Survey Research Methods (2019)
Vol. 13, No. 2, pp. 153-165
doi:10.18148/srm/2019.v13i2.7370

ISSN 1864-3361
http://www.surveymethods.org

Multivariate Tests for Phase Capacity

Taylor Lewis
RTI International
Washington, DC, U.S.A.

To combat the potentially detrimental 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 static data collection protocol. Consequently, (nonresponse-adjusted) point estimates calculated from the accumulating data begin to stabilize. This is the notion of phase capacity, suggesting some form of design change is warranted, such as switching modes, increasing the incentive, or simply discontinuing nonrespondent follow-up.Phase capacity testing methods that have appeared in the literature to date are generally only applicable to a single point estimate. It is unclear how to proceed if conflicting results are obtained following independent tests on two or more point estimates. The purpose of this paper is to introduce two multivariate phase capacity tests, one referred to as the Wald chi-square method and another referred to as the non-zero trajectory method. Both methods are designed to provide a universal, yes-or-no phase capacity determination for a battery of point estimates. The two competing methods' performance is compared via simulation and application using data from the 2011 Federal Employee Viewpoint Survey. All else equal, the Wald chi-square method is found to detect phase capacity sooner than the non-zero trajectory method.

Keywords: responsive survey design, adaptive design, stopping rules, nonresponse

1 Background

Unit nonresponse, which occurs when sampled individuals do not respond to a survey, is a widespread problem in social surveys. Response rates have been declining in surveys worldwide for the past several decades (Atrostic, Bates, & Silberstein, 2001; Beullens, Loosveldt, Vandenplas, & Stoop, 2018; Brick & Williams, 2013; Curtin, Presser, & Singer, 2005; de Leeuw & de Heer, 2002; Silver, 2014; Tourangeau & Plewes, 2013; Williams & Brick, 2017). Typically, a survey's data collection protocol involves making a sequence of follow-up attempts on nonrespondents, which can take on various forms depending on the survey's mode – for example, mailing a replacement questionnaire, placing a follow-up telephone call, or revisiting a residence. Each follow-up attempt tends to produce more survey completes, which have been referred to in the literature as incoming *waves* of data (e.g. Lewis, 2014; Rao, Glickman, & Glynn, 2008; Wagner & Raghunathan, 2010). On the surface, more follow-ups are desirable, as they increase the response rate, but they do not guarantee a reduction in nonresponse error (Lewis, 2017a). Additionally, they may be costly and can extend the data collection period, causing a delay in the reporting and analysis stages of the survey. Moreover, from a purely practical vantage point, empirical evidence (e.g., Potthoff, Manton, & Woodbury, 1993, Table 1; Lewis, 2017b, Table 1) suggests returns taper with each subsequent wave. Fewer and fewer new responses are obtained, in turn impacting point estimates less and less.

In an effort to reign in the increased data collection costs associated with the decline in response rates, Groves and Heeringa (2006) urge practitioners to adopt principles of *responsive survey design*, which Bethlehem, Cobben, and Schouten (2011) classify as a special case of *adaptive survey design* (Schouten, Peytchev, & Wagner, 2017; Wagner, 2008). The notion of adaptive survey design is to tailor features of the survey's data collection protocol to specific sample cases, in an acknowledgment that not all cases react to and weigh these features uniformly when deciding whether or not to participate (Groves, Singer, & Corning, 2000). When implemented using real-time information such as *paradata* (Couper, 1998; Kreuter, 2013), or data generated during the process of collecting survey data (e.g., interviewer observations, time stamps), it is referred to as *dynamic* adaptive survey design; when implemented based on historical information, it is referred to as *static* adaptive design.

Similarly in spirit to dynamic adaptive survey designs, the premise of responsive survey designs is to utilize real-time information to help inform data collection decisions and, if

Contact information: Taylor Lewis, Research Triangle Institute, 701 13th St, N.W., Suite 750, Washington, DC 20005-3967, email: thlewis@rti.org.

necessary, change course. One key differentiator, however, is that alternative interventions are implemented across two or more sequential *design phases*, which Groves and Heeringa (2006) define as mutually exclusive segments of the survey's overall data collection period with a fixed sampling frame and recruitment protocol. They term *phase capacity* the point during a design phase at which the additional responses cease influencing key statistics that have been adjusted for nonresponse in some way, such as via weighting or imputation. In lieu of terminating data collection or transitioning to a new design phase at some arbitrary threshold such as a target response rate, they recommend monitoring nonresponse-adjusted point estimates and intervening once phase capacity has been reached. But, as noted by Wagner and Raghunathan (2010), Groves and Heeringa (2006) stopped short of providing a formal, calculable rule or test for phase capacity. The concept is only illustrated visually in Figure 2 of their paper, in which they plot the trend line of one such point estimate derived from the National Survey of Family Growth over the course of the data collection period and comment on how it stabilizes well in advance of the design phase conclusion.

Lewis (2017b) compares two phase capacity testing methods that have emerged in the years following Groves and Heeringa's seminal paper, one based on multiple imputation for nonresponse (Rao et al., 2008; Rubin, 1987) and another based on weighting for nonresponse (Kalton & Flores-Cervantes, 2003; Lewis, 2014). A noted limitation of those methods, however, is that they are univariate in nature. They are designed to test whether a nonresponse-adjusted point estimate derived from data received between waves 1 and $k(k \geq 2)$, $\hat{\theta}_1^k$, differs significantly from the like derived using data received between waves 1 and $k-1$, $\hat{\theta}_1^{k-1}$. The assessment is based on a two-sample t test with underlying statistical hypotheses H_0: $\delta_{k-1}^k = \theta_1^{k-1} - \theta_1^k = 0$ vs. H_1: $\delta_{k-1}^k = \theta_1^{k-1} - \theta_1^k \neq 0$ and some prescribed significance level, typically $\alpha = 0.05$. In other words, the objective is to determine whether the observed point estimate change following receipt of the k^{th} wave's data, $\hat{\delta}_{k-1}^k = \hat{\theta}_1^{k-1} - \hat{\theta}_1^k$, is significantly different from 0. Once it is not, phase capacity is declared. While the approach is intuitive, as detailed in Section 3 of Lewis (2017b), the complex aspect is deriving an appropriate variance of $\hat{\delta}_{k-1}^k$ accounting for the covariance attributable to the fact that both point estimates use data from responses obtained between waves 1 and $k-1$.

Naturally, survey practitioners may not wish to limit focus on a single point estimate, but instead may be focused on a battery of $D(\geq 2)$ point estimates and their associated differences across adjacent waves of data collection. The battery could be comprised of distinct point estimates, separate population domains of interest for the same point estimate, or some combination of both. Although one could conduct a phase capacity test independently on each of the D differences, it is unclear how to proceed in the presence of contradictory results. For instance, suppose the test was conducted on $D = 3$ unique sample means. What is the decision on phase capacity when one mean changed significantly after incorporating the most recent wave's data, but not the other two? The purpose of this paper is to introduce two multivariate phase capacity tests to provide a single, yes-or-no determination for situations such as these. As will be noted, it is also possible to amend the tests to ascribe differential importance to the various point estimate changes.

This paper is organized as follows. Section 2 provides details regarding how to conduct the two phase capacity testing methods. In Section 3, the concept of multivariate phase capacity is illustrated by examining patterns in data collected as part of the U.S. Office of Personnel Management's 2011 Federal Employee Viewpoint Survey (www.opm.gov/fevs). Sections 4 and 5 present results from a simulation study and application, respectively, designed to compare and contrast the two phase capacity testing methods' performance. Section 6 concludes with a brief summary of the paper's key findings and outlines ideas for further research.

2 Methods

2.1 Wald Chi-Square Method

To facilitate exposition of the first multivariate phase capacity test, we must first introduce some matrix notation. Let $\boldsymbol{D}$ represent a $D \times 1$ vector of estimated point estimate differences as follows:

$$\boldsymbol{D} = \begin{bmatrix} \hat{\delta}_{(k-1)1}^k \\ \hat{\delta}_{(k-1)2}^k \\ \vdots \\ \hat{\delta}_{(k-1)D}^k \end{bmatrix} \tag{1}$$

One can conceptualize $\boldsymbol{D}$ as an estimate of Δ_{k-1}^k, a $D \times 1$ vector comprised of the unknown differences of interest $\delta_{(k-1)d}^k = \theta_{1d}^{k-1} - \theta_{1d}^k$. Furthermore, let $\boldsymbol{S}$ denote a symmetric $D \times D$ matrix with the D difference-specific variances terms along the diagonal and difference-to-difference covariances in the off-diagonal as follows:

$$\boldsymbol{S} = \begin{bmatrix} \text{var}(\hat{\delta}_{(k-1)1}^k) & \cdots & \text{cov}(\hat{\delta}_{(k-1)1}^k, \hat{\delta}_{(k-1)D}^k) \\ \vdots & \ddots & \vdots \\ \text{cov}(\hat{\delta}_{(k-1)D}^k, \hat{\delta}_{(k-1)1}^k) & \cdots & \text{var}(\hat{\delta}_{(k-1)D}^k) \end{bmatrix} \tag{2}$$

To illustrate how one would populate the terms in $\boldsymbol{D}$ and $\boldsymbol{S}$, suppose that a practitioner was interested in knowing whether $\hat{\bar{y}}_1^k$, an estimated mean based on data from waves 1 through wave k is significantly different from $\hat{\bar{y}}_1^{k-1}$, the like using data only through wave $k-1$. Let us assume that the two estimated means are weighted for the n cases in the sample by w_1^k and w_1^{k-1}, the nonresponse-adjusted weights

computed at the conclusion of waves k and $k-1$, respectively. Note that for cases responding at or before wave $k-1$, both weights would be positive values. For cases responding specifically during wave k, w_1^k would be positive but w_1^{k-1} would be 0. For cases that have yet to respond by wave k, both w_1^k and w_1^{k-1} would be 0. The row in the vector $\boldsymbol{D}$ corresponding to this point estimate would be $\hat{\delta}_{k-1}^k = \hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^k$.

One can employ principles of Taylor series linearization to populate $\boldsymbol{S}$. First, note how the difference can be expressed as a function of $p = 4$ estimated totals, since $\hat{\delta}_{k-1}^k = \hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^k = \frac{\sum_{i=1}^n w_{1i}^{k-1} y_i}{\sum_{i=1}^n w_{1i}^{k-1}} - \frac{\sum_{i=1}^n w_{1i}^{k} y_i}{\sum_{i=1}^n w_{1i}^{k}} = \frac{\hat{Y}_1^{k-1}}{\hat{N}_1^{k-1}} - \frac{\hat{Y}_1^{k}}{\hat{N}_1^{k}} = \frac{\hat{T}_1}{\hat{T}_2} - \frac{\hat{T}_3}{\hat{T}_4}$. When written in this manner, Wolter (2007, Section 6.5) demonstrates how a computational strategy originally proposed by Woodruff (1971) can greatly simplify matters by averting the need to explicitly estimate the $\binom{p}{2}$ covariance terms inherent in the variance estimate of the difference. The first step of the Woodruff strategy is to create a variate u_i equaling the sum of the difference function's partial derivatives multiplied by the corresponding estimated total. In the present case, $\text{var}(\hat{\delta}_{k-1}^k) \approx \text{var}\left(\sum_{i=1}^n \sum_{j=1}^p \frac{\partial \delta_{k-1}^k}{\partial T_j} t_{ji}\right)$, where t_{ji} represents the value of the j^{th} total in the function for the i^{th} primary sampling unit. To be specific, $t_{1i} = w_{1i}^{k-1} y_i$, $t_{2i} = w_{1i}^{k-1}$, $t_{3i} = w_{1i}^{k} y_i$, and $t_{4i} = w_{1i}^{k}$. After a little algebra, it can be shown $\sum_{j=1}^p \frac{\partial \delta_{k-1}^k}{\partial T_j} t_{ji} = u_i = \frac{1}{\hat{N}_1^{k-1}} w_{1i}^{k-1} y_i - \frac{\hat{Y}_1^{k-1}}{\left(\hat{N}_1^{k-1}\right)^2} w_{1i}^{k-1} - \frac{1}{\hat{N}_1^{k}} w_{1i}^{k} y_i + \frac{\hat{Y}_1^{k}}{\left(\hat{N}_1^{k}\right)^2} w_{1i}^{k}$. From this point, the estimated variance of the sum of the u_i's with respect to the sample design approximates $\text{var}(\hat{\delta}_{k-1}^k)$. For a simple random sample of size n, ignoring the finite population correction factor for the moment, this would be $\text{var}\left(\sum_{i=1}^n u_i\right) = \left(\frac{n}{n-1}\right) \sum_{i=1}^n \left(u_i - \frac{\sum_{i=1}^n u_i}{n}\right)^2$. For two sample means, say, indexed by d and d', the covariance term for the corresponding row and column of $\boldsymbol{S}$ matrix would be $\text{cov}\left(\sum_{i=1}^n u_{di}, \sum_{i=1}^n u_{d'i}\right) = \left(\frac{n}{n-1}\right) \sum_{i=1}^n \left(u_{di} - \frac{\sum_{i=1}^n u_{di}}{n}\right)\left(u_{d'i} - \frac{\sum_{i=1}^n u_{d'i}}{n}\right)$. This approach will work for any smooth, differentiable function of totals, which covers a wide range of point estimate differences. For those not meeting these criteria, a replication approach such as the jackknife or bootstrap could instead be employed. See Wolter (2007) for more details on these alternative variance estimation procedures.

The assessment of phase capacity hinges on the statistical hypotheses H_0: $\Delta_{k-1}^k = \mathbf{0}$ vs. H_1: $\Delta_{k-1}^k \neq \mathbf{0}$, where $\mathbf{0}$ is a $D \times 1$ vector of zeros. Phase capacity is declared at the conclusion of the first wave where the null hypothesis is not rejected. This determination is made based on the following Wald chi-square test statistic (Heeringa, West, & Berglund, 2017):

$$\chi_W^2 = \boldsymbol{D}^T \boldsymbol{S}^{-1} \boldsymbol{D} \tag{3}$$

Under the null hypothesis, χ_W^2 is a scalar distributed as a random chi-square variate with $D-1$ degrees of freedom, so one can use that reference distribution to determine an appropriate p-value.

In its basic form, the Wald chi-square method treats each point estimate equivalently, but there may be occasions when a practitioner wishes to assign differential degrees of importance. For instance, suppose the first of $D = 3$ estimates is deemed "most important". The practitioner still seeks an overall test of phase capacity, but would like any determination made to be twice as sensitive to changes in that estimate than the other two. This could be accommodated by introducing a matrix $\boldsymbol{C} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ into the corresponding test statistic as $\chi_W^2 = (\boldsymbol{C}^{TD})^T (\boldsymbol{C}^T \boldsymbol{S} \boldsymbol{C})^{-1} (\boldsymbol{C}^T \boldsymbol{D})$ The reference distribution would be the same with or without this matrix of relative importance weights.

2.2 Non-Zero Trajectory Method

The second multivariate phase capacity tests draws upon fundamentals of visualizing longitudinal data on change, as discussed in Chapter 2 of Singer and Willett (2003), by ascertaining whether there is a non-zero trajectory of change across all D estimates. The first step is to calculate the three most recent wave-over-wave relative percent changes in each of the D nonresponse-adjusted point estimates. Using the relative percent change ensures that all point estimate differences adhere to a common scale – for example, differences in proportions can be compared with differences in totals. One immediately evident difference of this method relative to the Wald chi-square method is that it requires $k \geq 4$, not $k \geq 2$.

The premise of the non-zero trajectory method is to model Δ_d, the d^{th} point estimate's relative percent change, as a simple linear function of the data collection wave. If we let w represent the data collection wave, a predictor variable taking the form of an integer one unit apart (e.g., 0, 1, and 2), and let I_d be a 0/1 indicator variable for the d^{th} point estimate (1 = yes; 0 = no), the following model is estimated:

$$\begin{aligned} \Delta_d = \beta_{01} + \beta_{02} + \ldots + \beta_{0D} + \beta_{11} \times w \times I_1 + \\ \beta_{12} \times w \times I_2 + \ldots + \beta_{1D} \times w \times I_d + \varepsilon_d \end{aligned} \tag{4}$$

Equation 4 can be thought of as a series of D simple linear regression models being fitted simultaneously, one for each of the D point estimates' relative percent change trend over the three most recent waves. The β_{0d} terms represent intercepts, the β_{1d} terms represent slopes, and ε_d represents a residual term. Phase capacity is declared as soon as all

intercept and slope terms in the model are statistically indistinguishable from 0. This determination is made by carrying out an F test based on the following underlying hypotheses H_0:$\beta_{01} = \beta_{02} = ... = \beta_{0D} = \beta_{11} = \beta_{12} = ... = \beta_{1D} = 0$ vs. H_1: at least one of $\beta_{01}, \beta_{02}, .., \beta_{0D}, \beta_{11}, \beta_{12}, ..., \beta_{1D}$ is not equal to 0.

To illustrate the non-zero trajectory method with the help of a simple artificial example, suppose the goal is to determine whether the $D = 3$ survey estimates (percentages) summarized in Table 1 have stabilized following the three most recent waves of nonrespondent follow-up.

In this case, the model would have a total of $2D = 6$ terms, $D = 3$ intercepts and $D = 3$ slopes. The model parameters can be estimated using standard matrix theory of ordinary least squares regression after first creating the outcome vector

$$\boldsymbol{\Delta} = \begin{bmatrix} 0.2 \\ 0.5 \\ 0.4 \\ 0.2 \\ 0.2 \\ 0.3 \\ 0.1 \\ 0.0 \\ 0.2 \end{bmatrix} \tag{5}$$

and design matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{bmatrix} \tag{6}$$

Specifically, one can obtain the estimated model parameters by finding the $2D \times 1$ vector $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Delta}$ and the corresponding $2D \times 2D$ covariance matrix by $\text{cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}_d^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, where $\hat{\sigma}_d^2$ is the estimated mean squared error of the model. Then, the phase capacity test statistic is

$$F = \hat{\boldsymbol{\beta}}^T \left(\text{cov}(\hat{\boldsymbol{\beta}})\right)^{-1} \hat{\boldsymbol{\beta}} \tag{7}$$

One can reference this test statistic against an F distribution with $2D$ numerator degrees of freedom and D denominator degrees of freedom at the desired significance level. Phase capacity is declared once this test statistic fails to be large enough to reject the null hypothesis.

Figure 1 is a visualization of the model being fitted from the data in Table 1; again, the model can be conceptualized as $D = 3$ simple linear regression models, one for each trend

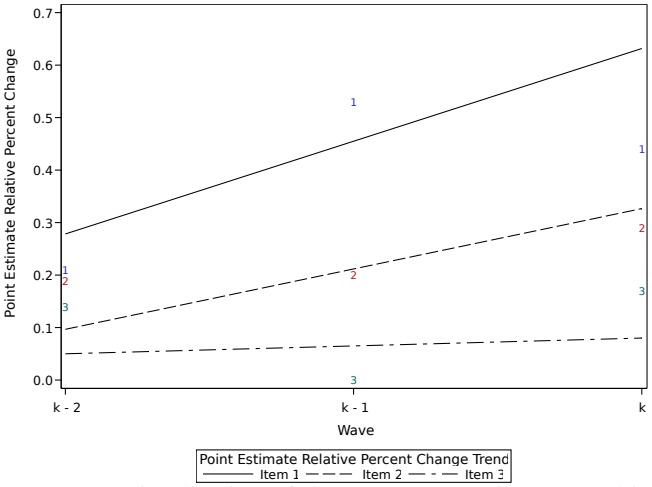

Figure 1. Visualization of the Non-Zero Trajectory Multivariate Phase Capacity Test

in the relative percent change of a given point estimate over the three most recent wave thresholds. Data points in the figure are labeled 1, 2, or 3 to reflect the survey item in Table 1 to which they correspond. All points fall above 0 on the y-axis scale, which is an indication of the increasing trend the point estimates exhibit following the most recent waves of data collection. Item 1 exhibits a more pronounced trend that do Items 2 or 3, but all are modest in magnitude.

As was noted regarding the Wald chi-square method, although each point estimate is treated equivalently by default, one could account for a vector of relative importance weights to assign differential degrees of importance. For instance, suppose the first of $D = 3$ estimates is deemed "most important", and so the practitioner would like any phase capacity determination made to be twice as sensitive to changes in that estimate relative to the other two. This could be accommodated by introducing the matrix

$$\boldsymbol{C} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{8}$$

where entries of 2 along the diagonal correspond to the model parameters of the most important item, and an entries of 1 correspond to the model parameters of the two other items. From here, the test statistic becomes $F = (\boldsymbol{C}^T\hat{\boldsymbol{\beta}})^T(\boldsymbol{C}^T\text{cov}\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{C})^{-1}(\boldsymbol{C}^T\hat{\boldsymbol{\beta}})$, which is still referenced against an F distribution with $2D$ numerator degrees of freedom and D denominator degrees of freedom.

Table 1
Illustrative Example Survey Data with Three Independent Point Estimates Based on Accumulating Data in the Four Most Recent Waves

Wave	Item 1	Item 1 Rel (%) Change	Item 2	Item 2 Rel (%) Change	Item 3	Item 3 Rel (%) Change
k - 3	75.2	–	83.6	–	88.5	–
k - 2	75.3	0.2	83.8	0.2	88.6	0.1
k - 1	75.7	0.5	83.9	0.2	88.6	0.0
k	76.1	0.4	84.2	0.3	88.7	0.2

3 Illustration of Multivariate Phase Capacity in the 2011 Federal Employee Viewpoint Survey

The purpose of this section is to motivate the notion of multivariate phase capacity using data drawn from the 2011 Federal Employee Viewpoint Survey (FEVS). We begin by providing background information on the FEVS. First launched in 2002 as the biennial Federal Human Capital Survey, the FEVS is now an annual organizational climate survey conducted by the U.S. Office of Personnel Management (OPM) on a sample of approximately 1.14 million full- or part-time employees (as of FEVS 2017) representing more than 80 distinct United States government agencies. With very few exceptions, the FEVS sampling frame is derived from a personnel database known as the Statistical Data Mart of the Enterprise Human Resources Integration (EHRI-SDM), a rich auxiliary data source containing numerous demographics and a detailed history of one's employment with the Federal government. Work-unit-level information is used to stratify the sampling frame in an effort to ensure sufficient responses are obtained for reports broken out for particular divisions of interest within the agency. For more detail on the FEVS sampling procedures, see pp. 2-3 of U.S. Office of Personnel Management (2016).

The FEVS instrument is comprised predominantly of attitudinal questions tapping at a diverse range of satisfaction dimensions and employee perceptions, such as one's level of enjoyment with the kind of work performed, opportunities for advancement, and confidence in senior leadership within the agency. Most survey questions are posed as statements to which respondents are asked to rate their level of agreement using a five-point Likert-type scale, such as one ranging from "Completely Agree" to "Completely Disagree", often with an explicit "Do Not Know" (DNK) or "No Basis to Judge" (NBTJ) option given. Statistical significance testing is often conducted after first dichotomizing an item's responses into either a positive or non-positive response, with a DNK or NBTJ election treated as missing. Specifically, if we let y_{id} denote a 0/1 indicator variable of a positive response for the i^{th} respondent to the d^{th} item and w_i denote the nonresponse-adjusted weight affixed to that respondent, then the so-called *percent positive estimate* for that item is defined as

$$\hat{p}_d = \frac{\sum_{i \in R_d} w_i y_{id}}{\sum_{i \in R_d} w_i} \times 100 \tag{9}$$

where $i \in R_d$ signifies the set of substantive responses, ignoring item nonresponse and questions for which a DNK or NBTJ response was given, both of which typically amount to less than 5% of cases for any particular item.

In addition to item-level summaries, thematically-linked groupings of FEVS items are used to form indices. Each index is computed as the average of a set of percent positive estimates, or

$$\hat{I} = \frac{1}{D} \sum_{d=1}^{D} \hat{p}_d \tag{10}$$

where D is the number of items comprising the index. One high-profile example is the Human Capital Assessment and Accountability Framework (HCAAF) established in the Chief Human Capital Officers Act of 2002 that led to the creation of four widely reported indices: (1) Leadership and Knowledge Management; (2) Results-Oriented Performance Culture; (3) Talent Management; and (4) Job Satisfaction. The item numbers, wording, and response scales for each of these four HCAAF indices are given in the Appendix.

The FEVS is a Web-based survey. On the first day of the fielding period, sampled individuals are emailed an invitation to participate with a personalized URL to access the survey. Thereafter, weekly reminder emails are sent to nonrespondents. The final reminder contains wording indicating that the survey closes at the end of the day. In FEVS 2011, agencies were given some leeway in determining their fielding period duration, but the median fielding period duration was eight weeks. As reported on p. 4 of U.S. Office of Personnel Management (2017), the response rate to the survey has oscillated between 45-50% in recent administrations. To compensate for unequal sampling probabilities and unit nonresponse, a three-stage weighting procedure is implemented exploiting EHRI-SDM variables known for the entire sample, such as gender, supervisory status, race/ethnicity, and

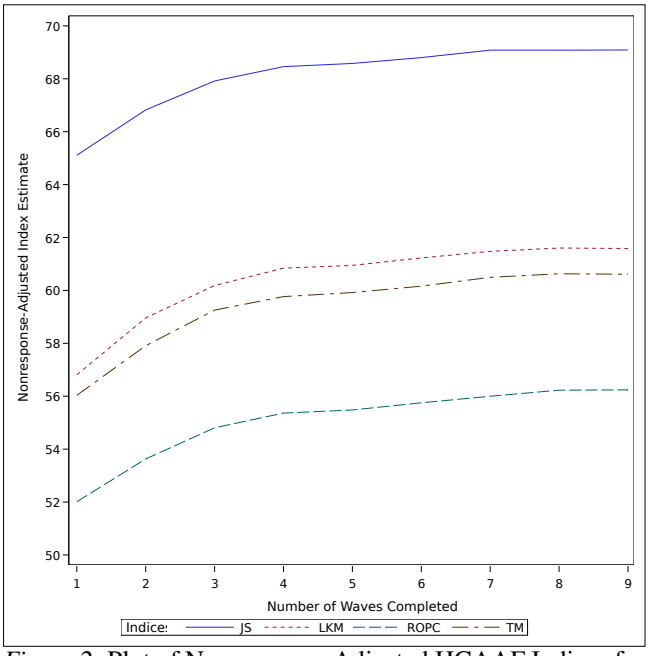

Figure 2. Plot of Nonresponse-Adjusted HCAAF Indices for an Example Agency Participating in the 2011 FEVS Using Accumulating Data as of the Given Data Collection Wave

tenure with the Federal government. A more detailed description of the FEVS weighting procedures can be found in Appendix E of U.S. Office of Personnel Management (2016).

As shown in Table 1 of Lewis (2017b), the first email invitation tends to generate the most completes, with returns steadily diminishing thereafter. As a result, percent positive estimates and, thus, indices, change less and less with each subsequent reminder, eventually settling into a state of phase capacity. Figure 2 illustrates this phenomenon for one agency participating in the 2011 FEVS. The figure shows the trend of nonresponse-adjusted HCAAF indices over data collection waves, defined here as responses obtained between two adjacent reminder emails. The trend for each index is upward, which is a reflection of early respondents tending to be less positive than the more reluctant respondents, echoing findings from Sigman, Lewis, Yount, and Lee (2014). Auxiliary variables from EHRI-SDM used to weight for unit nonresponse are unable to completely eradicate the trend, although changes are trivial after about wave 4 or 5.

4 Simulation Study

In this section of the paper, we present results from a simulation study conducted to compare and contrast the performance of the two proposed multivariate phase capacity tests. Rather than simulating data using one or more parametric distributions, we exploited actual data from the 2011 FEVS and the observed patterns of response. In particular, using data from the same three agencies as in Lewis (2017b), we treat the ultimate set of 2011 FEVS respondents from these three agencies as if they were complete samples, respectively, enabling an evaluation of (relative) nonresponse error. Agency 1 consists of a sample size of $n_1 = 8,105$ individuals, Agency 2 of $n_2 = 572$ individuals, and Agency 3 of $n_3 = 8,687$ individuals.

For each of 1,000 independent simulations, a response wave between 1 and 10 was randomly assigned to each individual based on one of two conditions summarized in Table 2. For Condition 1, an individual's response wave was independently generated using the same wave-specific respondent proportions of Agency 1 as shown in Table 1 of Lewis (2017b). A respondent was assigned as responding in wave 1 with probability 0.251, wave 2 with probability 0.175, and so on. For Condition 2, an individual's response wave was simulated in such a way that earlier respondents tended to be less positive, as determined by that individuals' responses to the $D = 7$ items comprising the Job Satisfaction Index. In particular, respondents were partitioned into two groups of roughly equal size based on an aggregate measure of their degree of satisfaction with respect to the index. This was accomplished by converting each response to the Likert-type scale into integers between 1 and 5 such that a 1 represented the most negative response (e.g., Very Dissatisfied) and a 5 represented the most positive response (e.g., Very Satisfied). The seven integers were then summed at the respondent level to create an aggregate measure of satisfaction ranging from a minimum of 7 (7×1) to a maximum of 35 (7×5). Two classes of respondents were then defined: (1) less satisfied respondents, or those respondents those whose aggregate measure fell below the median; and (2) more satisfied respondents, those whose aggregate measure fell above the median. An independently generated random uniform variate between 0 and 1 was first added to each aggregate measure to eliminate the possibility of ties and produce two groups of approximately equal size.

Despite being somewhat ad-hoc, we felt this classification scheme met the principal objective of simulating a scenario in which the outcome variables were associated with the response wave. To provide a few numbers with respect to the specifications given in Table 2, the less satisfied respondents were assigned wave 1 with probability 0.345, and the more positive respondents were assigned wave 1 with probability 0.156. These percentages were designed such that the expected marginal percentage of wave 1 respondents in the whole of Condition 2 matches that of Condition 1 – for example, $0.5 \times (34.5 + 15.6) \approx 25.1\%$.

The bifurcation of respondents based on the aggregate measure of satisfaction was not performed overall or by agency; rather, it was performed within one of 12 classes defined by the cross-classification of agency, minority status, and supervisory status. These 12 categorizations were also

Table 2
Summary of the Two Response Wave Distributions Used for the Simulation Study Comparing the Two Multivariate Phase Capacity Tests

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

used as weighting classes (Brick & Kalton, 1996) for Conditions 1 and 2. Within the c^{th} class, the sum of weights for respondents at the conclusion of each simulated wave was calibrated such that it matched the known population total within the class, N_c. For the Wald chi-square testing method, variance estimates were produced using Woodruff's (1971) method for Taylor series linearization as outlined in Section 2.1.

Results from the simulation study are summarized in Table 3. Using a significance level of $\alpha = 0.05$, the measure labeled "Mean Stop Wave" represents the average data collection wave at which phase capacity was declared over all 1,000 iterations. The standard deviation of this average is also reported. The measure labeled "Mean Nonresponse Error for Index" is the average magnitude of nonresponse error in the Job Satisfaction Index at the point phase capacity was determined, which, per Equation 10, can also be interpreted as the mean nonresponse error amongst the seven percent positive estimates comprising the index. Beneath that measure is the root mean squared error (RMSE) of the index at the point of phase capacity, averaged over all 1,000 simulations, where the RMSE is defined as square root of the sum of the following two quantities: (1) the nonresponse error of the index squared; and (2) the approximated variance of the index, which was derived via Taylor series linearization as detailed in Lewis (2012). The final quantity reported is the percentage of 95% confidence intervals formed about the index at the point phase capacity was declared that encompassed the index as calculated from the full sample.

In Condition 1, the first key finding is that both methods tend to detect phase capacity at their respective earliest possible points to do so: the second wave for the Wald chi-square method and the fourth for the non-zero trajectory method. For example, the mean stopping wave for Agency 1 was 2.05 for the former method and 4.16 for the latter. There is scant differentiation amongst the three agencies investigated for any particular method, but the non-zero trajectory method appears to exhibit more variability in the mean stopping wave relative to the Wald chi-square method. Not surprisingly, there is very little nonresponse error in the index introduced by curtailing the data collection period in Condition 1. Additionally, confidence intervals formed around the index estimated once phase capacity was first reached almost always cover the index value that would be obtained once responses for all sampled cases is obtained.

In Condition 2, the expected values of the seven percent positive estimates (and thus the index) were predisposed to increase with each subsequent wave of data incorporated. To the extent that the employees' varying degrees of satisfaction are not completely explained by the cross-classification of agency, minority status, and supervisory status, the variables used in the weighting class adjustment procedure, we would anticipate some degree of nonresponse error associated with stopping data collection early. Indeed, this is plainly observed in Table 3. Despite both methods generally calling for more than the absolute minimum number of waves, they often detect phase capacity prior to the tenth wave and, as such, are susceptible to nonresponse error and a decreased likelihood that the confidence interval formed about the index using the abridged data set contains the full-sample index value.

Interestingly, at least for Condition 2, both methods de-

clare phase capacity earlier for Agency 2 than the other two agencies. Under the Wald chi-square approach, the mean stopping wave for Agency 2 is 2.13, in contrast to 6.84 and 6.12 for Agency 1 and 3, respectively. This is coupled with a much larger mean nonresponse error over the 1,000 simulations. The value for Agency 2 (-5.76) is roughly 3 times the magnitude for Agency 1 (-1.55) and Agency 3 (-2.01). A similar story emerges comparing the 95% confidence interval coverage rates. A possible explanation is that the sample size for Agency 2($n_2 = 572$) is much smaller than the sample sizes for the other two agencies, both of which exceed 8,000. Similarly to what was concluded in a simulation study reported in Lewis (2017b), all else equal, a smaller sample size leads to a quicker determination of phase capacity.

5 Application

It was assumed in the simulation study design that nonresponse error can be extirpated altogether given enough waves of nonrespondent follow-up. Although this is not necessarily a realistic assumption, it enabled a comprehensive comparison of the two methods' performance. In this section, we evaluate the two methods via an application using the unaltered survey data and response patterns for the three example agencies participating in the 2011 FEVS. Moreover, instead of focusing exclusively on the seven items underlying the Job Satisfaction Index, we extend our investigation to include the other three HCAAF indices as well. To conduct wave-specific nonresponse adjustments, the SAS® macro %RAKING developed by Izrael, Hoaglin, and Battaglia (2000) was used to calibrate the weights of employees in the accumulating respondent sets such that they summed to known agency employment totals of the first level of work unit below agency, an indicator of whether the employee works at headquarters or in a field office, a minority status indicator, gender, and supervisory status (non-supervisor, supervisor, or executive).

Table 4 summarizes results from the 2011 FEVS application. The column labeled "Stop Wave" reports the wave at which phase capacity would be declared, as before using a significance level of $\alpha = 0.05$. This is flanked by the corresponding nonresponse-adjusted estimate of the given index and relative nonresponse error, where applicable. We say "where applicable" because phase capacity was not always declared prior to the final wave of data collection (e.g., LKM, ROPC, TM indices for Agency 1 under the non-zero trajectory method).

A ubiquitous finding is that the Wald chi-square method tends to declare phase capacity much sooner than the non-zero trajectory method. Indeed, there are no instances where the non-zero trajectory method calls for fewer waves of nonrespondent follow-up than the Wald chi-square method. This is at least partly influenced by the fact that the former requires a minimum of four waves, whereas the latter requires only two. Given the empirical tendency for the components of the index to increase with each new set of responses received, the earlier determination of phase capacity is coupled with a larger absolute magnitude of relative nonresponse error. For instance, we can note from Table 4 that the maximum absolute nonresponse error in the non-zero trajectory method is 0.5, whereas only two indices' nonresponse error measures fall within that bound for the Wald chi-square method.

6 Discussion

The purpose of this paper was to introduce and evaluate two multivariate methods to test for phase capacity in a survey design phase. Their advent was motivated by the objective of simultaneously assessing whether phase capacity has occurred for a battery of D nonresponse-adjusted point estimates. Faced with the unfortunate reality of falling response rates and increased levels of effort required to obtain a survey complete, these methods are designed to help practitioners of responsive and adaptive survey design be more nimble and efficient when allocating data collection resources. This is because the methods help identify the point at which when incoming responses have become "more of the same", thereby implying some form of design phase change is in order. The first method involved formulating a Wald chi-square test statistic in a straightforward multivariate extension of the two-sample t tests proposed in Lewis (2017b), whereas the second method drew upon basic tools for visualizing longitudinal data on change (Singer & Willett, 2003) to assess whether the trajectories of change for the D estimates are jointly indistinguishable from 0. For both, an insignificant test statistic is taken as evidence that all point estimates have stabilized to the point of phase capacity, and that some form of design phase change is warranted.

The two methods were contrasted by way of a simulation study and application using data from the 2011 Federal Employee Viewpoint Survey. Both simulation and application revealed that, all else equal, the non-zero trajectory method tends to dictate more wave of nonrespondent follow-up. The divergence in performance could be a function of two things. For one, the non-zero trajectory method requires a minimum of four waves of data, whereas the Wald chi-square method only requires two. Another reason is that the Wald chi-square method more directly accounts for the implicit covariance in point estimates due to the shared set of respondents through wave $k - 1$, and also the covariance of point estimates themselves (i.e., by way of the off-diagonal entries of $\boldsymbol{S}$). The non-zero trajectory method only indirectly accounts for these covariance sources by virtue of the relative percent changes in the point estimates converging to 0. Naturally, in settings where nonresponse error is reduced in magnitude with each new wave of data collection, even after applying nonresponse adjustments, the non-zero trajectory method will yield point estimates with a smaller relative error. But the

Table 3
Simulation Study Results Comparing the Two Multivariate Phase Capacity Tests

Wald Chi-Square Method				
Condition	Measure	Agency 1	Agency 2	Agency 3
1. Wave not associated with outcome variables	Mean Stop Wave	2.05	2.09	2.06
	Std. Dev. of Stop Wave	0.22	0.33	0.24
	Mean Nonresponse Error of Index	0.00	0.05	0.00
	Mean RMSE of Index	0.60	2.23	0.57
	95% CI Coverage Rate for Index	98.71	98.33	98.93
2. Wave associated with outcome variables	Mean Stop Wave	6.84	2.13	6.12
	Std. Dev. of Stop Wave	2.72	0.48	2.89
	Mean Nonresponse Error of Index	−1.55	−5.76	−2.01
	Mean RMSE of Index	1.71	6.08	2.14
	95% CI Coverage Rate for Index	64.98	8.64	56.13
Non-Zero Trajectory Method				
Condition	Measure	Agency 1	Agency 2	Agency 3
1. Wave not associated with outcome variables	Mean Stop Wave	4.17	4.17	4.16
	Std. Dev. of Stop Wave	0.46	0.45	0.44
	Mean Nonresponse Error of Index	0.00	0.02	−0.01
	Mean RMSE of Index	0.43	1.61	0.41
	95% CI Coverage Rate for Index	99.72	99.47	99.38
2. Wave associated with outcome variables	Mean Stop Wave	6.79	5.16	6.76
	Std. Dev. of Stop Wave	2.95	1.68	2.95
	Mean Nonresponse Error of Index	−1.22	−1.59	−1.15
	Mean RMSE of Index	1.38	2.23	1.31
	95% CI Coverage Rate for Index	47.64	74.54	46.89

Wald chi-square method's proclivity for declaring phase capacity sooner can prove efficient when there is no relationship between response wave and the outcome variables.

The methods introduced in this paper share certain limitations. One is that they are retrospective, meaning the phase capacity determination is made after the most recent wave(s) of data has arrived. Wagner and Raghunathan (2010) propose a prospective "stop-and-impute" univariate phase capacity test, with the goal of assessing whether a pending nonrespondent follow-up attempt is likely to significantly change a sample mean. Further research could look into ways the methods described in this paper could be adapted to forecast results from waves $k + 1$ and beyond, or ways in which Wagner and Raghunathan's approach could be extended to more than one sample mean or, more generally, to other types of point estimates. A second limitation is that both multivariate phase capacity tests introduced in this paper were implemented using a default significance level of $\alpha = 0.05$. This may not be appropriate in all settings; one may instead wish to adjust the significance level as a function of the sample size or the portion of new respondents obtained in the most recent wave(s). For example, it may be undesirable to call for continued follow-up attempts predominantly because a large underlying sample size is detecting statistical significant differences which are practically insignificant. On the other hand, it may be undesirable to base the phase capacity decision too heavily on imprecision in the point estimates attributable to a small sample size.

The research presented in this paper could also be extended in other ways. Future work could investigate point estimates other than ratios (or functions of ratios) as were exclusively considered herein, and for surveys other than the FEVS. Secondly, considering that the Wald chi-square method is a direct extension of the weighting variant of the univariate phase capacity testing method discussed in Lewis (2017b), further research could attempt to develop a more formal multivariate extension of the univariate phase capacity proposed by Rao et al. (2008) based on multiple imputation. Granted, the non-zero trajectory method can be used in combination with any nonresponse adjustment procedure. Another potentially worthwhile extension would be a multivariate variant of the phase capacity determination rule dis-

Table 4
Results from the 2011 FEVS Application Comparing the Two Multivariate Phase Capacity Tests

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

cussed in Vandenplas, Loosveldt, and Beullens (2017) in which changes in a point estimate are referenced against a pre-set standard error. Lastly, a different point of view for assessing multivariate phase capacity is to track overall or partial *R*-indicators (Schouten et al., 2012; Schouten, Cobben, & Bethlehem, 2009) as discussed in Moore, Durrant, and Smith (2016). Rather than focusing on the point estimates themselves, monitoring an *R*-indicator focuses on the other factor that can exacerbate nonresponse error: variability amongst the sample cases' estimated response propensities. It would be enlightening to learn, perhaps with theoretical derivations or with the help of a simulation study, conditions under which decisions based on these alternative criteria converge or diverge from the two tests introduced in this paper.

References

Atrostic, B., Bates, N., & Silberstein, A. (2001). Nonresponse in US government household surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics*, *17*(2), 209–226.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. Hoboken, NJ: Wiley.

Beullens, K., Loosveldt, G., Vandenplas, C., & Stoop, I. (2018). Response rates in the European Social Survey: Increasing, decreasing, or a matter of fieldwork efforts? *Survey Methods: Insights from the Field*. Retrieved from https://surveyinsights.org/?p=9673

Brick, J. M. & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, *5*(3), 215–238.

Brick, J. M. & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 36–59.

Couper, M. (1998). *Measuring survey quality in a casic environment*. Paper presented at the Joint Statistical Meetings of the American Statistical Association.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, *69*(1), 87–98.

de Leeuw, E. & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse*. New York: Wiley.

Groves, R. & 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*(3), 439–457.

Groves, R., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *The Public Opinion Quarterly*, *64*(30), 299–308.

Heeringa, S., West, B., & Berglund, P. (2017). *Applied survey data analysis* (2nd ed.). Boca Raton, FL: Chapman Hall/CRC Press.

Izrael, D., Hoaglin, D., & Battaglia, M. (2000). *A SAS macro for balancing a weighted sample*. Proceedings of the SAS Users Group International (SUGI) Conference. Retrieved from http://www2.sas.com/proceedings/sugi25/25/st/25p258.pdf

Kalton, G. & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2), 81–97.

Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information*. Hoboken, NJ: Wiley.

Lewis, T. (2012). *Incorporating the sampling variability from an employee perception survey into the ranking process of U.S. government agencies*. Proceedings of the Fourth International Conference on Establishment Surveys (ICES-IV).

Lewis, T. (2014). *Testing for phase capacity in surveys with multiple waves of nonrespondent follow-up*. Ph.D. thesis, University of Maryland.

Lewis, T. (2017a). Temporal perspectives of nonresponse during a survey design phase. *Methods, Data, Analyses*, *11*(2), 189–206.

Lewis, T. (2017b). Univariate tests for phase capacity: Tools for identifying when to modify a survey's data collection protocol. *Journal of Official Statistics*, *33*(3), 601–624.

Moore, J., Durrant, G., & Smith, P. (2016). Data set representativeness during data collection in three UK social surveys: Generalizability and the effects of auxiliary covariate choice. *Journal of the Royal Statistical Society: Series A*. online first edition.

Potthoff, R., Manton, K., & Woodbury, M. (1993). Correcting for nonavailability bias in surveys by weighting based on number of callbacks. *Journal of the American Statistical Association*, *88*(424), 1197–1207.

Rao, R., Glickman, M., & Glynn, R. (2008). Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statistics in Medicine*, *27*(12), 2196–2213.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., ... Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, *80*(3), 382–399.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*(1), 101–113.

Schouten, B., Peytchev, A., & Wagner, J. (2017). *Adaptive survey design*. Boca Raton, FL: Chapman Hall/CRC Press.

Sigman, R., Lewis, T., Yount, N., & Lee, K. (2014). Does the length of fielding period matter? Examining response scores of early versus late responders. *Journal of Official Statistics*, *30*(4), 651–674.

Silver, N. (2014). *Is the polling industry in stasis or in crisis?* Retrieved from https://fivethirtyeight.com/features/is-the-polling-industry-in-stasis-or-in-crisis/

Singer, J. & Willett, J. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford.

Tourangeau, R. & Plewes, T. (Eds.). (2013). *Nonresponse in social science surveys: A research agenda*. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/read/18293/chapter/1

U.S. Office of Personnel Management. (2016). *2016 Federa employee viewpoint survey: Technical report*. Retrieved from https://www.opm.gov/fevs/archive/2016FILES/2016_FEVS_Technical_Report.pdf

U.S. Office of Personnel Management. (2017). *2017 Federal employee viewpoint survey: Governmentwide management report*. Retrieved from https://www.opm.gov/fevs/archive/2017FILES/2017_FEVS_Gwide_Final_Report.PDF

Vandenplas, C., Loosveldt, G., & Beullens, K. (2017). Fieldwork monitoring for the European Social Survey: An illustration with Belgium and the Czech Republic in Round 7. *Journal of Official Statistics*, *33*(3), 659–686.

Wagner, J. (2008). *Adaptive survey design to reduce nonresponse bias*. Ph.D. thesis, University of Michigan.

Wagner, J. & Raghunathan, T. (2010). A new stopping rule for surveys. *Statistics in Medicine*, *29*(9), 1014–1024.

Williams, D. & Brick, J. M. (2017). Trends in U.S. face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, *62*(2), 186–211.

Wolter, K. (2007). *Introduction to variance estimation* (2nd ed.). 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*(334), 411–414.

Appendix

(Appendix tables on next pages)

Table A1

Items Comprising the U.S. Office of Personnel Management's Four Human Capital Assessment and Accountability Framework (HCAAF) Indices Derived from the 2011 Federal Employee Viewpoint Survey

Job Satisfaction Index (JS)		
Item	Wording	Response Scale[1]
4	My work gives me a feeling of personal accomplishment.	Strongly Agree - Strongly Disagree
5	I like the kind of work I do.	Strongly Agree - Strongly Disagree
13	The work I do is important.	Strongly Agree - Strongly Disagree, with DNK
63	How satisfied are you with your involvement in decisions that affect your work?	Very Satisfied - Very Dissatisfied
67	How satisfied are you with your opportunity to get a better job in your organization?	Very Satisfied - Very Dissatisfied
69	Considering everything, how satisfied are you with your job?	Very Satisfied - Very Dissatisfied
70	Considering everything, how satisfied are you with your pay?	Very Satisfied - Very Dissatisfied
Leadership and Knowledge Management Index (LKM)		
Item	Wording	Response Scale[1]
10	My workload is reasonable.	Strongly Agree - Strongly Disagree, with DNK
35	Employees are protected from health and safety hazards on the job.	Strongly Agree - Strongly Disagree, with DNK
36	My organization has prepared employees for potential security threats.	Strongly Agree - Strongly Disagree, with DNK
51	I have trust and confidence in my supervisor.	Strongly Agree - Strongly Disagree
52	Overall, how good a job do you feel is being done by your immediate supervisor/team leader?	Very Good - Very Poor
53	In my organization, leaders generate high levels of motivation and commitment in the workforce.	Strongly Agree - Strongly Disagree, with DNK
55	Managers/supervisors/team leaders work well with employees of different backgrounds.	Strongly Agree - Strongly Disagree, with DNK
56	Managers communicate the goals and priorities of the organization.	Strongly Agree - Strongly Disagree, with DNK
57	Managers review and evaluate the organization's progress toward meeting its goals and objectives.	Strongly Agree - Strongly Disagree, with DNK
61	I have a high level of respect for my organization's senior leaders.	Strongly Agree - Strongly Disagree, with DNK
64	How satisfied are you with the information you receive from management on what's going on in your organization?	Very Satisfied - Very Dissatisfied
66	How satisfied are you with the policies and practices of your senior leaders?	Very Satisfied - Very Dissatisfied

Continues on next page

Continued from last page

Item	Wording	Response Scale[1]
	Results-Oriented Performance Culture Index (ROPC)	
12	I know how my work relates to the agency's goals and priorities.	Strongly Agree - Strongly Disagree, with DNK
14	Physical conditions (for example, noise level, temperature, lighting, cleanliness in the workplace) allow employees to perform their jobs well.	Strongly Agree - Strongly Disagree, with DNK
15	My performance appraisal is a fair reflection of my performance.	Strongly Agree - Strongly Disagree, with DNK
20	The people I work with cooperate to get the job done.	Strongly Agree - Strongly Disagree
22	Promotions in my work unit are based on merit.	Strongly Agree - Strongly Disagree, with DNK
23	In my work unit, steps are taken to deal with a poor performer who cannot or will not improve.	Strongly Agree - Strongly Disagree, with DNK
24	In my work unit, differences in performance are recognized in a meaningful way.	Strongly Agree - Strongly Disagree, with DNK
30	Employees have a feeling of personal empowerment with respect to work processes.	Strongly Agree - Strongly Disagree, with DNK
32	Creativity and innovation are rewarded.	Strongly Agree - Strongly Disagree, with DNK
33	Pay raises depend on how well employees perform their jobs.	Strongly Agree - Strongly Disagree, with DNK
42	My supervisor supports my need to balance work and other life issues.	Strongly Agree - Strongly Disagree, with DNK
44	Discussions with my supervisor/team leader about my performance are worthwhile.	Strongly Agree - Strongly Disagree, with DNK
65	How satisfied are you with the recognition you receive for doing a good job?	Very Satisfied - Very Dissatisfied
	Talent Management Index (TM)	
Item	Wording	Response Scale[1]
1	I am given a real opportunity to improve my skills in my organization.	Strongly Agree - Strongly Disagree
11	My talents are used well in the workplace.	Strongly Agree - Strongly Disagree, with DNK
18	My training needs are assessed.	Strongly Agree - Strongly Disagree, with DNK
21	My work unit is able to recruit people with the right skills.	Strongly Agree - Strongly Disagree, with DNK
29	The workforce has the job-relevant knowledge and skills necessary to accomplish organizational goals.	Strongly Agree - Strongly Disagree, with DNK
47	Supervisors/team leaders in my work unit support employee development.	Strongly Agree - Strongly Disagree, with DNK
68	How satisfied are you with the training you receive for your present job?	Very Satisfied - Very Dissatisfied

[1] All items were posed on a five-point Likert-type response scale, with a "Do Not Know" (DNK) option provided where noted