Testing for Phase Capacity in Surveys with Multiple Waves of Nonrespondent Follow-Up

Taylor Lewis[1]
U.S. Office of Personnel Management, 1900 E St. NW, Washington, DC 20415

Abstract
To combat nonresponse, many surveys repeatedly follow up with nonrespondents, often targeting a response rate or a predetermined number of completes. Under a fixed data collection protocol, however, returns tend to diminish with each subsequent wave of data collected, and (nonresponse-adjusted) estimates eventually stabilize. This is the concept of phase capacity, suggesting some sort of design change is warranted (e.g., switch modes, increase the incentive, or discontinue follow-up altogether). The only known test for phase capacity appearing in the literature is one employing imputation models to adjust for nonresponse. This paper describes a test similar in spirit but applicable to surveys that conduct weighting adjustments to compensate for nonresponse. The two methods are compared via an application using data from a Web-based satisfaction survey of United States government employees. The weighting version of the phase capacity test proves more conservative in the sense that it tends to conclude more follow-up attempts are warranted.

Key Words: responsive survey design, multiple imputation, weighting, nonresponse

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) asserts 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 has risen 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 recent trends are alarming, survey researchers should take comfort in the fairly widespread, convincing evidence that lower response rates do not necessarily imply estimates are systematically less accurate (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. Indeed, each follow-up attempt tends to prompt more survey completes. We can think of these additional cases acquired as waves of incoming data. On the

[1] The opinions, findings, and conclusions expressed in this paper are those of the author alone and do not necessarily reflect those of the U.S. Office of Personnel Management.

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 over the since that time (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 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; 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 alter the course of data collection. 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 ends. This paper aims to fill this research gap by discussing and comparing two methods for formally testing for phase capacity.

As an aside, the often-utilized "stopping rule" label typically carries the connotation that the nonrespondent follow-up should be discontinued altogether once phase capacity has been reached. This is not exactly the case. More precisely, 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, 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

Background

To further elucidate the concept of phase capacity and introduce a real-world survey data set on which two proposed tests will be compared, 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 demographics. 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.

The sample frame for the FEVS is 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.

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 collected between any two weekly email invitations. Table 1 shows the wave-specific respondent counts and

corresponding relative percent increase for one example agency. It is plain to see how the relative increases quickly diminish within a few waves.

Table 1: Distribution of Responses by Wave for an Example Agency Participating in FEVS 2011.

Wave	*Respondents*	*Percent Increase*
1	2,178	--
2	1,516	69.6%
3	1,304	35.3%
4	959	19.2%
5	613	10.3%
6	510	7.8%
7	439	6.2%
8	381	5.1%
9	408	5.2%
10	379	4.6%
	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) at the agency level. 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 shifted to respondents within classes formed by the cross-classification of various demographic variables. 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 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 the response indicator back into the sample frame, a series of nonresponse-adjusted point estimates can be charted across time as additional waves of data are incorporated.

Figure 1 illustrates this 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, that estimates generated from earlier respondents tend to be lower than estimates generated from the ultimate set of respondents (Sigman et al., 2012).

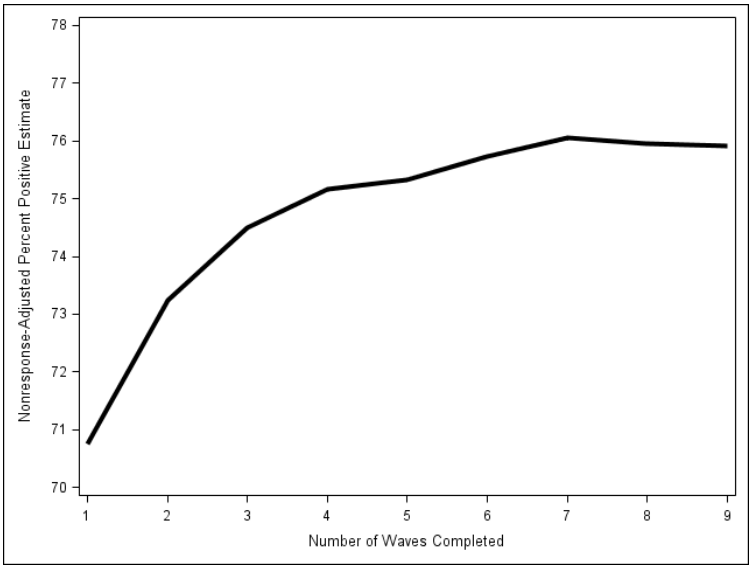

Figure 1: Plot of the Nonresponse-Adjusted Percent Positive Statistic for FEVS Item 4 in an Example Agency Using Cumulative Data as of the Given Wave of Nonrespondent Follow-Up.

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, preventing inefficacious nonrespondent follow-up attempts.

2. A Retrospective Test for Phase Capacity

2.1 Previous Methods

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 $\bar{\hat{Q}}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m$. 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 $\bar{\hat{U}}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{U}_m$ and (2) the between-imputation variance of the estimate $\hat{B}_M = \left(1+\frac{1}{M}\right)\sum_{m=1}^{M}\frac{\left(\hat{Q}_m - \bar{\hat{Q}}_M\right)^2}{M-1}$. That is, the overall multiple imputation variance formula is $\hat{T}_M = \bar{\hat{U}}_M + \hat{B}_M$. 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. For respondents up to and including wave $k-1$, $d_{mi} = 0$, but necessarily so for respondents during wave k and beyond.

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).

2.2 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, often 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 portion of the data.

Suppose 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 $T = 4$ estimated totals:

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

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 $\binom{T}{2}$ 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, $\text{var}(\hat{\delta}_{k-1}^{k}) \approx \text{var}\left(\sum_{i=1}^{n} \sum_{t=1}^{T=4} \frac{\partial \delta_{k-1}^{k}}{\partial T_t} \hat{t}_{ti} \right)$. After a little algebra, it can be shown this equals

$$u_i = \frac{1}{\hat{N}_1^{k-1}} w_{1i}^{k-1} y_i - \frac{\hat{Y}_1^{k-1}}{\left(\hat{N}_1^{k-1}\right)^2} w_{1i}^{k-1} - \frac{1}{\hat{N}_1^{k}} w_{1i}^{k} y_i + \frac{\hat{Y}_1^{k}}{\left(\hat{N}_1^{k}\right)^2} w_{1i}^{k} \qquad (2)$$

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 2 provides a visualization of this technique using a simple, hypothetical survey data set where $k = 2$.

Table 2: 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 from Table 2, 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 10 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: $\delta_1^2 = \bar{y}_1^1 - \bar{y}_1^2 \neq 0$. 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 surveys involve multiple 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 function of unbiased totals, which covers a wide range of statistics. This is a notable advantage over RGG's rule, whose difference variable approach was designed specifically for the comparison of two sample means.

Worthy of brief mention 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), 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 $\text{var}(\hat{\delta}_{k-1}^k)$ 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, …, 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, 2*R* replicate weights are constructed. One then conducts the full nonresponse adjustment routine on all replicate weights independently. After finding both $\hat{\theta}_{1r}^{k-1}$ and $\hat{\theta}_{1r}^{k}$ using the two sets replicate weights, the 2*R* estimates are consolidated by forming $\hat{\theta}_r = \hat{\theta}_{1r}^{k-1} - \hat{\theta}_{1r}^{k}$. Ultimately, the average squared deviation of these *R* estimated differences from the full-sample difference $\hat{\theta} = \hat{\theta}_1^{k-1} - \hat{\theta}_1^{k}$ approximates the variance of the two sample means' difference. Other replication approaches, such as the *jackknife* (Ch. 4 of Wolter, 2007) or the *bootstrap* (Ch. 5 of Wolter, 2007), could be conducted in a similar manner.

3. Application to the Federal Employee Viewpoint Survey

We next discuss an application of these methods using data from three agencies participating in FEVS 2011. As before, the estimates under investigation are sample means—namely, the seven percent positive estimates for items constituting OPM's Job Satisfaction index. The fundamental objective was to evaluate the performance of the two competing phase capacity tests. 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 RGG's 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).

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 MI approach. As with the simulation, Taylor series linearization was utilized to approximate the variance of the adjacent-wave weighted mean difference.

Table 3 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 the interpretation of nonresponse error in this application is perhaps more aptly described as *relative* nonresponse error, because we define it here 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. Note, also, that the two nonresponse-adjusted 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 how 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.

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), 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 or 2 percentage points could make a difference when assessing whether a change relative to the previous years' survey results was statistically significant, a popular technique human resources managers use to flag items deserving celebration or requiring intervention.

Another observation worth mentioning 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.

Table 3: Results from a 2011 Federal Employee Viewpoint Survey Phase Capacity Test 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

4. Discussion

This paper introduced an adaptation of the phase capacity test 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 and examples focused on the sample mean, the weighting variant is more flexible in that it can easily be altered to accommodate other estimators, whereas the M difference variable approach outlined in Rao, Glickman, and Glynn (2008) is geared specifically towards investigating a sample mean difference.

The primary conclusion is that the weighting version is more sensitive to point estimate deviations. This is due to the fact that $\text{var}(\hat{\delta}_{k-1}^{k}) = \text{var}(\hat{\bar{y}}_1^{k-1}) + \text{var}(\hat{\bar{y}}_1^{k}) - 2\,\text{cov}(\hat{\bar{y}}_1^{k-1}, \hat{\bar{y}}_1^{k})$, the variance term in the denominator of the *t* statistic is smaller in the RGG version than in the RGG version. Because the FEVS application was designed such that the wave-specific sample means would have variances equal in expectation, we must attributed the difference to how the covariance term embedded in $\text{var}(\hat{\delta}_{k-1}^{k})$ is implicitly calculated. We leave for further research the task of developing a more formal theoretical understanding as to why this is so. 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 a cursory analysis suggested the jackknife and bootstrap methods approaches mirrored the performance of the Taylor Series linearization method presently utilized, a more rigorous study investigating other estimators would be useful to rule out potential anomalies.

This work is derived from a more comprehensive treatment on the topic currently being undertaken as part of this author's PhD thesis for the Joint Program in Survey Methodology (JPSM) at the University of Maryland, College Park. Still in preparation at the time of this writing, the thesis details a simulation study conducted to further evaluate these two phase capacity tests. Findings from that simulation study echo many of the same findings from the FEVS application presented in this paper. Additional research covered in the thesis includes (1) a multivariate extension of the methods proposed in this paper, with the aim of providing a single yes-or-no phase capacity determination for two or more estimates simultaneously monitored and (2) an extension of a prospective test for phase capacity proposed by Wagner and Raghunathan (2010), with the aim of determining whether a pending wave of data collection will significantly influence key estimates.

References

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.

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.

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.

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.

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

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.

Kreuter, F. (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: 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., and Vartivarian, S. (2005). "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology*, **31**, pp. 161–168.

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.

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.

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.

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).

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.

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.