DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 3, 2017, pp. 601–624, http://dx.doi.org/10.1515/JOS-2017-0029

Univariate Tests for Phase Capacity: Tools for Identifying When to Modify a Survey's Data Collection Protocol

Taylor Lewis[1]

To mitigate the potentially harmful effects of nonresponse, most surveys repeatedly follow up with nonrespondents, often targeting a 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, as each subsequent wave tends to consist of fewer responses than the last. Consequently, point 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. A previously proposed test for phase capacity calls for multiply imputing nonrespondents' missing data to assess, retrospectively, whether the most recent wave of data significantly altered a key, nonresponse-adjusted point estimate. This study introduces a more flexible adaptation amenable to surveys that instead reweight the observed data to compensate for nonresponse. Results from a simulation study and application indicate that, all else equal, the weighting version of the test is more sensitive to point estimate changes, thereby dictating more follow-up attempts are warranted.

Key words: Responsive survey design; multiple imputation: weighting; nonresponse.

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 over the past few decades (Atrostic et al. 2001; De Leeuw and De Heer 2002; Curtin et al. 2005; Brick and Williams 2013; Tourangeau and Plewes 2013). Typically, a survey's data collection protocol involves making a sequence of follow-ups on nonrespondents, 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 produce more survey completes, which can be conceptualized as incoming *waves* of data. More follow-ups are ostensibly desirable, as they serve to reduce the nonresponse rate, but they do not guarantee a reduction in nonresponse error. More follow-ups come at a cost, however, and can extend the data

[1] U.S. Office of Personnel Management (OPM), 1900 E Street, NW, Washington, DC 20415. U.S.A. Email: Taylor.Lewis@opm.gov

Acknowledgments: The findings and conclusions in this article are those of the author and do not necessarily represent the views of the U.S. Office of Personnel Management. This work was carried out as part of a PhD dissertation for the Joint Program in Survey Methodology at the University of Maryland, College Park. The author would like to thank dissertation committee co-chairs Frauke Kreuter and Partha Pahiri as well as committee members Richard Valliant, James Wagner, and Michael Rendall.

collection period, delaying subsequent stages of the survey process, such as the reporting and analysis stages. Moreover, from a purely practical standpoint, empirical evidence (e.g., Table 1 in Potthoff et al.1993) suggests returns diminish with each subsequent wave. 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 obtain 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 et al. (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 strand 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. 2014). 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 limitation 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 procedure. 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) urge researchers to employ principles of *responsive survey design*, which Bethlehem et al. (2011) note is a special case of *adaptive survey design* (Wagner 2008). The basic premise of responsive survey design is to utilize paradata in real-time to help inform data collection decisions and, if necessary, change course. Groves and Heeringa (2006) refer to a *design phase* as 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. They argue 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 once 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 article, in which they plot the trend of a key, nonresponse-adjusted point estimate over the data collection period and comment on how it stabilizes well before the

design phase concludes. The methods discussed in this article aim to fill this gap in the literature.

As an aside, it should be acknowledged that the survey methodology literature is replete with strategies for allocating resources while following up with nonrespondents to maximize precision, minimize costs, and/or minimize nonresponse error (e.g., Hansen and Hurwitz 1946; Filion 1976; Deming 1953; El-Badry 1956; Elliott et al. 2000; Peytchev et al. 2009; Beaumont et al. 2014; Schouten and Schlomo 2015). Testing for phase capacity has a subtly different objective: to determine the point(s) during a fixed data collection protocol when it would be prudent to introduce some form of change, perhaps by pursuing one of these resource allocation strategies. Phase capacity does not necessarily mean that the point estimate is devoid of nonresponse error, but it does suggest future follow-up attempts using the same recruitment protocol will be limited in their ability to reduce nonresponse error. If the point estimate itself is hardly changing, it follows that any nonresponse error associated with it is also hardly changing.

Another critical point worth articulating is that phase capacity tests are often referred to in the literature as "stopping rules." We hesitate to adopt that terminology here because it carries a connotation that the nonrespondent follow-up campaign should be discontinued altogether once phase capacity has been reached. More precisely, phase capacity indicates that 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 article, but alternative interventions include switching modes (De Leeuw 2005) or increasing the incentive to participate (McPhee and Hastedt 2012).

This article is structured as follows. Section 2 demonstrates the phenomenon of phase capacity within the context of a real-world survey. In Section 3, two tests for phase capacity are introduced. A simulation study conducted to compare and contrast their performance is detailed in Section 4, followed by an application to actual survey data in Section 5. Section 6 concludes with a summary of the key findings and a discussion of potential avenues for further research.

2. Illustrating Phase Capacity in the Federal Employee Viewpoint Survey

To further elucidate the concept of phase capacity and introduce a real-world survey data set on which the two phase capacity 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 administered in 2002 by the U.S. Office of Personnel Management (OPM). Initially conducted 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-type 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 were missing. The key estimate from each item thus reduces to the proportion (or percentage) of employees who react positively

to the statement posed. The terminology used to describe this statistic is the "percent positive" for a particular survey item. Although this dichotomization foregoes some information, Jacoby and Matell (1971) argue that it does not cause any significant decrement in reliability or validity.

The FEVS sampling frame 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.

The sampling frame contains a plethora of auxiliary variables known for both respondents and nonrespondents, a subset of which is utilized in a three-step weighting procedure to compensate for unit nonresponse (Kalton and Flores-Cervantes 2003). As described in Appendix E of OPM (U.S. Office of Personnel Management 2015), 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 third step, weights are raked such that they aggregate to certain known frame totals for the agency as a whole.

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 somewhat, but the median duration was six weeks. The FEVS 2011 data collection strategy fits well within the paradigm of a stable recruitment protocol with multiple nonrespondent follow-up attempts. On the survey start date, an initial email invitation containing the website URL and login 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 clear demarcation of a data collection wave is the set of responses obtained between any two weekly email solicitations. Table 1 shows the wave-specific respondent counts and corresponding relative percent increase for the three agencies analyzed in this article. It is plain to see how the relative increases quickly diminish after the first few waves, suggesting that the impact on percent positive estimates diminishes correspondingly. At the conclusion of the last respective wave undertaken, all three agencies had achieved about a 50% response rate.

The survey reminder schedule is generally fixed for each agency prior to the start of the survey, but for many percent positive estimates, phase capacity occurs before the final reminder email is sent. Since responses are electronically recorded real-time and all weighting adjustments can be made after merging the response indicator back onto the sampling frame, a series of nonresponse-adjusted point estimates can be plotted as each

Table 1. FEVS 2011 respondent counts by data collection wave (a calendar week) for three example agencies.

	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	

incoming wave of data is incorporated. Figure 1 illustrates this type of plot for an example agency based on the percent positive statistic associated with Item 4 on the survey instrument, which asks employees their level of agreement with the statement "My work gives me a feeling of personal accomplishment." The figure shows how the point estimate increases over the course of data collection, even after adjusting for unit nonresponse. By about Wave 6, however, it has more or less stabilized. While the tendency for nonresponse-adjusted estimates to change 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)), this particular pattern – estimates derived from earlier respondents tending to be less

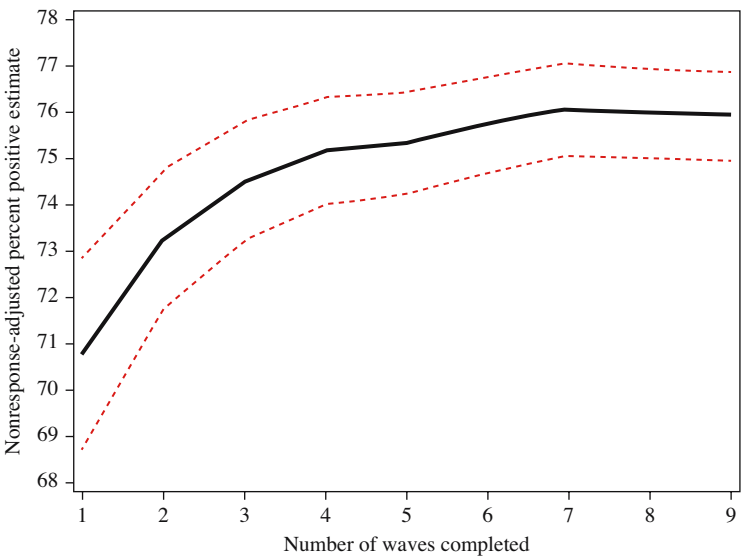

Fig. 1. Trends of nonresponse-adjusted percent positive statistic and 95% confidence interval for FEVS 2011 Item 4 using cumulative data as of the given wave of nonrespondent follow-up for an example agency.

positive than estimates generated using the ultimate set of respondents – is observed for numerous FEVS items (Sigman et al. 2014).

3. Tests for Phase Capacity

3.1. A Method Based on Multiple Imputation

Rao, Glickman, and Glynn (RGG) (2008) was the first known proposal for 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 objective 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. The key estimate of interested was a sample mean, the proportion of women using birth control. They derived three rules to assess whether the estimated proportion changed substantively following the completion of wave k ($k \geq 2$) relative to the value following the completion of wave $k-1$.

RGG's third rule performed best in simulation and application. It adjusted for nonresponse by multiply imputing (Rubin 1987) the 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 observed relationship between $\mathbf{X}$ and $\mathbf{Y}_1$ to derive plausible values of $\mathbf{Y}_0$ given $\mathbf{X}$. *Multiple imputation* (MI) is a technique whereby missing values are imputed M times ($M \geq 2$), resulting in M completed data sets. Rubin (1987) advocates this technique over single imputation because an augmentation to the variance formula allows one to better reflect the missing data uncertainty.

Let $\hat{Q}_m$ denote the *mth* completed data set estimate for a generic 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 *mth* completed data set estimated variance for $\hat{Q}_m$. The MI variance is the sum of two components: (1) the average of the M completed data set variances, $\hat{\bar{U}}_M = \frac{1}{M}\sum_{m=1}^{M} \hat{U}_m$; and (2) the between-imputation variance of the estimate, $\hat{B}_M = \sum_{m=1}^{M} \frac{\left(\hat{Q}_m - \hat{\bar{Q}}_M\right)^2}{M-1}$. Taken together, the overall multiple imputation variance is $\hat{T}_M = \hat{\bar{U}}_M + \left(1 + \frac{1}{M}\right)\hat{B}_M$, where the term $\left(1 + \frac{1}{M}\right)$ represents a finite imputation correction factor, which converges to 1 as $M \rightarrow \infty$.

RGG's third rule proceeds as follows. First, one imputes the missing data for nonrespondents using all available information. Next, responses obtained during wave k, specifically, are deleted and imputation is performed using a model fit using only the respondents as of wave $k-1$. The result is $2M$ completed data sets, which are clearly dependent, because the two underlying models used to create them are based on the shared set of respondents through wave $k-1$. To avert explicit calculation of a covariance, RGG assert one can 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 or data used in the imputation process and the subscript denotes the *mth* completed data set value (imputed or observed), for the *ith* individual. A contrived data set is presented in an appendix to help visualize 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 average of the M values of $\text{var}(\hat{\bar{d}}_m)$ and the sample variance of the M values of $\hat{\bar{d}}_m$ times the finite imputation correction factor. The former is the within-imputation variance component and the latter is the between-imputation variance component.

3.2. A New Method Based on Weighting

RGG's phase capacity test is restrictive in the sense that not all surveys employ MI to address unit nonresponse. Many surveys, including the FEVS, instead adjusting the weights of respondents (Kalton and Flores-Cervantes 2003) such that they better represent the original sample or target population. In this section, we introduce an adaptation of the MI phase capacity test amenable to surveys that reweight the observed data to combat nonresponse.

Suppose for the moment that one is interested in determining whether $\hat{\bar{y}}_1^k$, the sample mean using data from waves 1 through wave k, is not significantly 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 at the conclusion of the two respective, adjacent waves. For cases responding at or before wave $k-1$, both weights are positive. For cases responding specifically during wave k, w_{1i}^k is positive while $w_{1i}^{k-1} = 0$. For cases that have yet to respond by wave k, both w_{1i}^k and w_{1i}^{k-1} are 0.

From here, just as in the MI version of the test, the objective is to standardize the difference between the two sample means. Fundamentals of Taylor series linearization can be employed to derive an estimated variance of the difference after first observing how the difference can be expressed as a function of $p = 4$ estimated totals:

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

When written in this fashion, Wolter (2007, sec. 6.5) demonstrates how a computational algorithm attributable to Woodruff (1971) can greatly simplify the Taylor series variance estimation 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 variate u_i at the Primary Sampling Unit (PSU) level equaling the sum of the function's partial derivatives multiplied by the corresponding estimated total. In the

present case, $\text{var}\left(\hat{\delta}_{k-1}^{k}\right) \approx \text{var}\left(\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{\partial \delta_{k-1}^{k}}{\partial T_j}t_{ji}\right)$, where t_{ji} represents the PSU-level estimate of the jth total in the function. Specifically, $t_{1i} = w_{1i}^{k-1}y_i$, $t_{2i} = w_{1i}^{k-1}$, $t_{3i} = w_{1i}^{k}y_i$, and, $t_{4i} = w_{1i}^{k}$. After a little algebra, it can be shown

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

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

Using figures in 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 $\text{var}\left(\hat{\delta}_1^2\right)$ 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

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*
Sampling unit 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 $\hat{Y}_1^2 = 100.86$.

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 number of respondents in 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, indicating that phase capacity has occurred.

When thinking about the structure of the quotient that makes up the phase capacity test statistic, one can reason how the precision (or lack thereof) of $\text{var}\left(\hat{\delta}_1^{k-1}\right)$, $\text{var}\left(\hat{\delta}_1^{k}\right)$, and, therefore, $\text{var}\left(\hat{\delta}_{k-1}^{k}\right) = \text{var}\left(\hat{\delta}_1^{k-1}\right) + \text{var}\left(\hat{\delta}_1^{k}\right) - 2\text{cov}\left(\hat{\delta}_1^{k-1}, \hat{\delta}_1^{k}\right)$, is a key factor in determining how many follow-up attempts are deemed necessary. In particular, a high degree of precision renders one more likely to reject the null hypothesis and continue, whereas imprecision renders one more likely to fail to reject the null hypothesis and discontinue. A driver of the former might be, say, a small relative increase in new respondents following a particular data collection wave. A driver of the latter might be small respondent counts, say, in early data collection waves. On the one hand, a high degree of precision could be perceived as an advantage, as there is seemingly less risk for residual nonresponse error; however, this may lead the phase capacity test to detect differences that are statistically significant, but not practically significant. On the other hand, it seems ill-advised to have a lack of precision alone be the sole determinant of phase capacity. As such, it may be prudent for practitioners to designate an attainable precision target or minimum number of data collection waves that must be attempted prior to adhering to the conclusions of the tests. Naturally, this will depend on the point estimate of interest, the data collection mode, and the analytic objectives of the survey, among other factors.

One potentially useful diagnostic tool for thinking about the role of (im)precision is to plot a statistical power curve for detecting a statistically significant difference in the adjacent wave point estimates. Figure 2 plots one such curve using the data from Table 2.

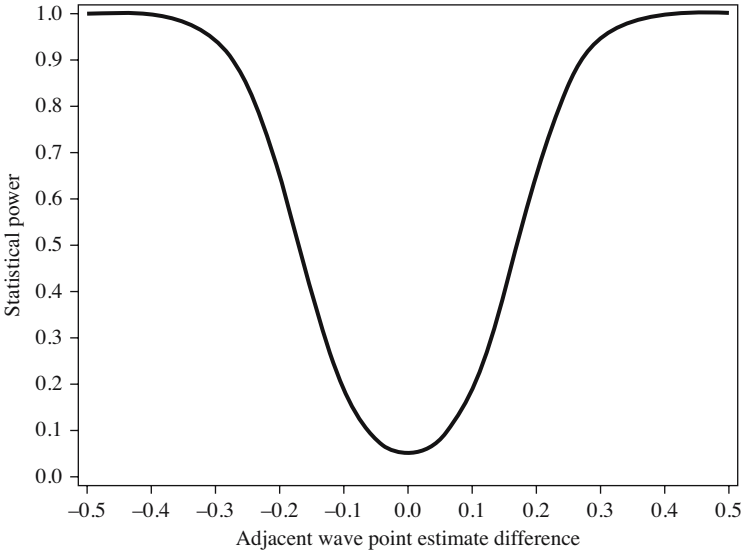

Fig. 2. Statistical power curve based on the example survey data presented in Table 2.

For hypothetical values of δ plotted along the x-axis, the statistical power for detecting a point estimate difference at the $\alpha = 0.05$ significance level is plotted along the y-axis. For a two-sided alternative hypothesis test with cut-off $t_{df,1-\alpha/2}$, the statistical power curve can be found by solving for $1 - \Pr\left[-t_{df,1-\alpha/2} - \frac{\delta}{\sqrt{\text{var}\left(\hat{\delta}_{k-1}^{k}\right)}} < t < t_{df,1-\alpha/2} - \frac{\delta}{\sqrt{\text{var}\left(\hat{\delta}_{k-1}^{k}\right)}}\right]$. With respect to the data in Table 2, $t_{9,0.975} = 2.262$, and $\text{var}\left(\hat{\delta}_{k-1}^{k}\right) = 0.00566$.

The curve in Figure 2 indicates that one is virtually guaranteed to detect a difference of $+/-$ 0.4. Acknowledging the tendency for point estimate changes to decrease as k increases, practitioners may find information gleaned from this kind of plot useful in reinforcing or overturning the phase capacity test declaration. For example, if a meaningful change would have to be something much larger than $+/-$ 0.4, it may weaken the case for continuing under the same data collection protocol, despite the phase capacity test indicating one should do so. Of course, information from this kind of plot could also be used from the other end of the spectrum, such as to overturn an early-wave phase capacity declaration.

While the set-up thus far has pertained only to simple random sample designs, complex survey features can easily be accommodated. For instance, many survey samples involve hierarchical stages of clustering, often within strata. To simplify the variance approximation process, the "ultimate cluster" assumption (Heeringa et al. 2010, p. 67) is frequently adopted in which case the u_i's are constructed as illustrated above at the PSU level and stratum-specific variances are estimated and assimilated into an overall variance estimate. And although the present exposition focuses 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 major advantage over the MI version of the test, which was designed specifically to test for a difference in means.

Another avenue for estimating $\text{var}\left(\hat{\delta}_{k-1}^{k}\right)$ is to employ a *replication* approach (Rust 1985), such as the *jackknife* (Wolter 2007, chap. 4) or the *bootstrap* (Efron and Tibshirani 1993). The idea is to form two sets of R replicate weights, one for respondents through wave $k-1$ and another for all respondents through wave k. One then conducts the full nonresponse adjustment procedure independently on all replicate weights. This enables one to account for the added variance attributable to the nonresponse adjustment procedure (Valliant 2004). After finding both $\hat{\theta}_1^{(k-1)r}$ and $\hat{\theta}_1^{kr}$ using the two sets of replicate weights, the $2R$ estimates are consolidated by forming $\hat{\theta}^r = \hat{\theta}_1^{(k-1)r} - \hat{\theta}_1^{kr}$. From there, $\text{var}\left(\hat{\delta}_{k-1}^{k}\right)$ is estimated by a simple function of the squared deviation of these R differences from the full-sample difference $\hat{\theta} = \hat{\theta}_1^{k-1} - \hat{\theta}_1^{k}$, depending on the particular method.

As one final aside, there is an alternative computational strategy practitioners may find easier to apply than the method outlined whenever the point estimate of interest 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 respondent 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 in a simple random sample design, one would treat the unique respondent identifier as the PSU. The next step is to assign an indicator variable in the stacked data set taking on a value of 0 for

cases originally from the respondent data set as of wave k and a value of 1 for cases originally from the respondent data set as of wave $k - 1$. 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 the model is zero is equivalent to the two-sample t test described earlier in this section.

4. A Simulation Study Comparing the Two Phase Capacity Tests

In this section, we report results from a simulation study conducted to compare and contrast the two phase capacity tests. The goal was to evaluate their performance in four settings based on the cross-classification of two conditions: (1) whether or not the response wave is associated with a covariate known for the entire sample; and (2) whether or not a continuous outcome variable is associated with the response wave. Using a randomly generated covariate, each simulated survey sample was split into two classes within which both a weighting adjustment and multiple imputation routine could be performed. In effect, data were assumed Missing Completely At Random (MCAR) (Little and Rubin 2002) within each class. For the weighting version of the test, a single adjustment factor proportional to the inverse of the class-specific response rate was used to inflate the weights of respondents back the count from the initial sample. For the MI version of the test, the *Approximate Bayesian Bootstrap* (ABB) was carried out within each class. Rubin and Schenker (1986) prove that the expected value of the MI variance of a sample mean after implementing the ABB is approximately equal to the variance of the sample mean using only the observed portion of the data, $\mathbf{Y}_1$. Considering that, within a class, a constant weight adjustment will have no effect on the estimated variance of a mean, we can reason that the two techniques should be completely balanced in terms of their expected post-adjustment effect on the estimated variance of $\hat{\bar{y}}_1^{k-1}$ and $\hat{\bar{y}}_1^{k}$.

To partition each sample into two classes of roughly equal size, a random uniform variate x_i between 0 and 1 was first generated. A sample case was assigned to the first class if this number was less than 0.5, and the second class otherwise. Table 3 summarizes the two wave-of-response distributions that were defined using the empirical response distribution of Agency 3 in FEVS 2011 reported in Table 1. For the condition where the response wave is not associated with x_i, sample cases were assigned response waves in proportion to the empirical FEVS distribution. For the second condition where the response wave is associated with the covariate, if $x_i < 0.5$, the sample case was predisposed to respond sooner than when $x_i \geq 0.5$. In doing so, however, the expected marginal distribution was designed to match that of the first condition – for instance, $0.5*(34.5\% + 15.6\%) \approx 25.1\%$ and $0.5*(20.7\% + 14.2\%) \approx 17.5\%$.

In contrast to the RGG (2008) simulation study design, the outcome variable y_i was assigned as continuous rather than dichotomous. This was done to magnify the potential differences to be observed in sample means over the simulated data collection period. 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

Table 3. Summary of the two wave-of-response distributions used for the simulation study comparing the two phase capacity tests.

	Wave not associated with covariate	Wave associated with covariate	
Wave	for any x_i	for $x_i < 0.5$	for $x_i \geq 0.5$
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%

wave, $y_i = 10x_i + wave_i + \varepsilon_i$. Thus, the wave-specific mean outcome increases linearly in expectation.

Each of the four conditions were simulated 1,000 times for sample sizes $n = 500$ and $n = 5{,}000$. A practical issue when employing MI 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 deemed not a parameter worthy of manipulating in the simulation. Another consideration was the variance approximation method for $\text{var}\left(\hat{\delta}_{k-1}^{k}\right)$. Although the exposition in the previous section focused predominantly on the Taylor series linearization approach, it was noted that a replication approach would also be viable. To this end, a nonparametric bootstrap estimator was investigated during initial analyses, but results did not differ substantively from those obtained via Taylor series linearization. As such, the particular variance approximation method implemented was deemed immaterial for purposes 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 4a and 4b summarize results from the simulation study. The former presents a summarization where $n = 500$ and the latter where $n = 5{,}000$. 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" reports the average distance between the 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 line labeled "95 Percent Coverage" measures the percentage of abridged data set sample mean confidence intervals that encompass the full-sample mean.

Table 4a. Simulation study results comparing the performance of the two phase capacity tests (n = 500).

Condition	Measure	MI		Weighting	
		$\varepsilon_i \sim N(0,1)$	$\varepsilon_i \sim N(0,9)$	$\varepsilon_i \sim N(0,1)$	$\varepsilon_i \sim N(0,9)$
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 4b. Simulation study results comparing the performance of the two phase capacity tests (n = 5,000).

Condition	Measure	MI		Weighting	
		$\varepsilon_i \sim N(0,1)$	$\varepsilon_i \sim N(0,9)$	$\varepsilon_i \sim N(0,1)$	$\varepsilon_i \sim N(0,9)$
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

One overarching finding is that when the outcome is not related to the wave of response, as simulated in Conditions 1 and 3, both the MI 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 intervals formed earlier in the simulated 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 the expected value of the outcome variable 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 plagued by substantial nonresponse error and an unsatisfactory confidence interval coverage rate. That said, there is a fair amount of variability in terms of the mean stopping wave in the $n = 500$ setting. Other findings of note are that the mean stopping wave for Condition 2 is somewhat less than Condition 4, and that phase capacity tends to be detected earlier when the ε_i terms are governed by less variability.

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 4b, the mean stopping point is almost always the tenth (and final) wave. The most probable explanation for this difference observed across sample sizes is that a larger sample size results in more precision for the underlying estimates of $\text{var}\left(\hat{\bar{y}}_1^{k-1}\right)$, $\text{var}\left(\hat{\bar{y}}_1^{k}\right)$, and, therefore, $\text{var}\left(\hat{\delta}_{k-1}^{k}\right) = \text{var}\left(\hat{\bar{y}}_1^{k-1}\right) + \text{var}\left(\hat{\bar{y}}_1^{k}\right) - 2\text{cov}\left(\hat{\bar{y}}_1^{k-1}, \hat{\bar{y}}_1^{k}\right)$. Considering these terms comprise the denominator of the phase capacity test statistic quotient, it follows that this would render one more likely to *fail* to reject the test.

Another finding that emerges from comparing 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 the MI version. 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 $\text{var}\left(\hat{\delta}_{k-1}^{k}\right)$. This is confirmed by Figure 3, which overlays the two average values of $\text{var}\left(\hat{\delta}_{k-1}^{k}\right)$ at each wave threshold over all 1,000 iterations of each simulation condition where $n = 500$ and $\varepsilon_i \sim N(0,1)$. One can observe how the variance is consistently smaller for the weighting version until the two converge near the final wave threshold.

Recall how both tests avoid explicit calculation of the term $\text{cov}\left(\hat{\bar{y}}_1^{k-1}, \hat{\bar{y}}_1^{k}\right)$ embedded within $\text{var}\left(\delta_{k-1}^{k}\right) = \text{var}\left(\hat{\bar{y}}_1^{k-1}\right) + \text{var}\left(\hat{\bar{y}}_1^{k}\right) - 2\text{cov}\left(\hat{\bar{y}}_1^{k-1}, \hat{\bar{y}}_1^{k}\right)$. 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 estimate must be attributable to $\text{cov}\left(\hat{\bar{y}}_1^{k-1}, \hat{\bar{y}}_1^{k}\right)$. It appears the covariance from the weighting version of the phase capacity test is larger in magnitude than the covariance calculated using the MI version.

5. Federal Employee Viewpoint Survey Application

We next discuss an application of the two phase capacity tests using actual survey data from three example agencies participating in FEVS 2011. The respondent counts by wave

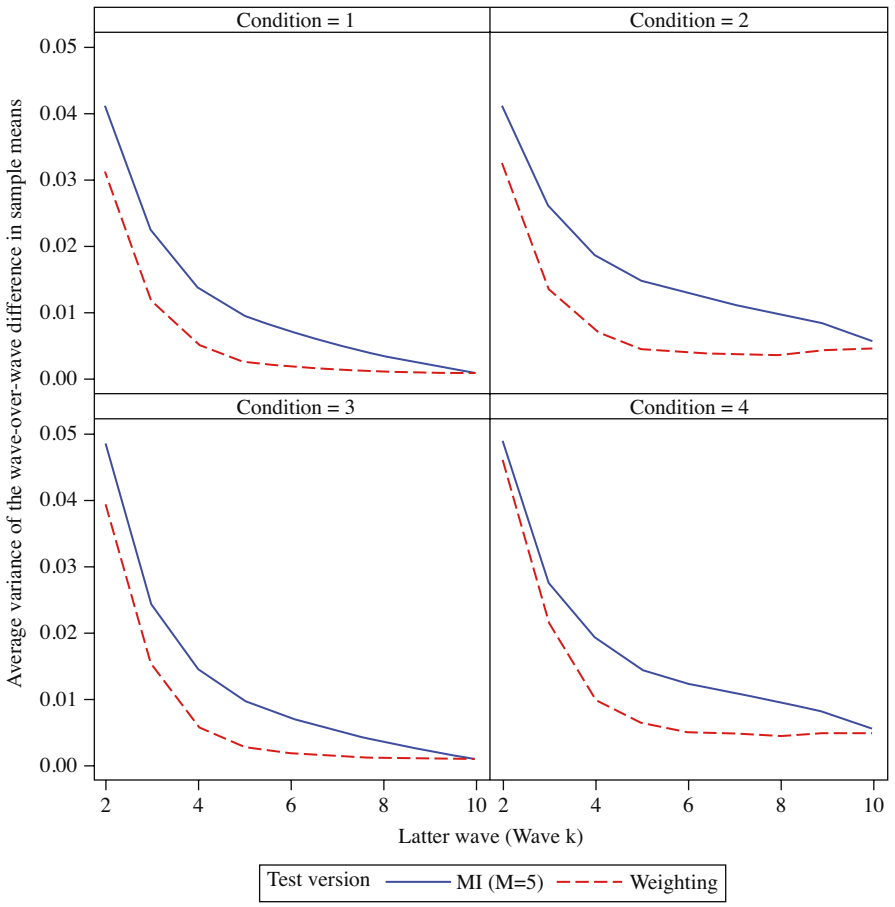

Fig. 3. Average estimated variance of the difference between two adjacent wave sample means by phase capacity testing method for the simulation study setting where $n = 500$ and $\varepsilon_i \sim N(0,1)$.

for these three agencies' were summarized in Table 1. As before, the point estimates investigated are sample means – namely, the seven percent positive estimates for items constituting the Job Satisfaction Index of the Human Capital Assessment and Accountability Framework (HCAAF) (U.S. Office of Personnel Management 2015, 28). This subset of items was chosen because we felt it best captures the essence of the FEVS, an 84-item survey tapping at various dimensions of an employees' overall satisfaction level with his or her job.

The interpretation of nonresponse error is different in the 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 condition, and it was further assumed that a 100% response rate could be achieved with enough follow-up attempts. This was not necessarily realistic, but permitted a gold standard upon which to benchmark the point estimates derived from the abridged sample data sets. Here, given the retrospective nature of the application, nonresponse error is the difference between the point estimate computed once phase capacity has been declared and the like computed at the conclusion of the agency's FEVS 2011 field period.

As with the simulation study, the fundamental objective was to evaluate the performance of the two competing tests of phase capacity. To foster a balanced comparison, a shared set of five auxiliary variables was used in both nonresponse adjustment procedures: (1) agency-subelement; (2) an indicator of whether the employee works at the agency headquarters or in a field office; (3) gender; (4) a minority/non-minority indicator variable; and (5) supervisory status (non-supervisor, supervisor, and executive). For the MI 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. At the end of each wave, the seven positive/non-positive indicators for were multiply imputed $M = 5$ times using the %IMPUTE module within IVEware (http://www.isr.umich.edu/src/smp/ive/), a free, SAS-callable set of macros developed by researchers at the Institute for Social Research at the University of Michigan. The %IMPUTE module implements the Sequential Regression Multiple Imputation (SRMI) algorithm discussed in Raghunathan et al. (2001). For the weighting version of the phase capacity test, base weights of respondents at the end of a given wave were raked to marginal, agency-level totals aggregated from the sampling frame. The totals were derived from the same set of categorical variables serving as main effects in the imputation models used in the MI version. The SAS macro developed by Izrael et al. (2000) was used to carry out the raking process. As in the simulation study, Taylor series linearization was utilized to estimate variances of the adjacent-wave weighted mean differences.

Table 5 summarizes 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 that 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. 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 using $M = 5$ should not substantively alter the main findings.

In many respects, the conclusions to be gleaned from Table 5 coincide with the main takeaways from the simulation study. The weighting version of the test tends to necessitate more wave of nonresponse follow-up than the MI version, which surpasses the second wave only in a few instances. Due to the proclivity of the FEVS percent positive estimates to increase with each additional wave (i.e., as demonstrated in Figure 1), it is of little surprise to observe that the nonresponse error is smaller for the weighting test. The differences are relatively small, however. For example, the average percentage point difference for the seven estimates analyzed for Agency 1 is -1.4. This is the largest average difference for any of the three agencies examined. Still, 1 to 2 percentage points is enough to declare a statistically significant change relative to the previous years' survey.

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.

Table 5. Results from a FEVS 2011 application using data from three example agencies to compare the two phase capacity tests.

	MI ($M = 5$)			Weighting		
Item	Stopping wave	Point estimate	Relative NR error	Stopping wave	Point 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

6. Discussion

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 sampling units respond in the first recruitment attempt, and a sequence of follow-up attempts typically ensues. Survey sponsors often sanction these to continue indefinitely in pursuit of a target response rate or minimum respondent count, but this is not necessarily guaranteed to reduce nonresponse error. Groves and Heeringa (2006) encourage practitioners to employ paradata and other real-time information to help guide decisions about the data collection process and, in particular, when to transition to a new design phase. They defined the notion of phase capacity as the point during a design phase when point estimates stabilize. Absent in their article, however, is a clearly defined, calculable rule practitioners can follow to determine whether phase capacity has been reached. The primary objective of this article was to compare and contrast two techniques to do so, one based on multiple imputation to handle unit nonresponse and one based on weighting.

Evaluating the two tests via a simulation study and an application using actual survey data from the 2011 FEVS, the weighting version was found to be more sensitive to sample

mean changes and, thus, more conservative in declaring phase capacity. It was argued this is due to the two tests' implicit accounting of the covariance between the two adjacent-wave means. We leave for further research the task of developing a more formal theoretical understanding as to why the covariance is not accounted for equivalently in the two tests. Further research could also study the behavior of the weighting version of the phase capacity test for alternative point estimates or variance estimation methods. Although an admittedly cursory analysis indicated certain replication approaches mirrored the performance of the Taylor Series linearization approach utilized exclusively in this research, a more rigorous study would help to rule out any potential anomalies.

The two tests described in this article share two general limitations. One is that they are inherently univariate, meaning they focus on a single point estimate. It is not immediately obvious how one would proceed if a phase capacity test was conducted on two or more point estimates, which could take the form of the same point estimate computed for two or more population domains, resulting in contradictory conclusions. The ideas of the weighting test for phase capacity are extended to multivariate settings such as this in Chapter 4 of Lewis (2014) and Lewis (2015). The second general criticism is that the two tests are retrospective in nature. Knowing the most recent wave's data did not significantly alter 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. To be sure, more research is needed on prospective tests for phase capacity.

Despite our aversion to the phrase "stopping rule," a limitation of this article is that the sole design phase transition considered was, in fact, terminating the nonrespondent follow-up process. More research is warranted to understand how these techniques perform under alternative design phase changes such as changing the incentive or switching modes. Another intriguing idea, at least for an annual survey like the FEVS, would be to research whether information from prior administrations could be somehow incorporated into tests for phase capacity.

In closing, we feel compelled to acknowledge that the actual adoption of a phase capacity testing approach to guide the FEVS data collection process would face headwinds. One reason is the survey administration team's dogma that each agency should be treated equitably. In FEVS 2011 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 FEVS 2012, 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. Another point of contention for a shortened data collection period is the tendency for point estimates to increase over time, as shown in Figure 1, even after adjusting for nonresponse. From the perspective of an agency's senior leadership and stakeholders, higher scores are more desirable, as they are indicative of a more satisfied, engaged workforce. Gaining an extra one-half percentage point, say, even if failing to be a statistically significant change, could be enough to increase the agency's standing in the Partnership for Public Service's Best Places to Work in the Federal Government® rankings (www.bestplacestowork.org).

Appendix

Visualization of the Rao, Glickman, and Glynn (2008) multiple imputation test for phase capacity.

	Observed data				Completed data sets using wave $k-1$ respondents			Completed data sets using wave k respondents			Difference variables		
Sampling unit 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	?	?		?	?		?	?		?

7. References

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

Bates, N. and K. Creighton. 2000. "The Last Five Percent: What Can We Learn from Difficult/Late Interviews?" In Proceedings of the Section on Government Statistics and the Section on Social Statistics: American Statistical Association, 2000: 120–125, Washington, DC. Available at: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/IHSNG-asa2000proceedings.pdf (accessed April 2017).

Beaumont, J., C. Bocci, and D. Haziza. 2014. "An Adaptive Data Collection Procedure for Call Prioritization." *Journal of Official Statistics* 30: 607–621. Doi: http://dx.doi.org/10.2478/jos-2014-0040.

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: Wiley.

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

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

Brick, J.M. and D. Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys." *The Annals of the American Academy of Political and Social Science* 645: 36–59. Doi: http://dx.doi.org/10.1177/0002716212456834.

Couper, M. 1998. "Measuring Survey Quality in a CASIC Environment." In Proceedings of the Section on Survey Research Methods: American Statistical Association, 1998: 41–49, Washington, DC.

Curtin, R., S. Presser, and E. Singer. 2000. "The Effect of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64: 413–428. Doi: http://dx.doi.org/10.1093/poq/nfi002.

Curtin, R., S. Presser, and E. Singer. 2005. "Changes in Telephone Survey Nonresponse over the Past Quarter Century." *Public Opinion Quarterly* 64: 87–98. Doi: http://dx.doi.org/10.1093/poq/nfi002.

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

de Leeuw, E. and W. de Heer. 2002. "Trends in Household Survey Nonresponse: a Longitudinal and International Comparison." In *Survey Nonresponse*, edited by R. Groves, D. Dillman, J. Eltinge, and R. Little, 41–54. New York, NY: Wiley.

Deming, W. 1953. "On a Probability Mechanism to Attain an Economic Balance between the Resultant Error of Response and Bias of Nonresponse." *Journal of the American Statistical Association* 48: 743–772.

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

El-Badry, M. 1956. "A Sampling Procedure for Mailed Questionnaires." *Journal of the American Statistical Association* 51: 209–227. Doi: http://dx.doi.org/10.1080/01621459.1956.10501321.

Elliott, M., R. Little, and S. Lewitzky. 2000. "Subsampling Callbacks to Improve Survey Efficiency." *Journal of the American Statistical Association* 95: 730–738. Doi: http://dx.doi.org/10.2307/2669453.

Filion, F. 1976. "Exploring and Correcting for Nonresponse Bias Using Follow-Ups of Nonrespondents." *Pacific Sociological Review* 19: 401–408. Doi: 10.2307/1388756.

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

Graham, J., A. Olchowski, and T. Gilreath. 2007. "How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory." *Prevention Science* 8: 206–213. Doi: http://dx.doi.org/10.1007/s11121-007-0070-9.

Groves, R. and S. Heeringa. 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: 439–457. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

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

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

Izrael, D., D. Hoaglin, and M. Battaglia. 2000. "A SAS Macro for Balancing a Weighted Sample." In Proceedings of the SAS Users Group International (SUGI) Conference, Cary, NC: SAS Institute Inc. 1350–1355. Available at: http://www2.sas.com/proceedings/sugi25/25/st/25p258.pdf (accessed April 2017).

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

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

Keeter, S., C. Kennedy, M. Dimock, J. Best, and P. Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly* 70: 759–779. Doi: http://dx.doi.org/10.1093/poq/nfl035.

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

Kreuter, F. and C. Casas-Cordero. 2010. "Paradata." Working Paper 136. RatSWD Working Paper Series. Available at: http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_136.pdf (accessed April 2017).

Lewis, T. 2014. *Testing for Phase Capacity in Surveys with Multiple Waves of Nonrespondent Follow-Up*. PhD Thesis. University of Maryland, College Park. Doi: http://dx.doi.org/10.13016/M2WW46.

Lewis, T. 2015. "Multivariate Tests for Phase Capacity." Paper presented at the 2015 FedCASIC Workshops, Washington, DC. Available at: http://www.census.gov/fedcasic/fc2015/ppt/05_lewis.pdf (accessed April 2017).

Lin, I.-F. and N. Schaeffer. 1995. "Using Survey Participants to Estimate the Impact of Nonparticipation." *Public Opinion Quarterly* 59: 236–258. Doi: http://dx.doi.org/10.1086/269471.

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

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

McPhee, C. and S. Hastedt. 2012. "More Money? The Impact of Larger Incentives on Response Rates in a Two-Phase Mail Survey." In Proceedings from the Federal Committee on Statistical Methodology (FCSM) Research Conference. Washington, DC. Available at: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/Hastedt_2012FCSM_I-A.pdf (accessed April 2017).

O'Quigley, J., M. Pepe, and L. Fisher. 1990. "Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer." *Biometrics* 46: 33–48. Doi: http://dx.doi.org/10.2307/2531628.

Peytchev, A., R. Baxter, and L. Carley-Baxter. 2009. "Not All Survey Effort Is Equal: Reduction of Nonresponse Bias and Nonresponse Error." *Public Opinion Quarterly* 73: 785–806. Doi: http://dx.doi.org/10.1093/poq/nfp037.

Politz, A. and W. Simmons. 1949. "An Attempt to Get the Not-at-Homes into the Sample without Callbacks." *Journal of the American Statistical Association* 44: 9–16. Doi: http://dx.doi.org/10.1080/01621459.1949.10483288.

Potthoff, R., K. Manton, and M. Woodbury. 1993. "Correcting for Nonavailability Bias in Surveys Weighting Based on the Number of Callbacks." *Journal of the American Statistical Association* 88: 1197–1207. Doi: http://dx.doi.org/10.1080/01621459.1993.10476399.

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

Rao, R., M. Glickman, and R. Glynn. 2004. "Use of Covariates and Survey Wave to Adjust for Nonresponse." *Biometrical Journal* 46: 579–588. Doi: http://dx.doi.org//10.1002/bimj.200310049.

Rao, R., M. Glickman, and R. Glynn. 2008. "Stopping Rules for Surveys with Multiple Waves of Nonrespondent Follow-Up." *Statistics in Medicine* 27: 2196–2213. Doi: http://dx.doi.org/10.1002/sim.3063.

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

Rubin, D. and N. Schenker. 1986. "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81: 366–374. Doi: http://dx.doi.org/10.1080/01621459.1986.10478280.

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

Schenker, N., T. Raghunathan, P.-L. Chiu, D. Makuc, G. Zhang, and A. Cohen. 2006. "Multiple Imputation of Missing Income Data in the National Health Interview Survey." *Journal of the American Statistical Association* 101: 924–933. Doi: http://dx.doi.org/10.1198/016214505000001375.

Schouten, B. and N. Schlomo. 2015. "Selecting Adaptive Survey Design Strata with Partial R-Indicators." *International Statistical Review*, Online First Edition. Doi: http://dx.doi.org/10.1111/insr.12159.

Sigman, R., T. Lewis, N. Yount, and K. Lee. 2014. "Does The Length of Fielding Period Matter? Examining Response Scores of Early versus Late Responders." *Journal of Official Statistics* 30: 651–674. Doi: https://doi.org/10.2478/jos-2014-0042.

Tourangeau, R. and T. Plewes (eds.). 2013. *Nonresponse in Social Science Surveys: A Research Agenda*. Washington, DC: The National Academies Press. Available at: http://www.nap.edu/read/18293/chapter/1 (accessed April 2017).

Valliant, R. 2004. "The Effect of Multiple Weighting Steps on Variance Estimation." *Journal of Official Statistics* 20: 1–18.

U.S. Office of Personnel Management (OPM). 2015. "2015 Federal Employee Viewpoint Survey Technical Report." Available at: http://www.fedview.opm.gov/2015FILES/2015_OPM_Technical_Report.pdf (accessed April 2017).

Wagner, J. 2008. *Adaptive Survey Design to Reduce Nonresponse Bias*. PhD Thesis. University of Michigan, Ann Arbor.

Wagner, J. 2010. "The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data." *Public Opinion Quarterly* 74: 223–243. Doi: http://dx.doi.org/10.1093/poq/nfq007.

Wagner, J. and T. Raghunathan. 2010. "A New Stopping Rule for Surveys." *Statistics in Medicine* 29: 1014–1024. Doi: http://dx.doi.org/10.1002/sim.3834.

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: 411–414. Doi: http://dx.doi.org/10.2307/2283947.

Received March 2016
Revised April 2017
Accepted May 2017