a New Day for Federal Service

# Multivariate Tests for Phase Capacity

5th Workshop on Adaptive and Responsive Design

University of Michigan

Ann Arbor, MI

November 7, 2017

Taylor Lewis[1]

Senior Data Scientist

U.S. Office of Personnel Management

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

# Outline

I. Background

II. Brief Summary of Prior Research – Univariate Phase Capacity Tests

III. Multivariate Extensions of Phase Capacity Tests:
   1. Wald Chi-Square Method
   2. Non-Zero Trajectory Method

IV. Retrospective Application using the 2011 Federal Employee Viewpoint Survey

V. Limitations and Further Research

# I. Background

# Nonresponse and Nonrespondent Follow-Up

- Invariably, not all sampled units respond to the initial survey solicitation
- Most surveys repeatedly follow-up with nonrespondents making additional mailings, phone calls, household visits, etc., sometimes with a preset response rate target in mind
- Each subsequent reminder brings in a new “wave” of data, which tends to be progressively smaller in size, thereby impacting estimates less and less
- Other temporal delineations of waves possible

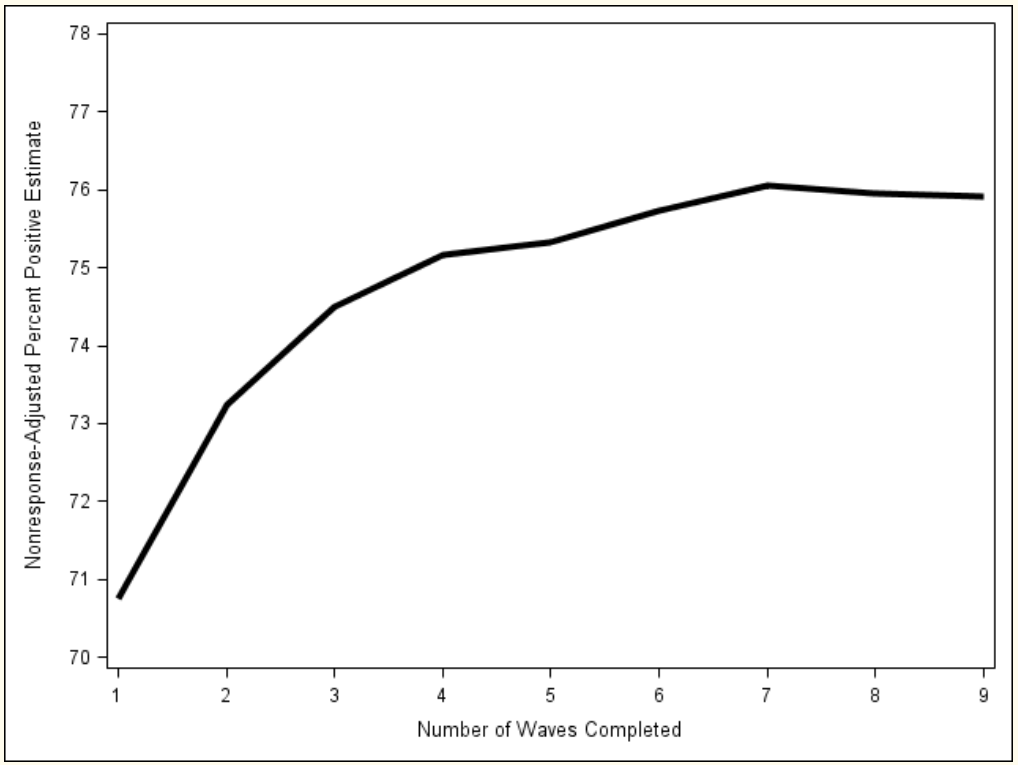
# The Notion of Phase Capacity

- In their discussion of responsive survey design, Groves and Heeringa (2006) define the following key terms:
  - *design phase* – spell of data collection period with stable frame, sample, and recruitment protocol
  - *phase capacity* – point during a design phase at which additional responses cease influencing key statistics
- Rather than fixating on a target response rate, they argue one should change design phases (e.g., switch mode, increase incentive) or discontinue nonrespondent follow-up altogether once phase capacity has been reached
- Problem for practitioners: no calculable rule given

# Illustration of Phase Capacity in the Federal Employee Viewpoint Survey (FEVS)

- The FEVS is an annual organizational climate survey administered by the U.S. Office of Personnel Management (OPM) to a sample of 800,000+ federal employees from 80+ agencies
- Web-based instrument comprised mainly of attitudinal items posed on a five-point Likert scale
- Key statistics are “percent positive” estimates based on the dichotomization of, for example, “Completely Agree” or “Agree” elections versus all other possible response choices
- Nonrespondents are sent weekly reminder emails

# Example of a Nonresponse-Adjusted Percent Positive Trend Using Cumulative Responses




← Goal is to identify point estimate stability at earliest possible wave

Note: estimate stability does not necessarily imply that the value converged upon is free of nonresponse error; it implies that additional follow-ups under the same protocol will continue to be inefficacious

# II. Brief Summary of Prior Research – Univariate Phase Capacity Tests

# Previously Proposed Univariate Tests

- Rao, Glickman, and Glynn (RGG) (2008) (termed "stopping rules") – best-performing method used multiple imputation (MI)

- Idea is to multiply impute (Rubin, 1987) the missing data $M$ ($M \geq 2$) times for nonrespondents as of wave $k$, then delete responses obtained during wave $k$, specifically, and repeat for nonrespondents as wave $k - 1$ → result is $2M$ completed data sets and two nonresponse-adjusted, MI point estimates

- A $t$-test is carried out by dividing the two point estimates' difference by an estimate of the MI variance of the difference

- Phase capacity declared once the test statistic is insignificant

# Previously Proposed Univariate Tests (2)

- RGG approach is limited in that it is only designed to track a sample mean and inapplicable to surveys that conduct weighting adjustments for nonresponse

- Lewis (2017) describes a new method circumventing these limitations: same premise, except nonresponse-adjusted point estimates are formulated based on two sets of weights, one for respondents through wave $k$ and another for respondents through wave $k - 1$

- As with the RGG approach, tricky part is deriving a variance factoring in the covariance attributable to shared respondent set through wave $k - 1$

- Two viable methods to do so: (1) Taylor series linearization; (2) replication

# III. Multivariate Extensions of Phase Capacity Tests

# Background

- A practical limitation of both the RGG approach and Lewis' variant is that they are univariate in nature → how would one proceed if independently conducted on two or more point estimates with conflicting results?

- Conference paper discusses to proposals to provide a single yes/no answer for a battery of *D* point estimates:
  1. Wald Chi-Square Method – direct multivariate extension of two-sample *t*-test using matrix algebra
  2. Non-Zero Trajectory Method – based on ideas of longitudinal data analysis (Singer and Willett, 2003), jointly fit *D* simple linear regression models of point estimates' relative percent change

- Both methods default to treating each point estimate difference equivalently, but differential importance can be assigned to each via a contrast vector

# Wald Chi-Square Method

- Let **D** denote a *D* x 1 matrix of nonresponse-adjusted point estimate differences, and let **S** denote the corresponding *D* x *D* variance-covariance matrix

- Entries of **S** can be obtained via Taylor series linearization or replication (see Section 3.2 of Lewis, 2017)

- Supposing the goal is to test for no significant differences, the test statistic is

$$\chi^2_W = \mathbf{D}^{\mathbf{T}}\mathbf{S}^{-\mathbf{1}}\mathbf{D}$$

  which is referenced against a chi-square distribution with *D* – 1 degrees of freedom

- Phase capacity declared whenever test statistic is not significant

# Non-Zero Trajectory Method

- Find the *D* differences' 3 most recent relative percent changes (to harmonize potential scale incongruities):

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

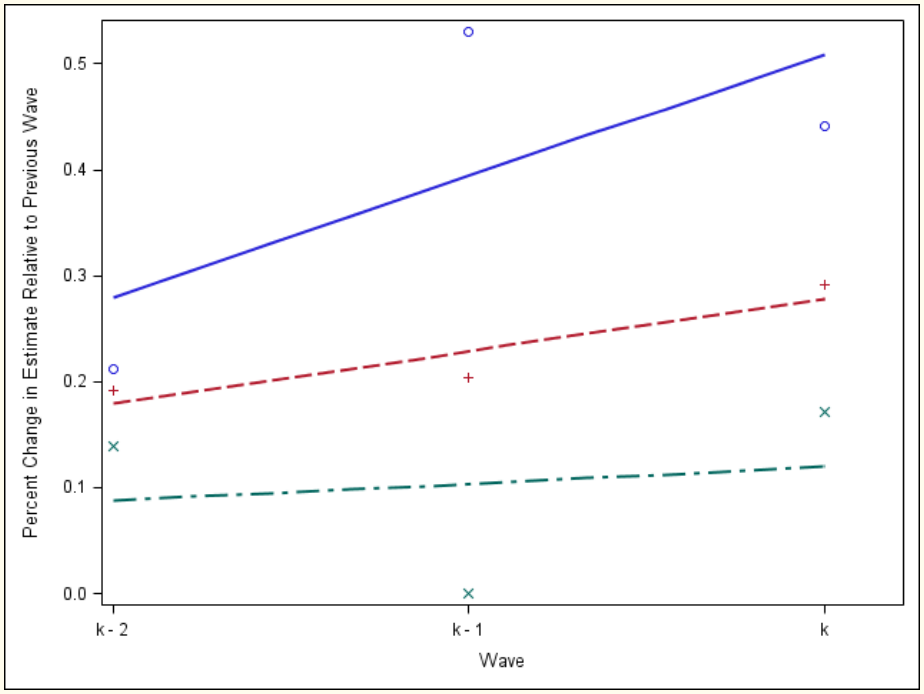
- Treating *w* as a wave indicator one unit apart (e.g., 1, 2, 3), one then estimates the following model:

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

where the first set of *D* terms represent estimate-specific intercepts, and the second set represents estimate-specific slopes

- Disadvantage: at least 4 waves needed (Wald needs 2)

# Visualization of Non-Zero Trajectory Method




- When point estimates have stabilized, all intercept/slope terms should be insignificantly different from zero; we can test for this using the following *F* test:

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

which can be referenced against an *F* distribution with *D* numerator and and 2*D* denominator degrees of freedom

# IV. Retrospective Application using the 2011 Federal Employee Viewpoint Survey

# FEVS 2011 Application Details

- Batteries of point estimates investigated were the four Human Capital Assessment and Accountability Framework (HCAAF) indices, which are averages of the percent positive estimates of thematically-linked items (e.g., Job Satisfaction, Talent Management)

- Using timestamp information for three agencies, respondents were partitioned into waves, and each successive (cumulative) set of respondents was assigned a set of weights raked to known marginal distributions from sample frame (e.g., agency component, minority status, gender, and supervisory status)

- Retroactively implemented the two methods for each agency x index combination to compare and contrast performance

# FEVS 2011 Application Results

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

- Wald method concludes phase capacity earlier, in part because it requires fewer waves (2 vs. 4 for NZT); this results in larger residual differences relative to the final wave estimate (see NR Error column) – recall there is an upward trend in the point estimates underlying indices

# V. Limitations and Further Research

# Practical Limitations

- Actual adoption of these approaches in FEVS would face resistance because:
  - Desirable to treat each agency equitably; beginning in FEVS 2012, field period was preset to 6 weeks for all agencies
  - Higher scores are better, and so there may be opposition to any change, shortened field period included, believed to reduce point estimates

- Data must be collected/processed real-time, and it was tacitly assumed that the full sample is “active” – may be impractical for in-person surveys covering a vast geographical expanse taking weeks or months for interviewers to exhaust sample cases, although tests could be applied to subsamples

# Practical Limitations (2)

- Even when entire sample is “active,” may not be feasible to send reminders simultaneously as in the FEVS Web mode – alternative data collection wave definition may be a plausible work-around
- Despite aversion to phrase *stopping rule*, stopping was the only design phase change investigated in this research
- Would be interesting to investigate in a mixed-mode survey setting or in surveys with two stages of data collection, such as the National Immunization Survey (NIS) or the Residential Energy Consumption Survey (RECS)
- In those settings, differential sensitivities may be desired

# Further Research Ideas

- All phase capacity testing methods discussed today are retrospective in nature; future research could develop prospective variants in the spirit of the one proposed by Wagner and Raghunathan (2010)

- Compare performance with another recently proposed phase capacity testing method by Moore et al. (2016) that considers CV thresholds in an overall and partial R-indicator (Schouten et al., 2009; Schouten et al., 2012)

- Given the survey is annual with substantial overlap in the sample composition, carry forward prior year(s) information to facilitate the phase capacity determination

- Time series/forecasting methods and/or Bayesian approaches

# Thanks!

Questions/Comments?
Taylor.Lewis@opm.gov

# References


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

Lewis, T. (2017). "Univariate Tests for Phase Capacity: Tools for Identifying When to Modify a Survey's Data Collection Protocol," *Journal of Official Statistics*, **33**, pp. 601 – 624.

Moore, J., Durrant, G., and 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 Statistics Society: Series A*, online first edition.

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

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

Schouten, B., Cobben, F. and Bethlehem, J. (2009). "Indicators for the Representativeness of Survey Response," *Survey Methodology*, **35**, pp. 101 – 113.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. and Skinner, C. (2012). "Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-indicators and Partial R-indicators," *International Statistical Review*, **80**, pp. 382-399.

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

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