 – 31.

Peytchev, A., Baxter, R., and Carley-Baxter, L. (2009). "Not All Survey Effort Is Equal: Reduction of Nonresponse Bias and Nonresponse Error," *Public Opinion Quarterly*, **73**, pp. 785 – 806.

Potthoff, R., Manton, K., Woodbury, M. (1993). "Correcting for Nonavailability Bias in Surveys Weighting Based on the Number of Callbacks," *Journal of the American Statistical Association*, **88**, pp. 1197 – 1207.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001). "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, **27**, pp. 85 – 95.

Rao, R., Glickman, M., and Glynn, R. (2004). "Use of Covariates and Survey Wave to Adjust for Nonresponse," *Biometrical Journal*, **46**, pp. 579 – 588.

Rao, R., Glickman, M., and Glynn, R. (2008). "Stopping Rules for Surveys with Multiple Waves of Nonrespondent Follow-Up," *Statistics in Medicine*, **27**, pp. 2196 – 2213.

Reiter, J., Raghunathan, T., and Kinney, S. (2006). "The Importance of Modeling the Sample Design in Multiple Imputation," *Survey Methodology*, **32**, pp. 143 – 149.

Rosenbaum, P., and Rubin, D. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, **70**, pp. 41 – 55.

Rubin, D., and Schenker, N. (1986). "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association*, **81**, pp. 366 – 374.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley.

Rust, K. (1985). "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics*, **1**, pp. 381 – 397.

Schenker, N., Raghunathan, T., Chiu, P.-L., Makuc, D., Zhang, G., and Cohen, A. (2006). "Multiple Imputation of Missing Income Data in the National Health Interview Survey," *Journal of the American Statistical Association*, **101**, pp. 924 – 933.

Sigman, R., Lewis, T., Dyer, N., and Lee, K. (2012). "Does The Length of Fielding Period Matter? Examining Response Scores of Early versus Late Responders." *Proceedings of the Fourth International Conference on Establishment Surveys* (ICES-IV).

Singer, J., and Willett, J. (2003). *Applied Longitudinal Data Analysis*. New York, NY: Oxford.

Torrieri, N. (2014). "American Community Survey: Design and Methodology (January 2014)." Available on-line at: http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology_report_2014.pdf

Valliant, R., Dever, J., and Kreuter, F. (2013). *Practical Tools for Designing and*

Weighting Survey Samples. New York, NY: Springer.

Wagner, J. (2010). "The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data," *Public Opinion Quarterly*, **74**, pp. 223 – 243.

Wagner, J., and Raghunathan, T. (2010). "A New Stopping Rule for Surveys," *Statistics in Medicine*, **29**, pp. 1014 – 1024.

Wagner, J., and Hubbard, F. (2014). "Unbiased Estimates of Propensity Models During Data Collection," *Journal of Survey Statistics and Methodology*, **2**, pp. 323 – 342.

Westat. (2007). *WesVar® 4.3 User's Guide*. Retrieved April 9, 2012 at: http://www.westat.com/Westat/pdf/wesvar/WV_4-3_Manual.pdf

Wolter, K. (2007). *Introduction to Variance Estimation*. *Second Edition*. New York, NY: Springer.

Woodruff, R. (1971). "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association*, **66**, pp. 411 – 414.