Westat

Washington Statistical Society • September 9, 2015

# Non-Probability Sampling Assumptions and Methods

J. Michael Brick
Westat

The views presented in this paper are those of the author(s) and do not represent the official views of any Federal Government agency/department or Westat.

# The Non-Probability Sampling Explosion

- Global $$$ for online research 19% to 35% from 2006-12
- 43% of all surveys conducted online in 2012
- Online surveys used by all types of organizations
  - Commercial
  - Academic
  - Government

# Non-Probability Sampling (NPS) Literature

- Two AAPOR panels
- Monograms
- Ever increasing number of journal articles from many disciplines
- International scope

# What Is THE Issue

- Representation
- Probability sampling is strong on representation
  - Fixed sampling frame and probabilities of selection basis for inference that is relatively robust despite problems
- Non-probability sampling weaker on representation
  - Models and assumptions that are hard to justify or test

# NPS Online Design Approaches

- Matching
  - Identify units from a probability sample or census that have characteristics highly related to the key survey outcome variables and locate NPS respondents matching those characteristics
- Quotas
  - Essentially the same as matching but typically based on demographic variables
- Blending
  - Combining samples; sometimes NPS with probability sample and sometimes multiple NPS

# Typical NPS Weighting Approaches

- Weight observed sample with initial weights of unity
  - Unweighted
  - Poststratification or raking
  - Inverse Probability Weighting (IPW)

# Poststratification or Raking

- **Consider Outcome model**
  $E_O y_k = \mu + \alpha_g = \mu_g$ for all $k \in s_g$ , $g$=1,…,$G$

- **Poststratification (unweighted poststratification cell mean adjusted to population total for the cell) is unbiased under this model**
- **Poststratification is criticized as not accounting for selection bias**

# Inverse Probability Weighting

- Consider Missingness Model

$$E_M(R_k \mid \mathbf{Z}) = \pi_{\mathbf{Z}_k}$$

where $\pi_{\mathbf{Z}_k}$ is propensity of unit *k*

- Inverse of propensity score adjustment (observation weighted using reference sample, see Lee (2006)) is unbiased under this model

- IPW criticized as being unstable when propensities are extreme

Westat

# A Compositional Model

- First IPW then poststratification to give $\{w_k\}$
- Lee and Valliant (2009) describe this weighting method
- Related to calibration and doubly robust augmented IPW (AIPW), but called compositional because only counts of population controls allowed (GREG not in this class)

# Properties

1) $w_k > 0 \ \forall \ k \in s$
2) $\sum_{k \in s} w_k \boldsymbol{\delta}_k = \mathbf{N}$ where ***N*** is a vector of pop totals
3) Estimates of totals are linear or smooth function of estimated totals
4) Unbiased and consistent if either outcome or missingness model holds

# Marginal Structural Model

- Structural model specified by mean and variance models.
- Assume a population structure with clustering generates the data and observations within cluster may be correlated (for variance computation).
- Resample clusters to estimate variance of estimates

- Under the models $\hat{y}_{com}$ is unbiased and consistent and, with large samples, 95% CI is

$$\hat{y}_{com} \pm 2\sqrt{v(\hat{y}_{com})}$$

# Case Study

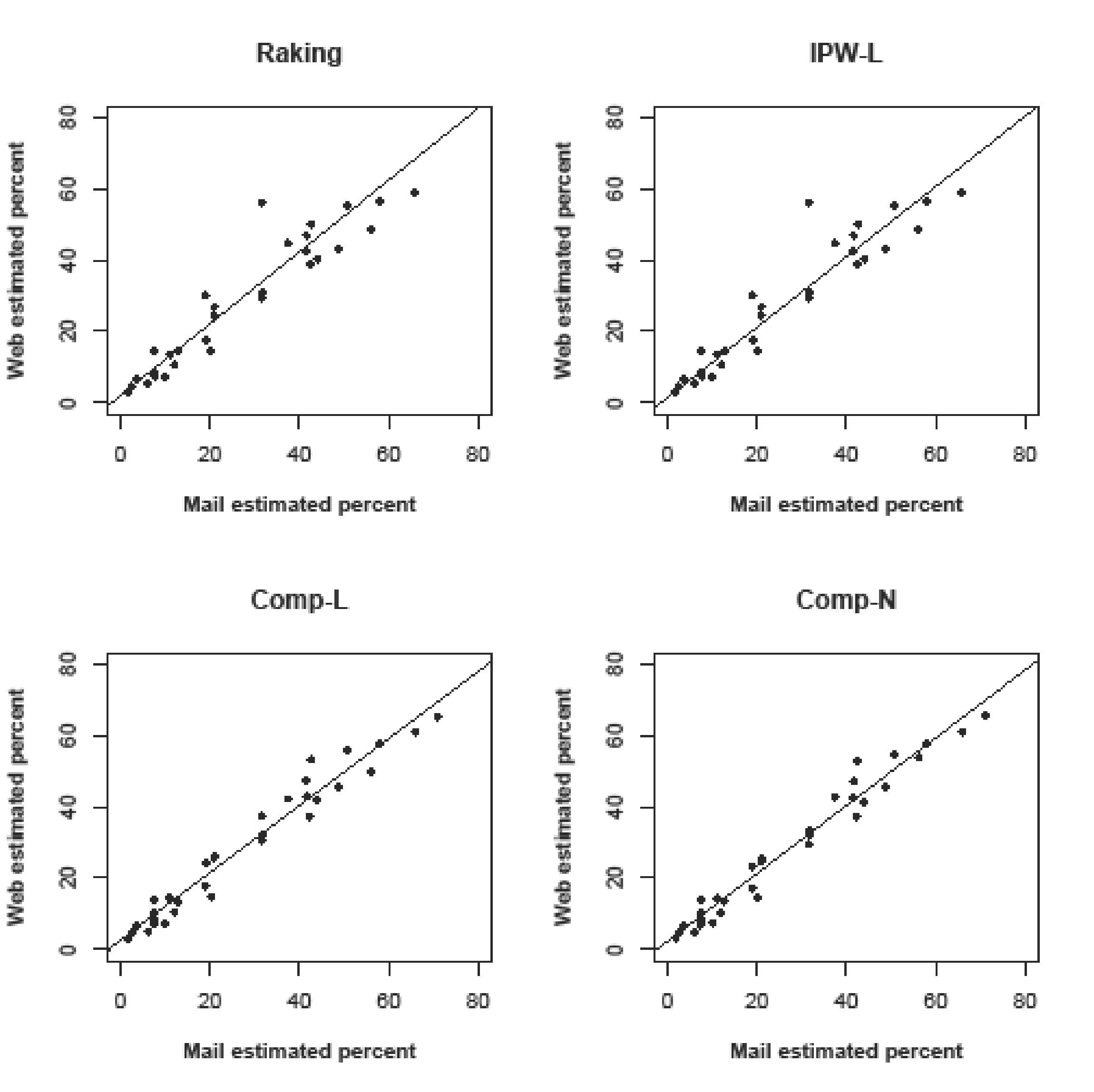
- Collaboration between Pew Research Center, SurveyMonkey, and Westat.
- SurveyMonkey Audience Panel (9/14)
  - 5,301 adult respondents
- ABS (mail) survey (9-10/14) RR=29%
  - 2,668 respondents
  - Serves as reference sample

# Weighting Methods

| | IPW | Raking |
|---|---|---|
| Raking | None | 7 dimensions |
| IPW-L | Logistic - 4 groups | None |
| Comp-L | Logistic - 4 groups | 7 dimensions |
| Comp-N | Exact - 16 groups | 7 dimensions |

- Variance computed using jackknife based on MSA of respondent

# Comparing Web and Mail Substantive Estimates



# Diagnostics

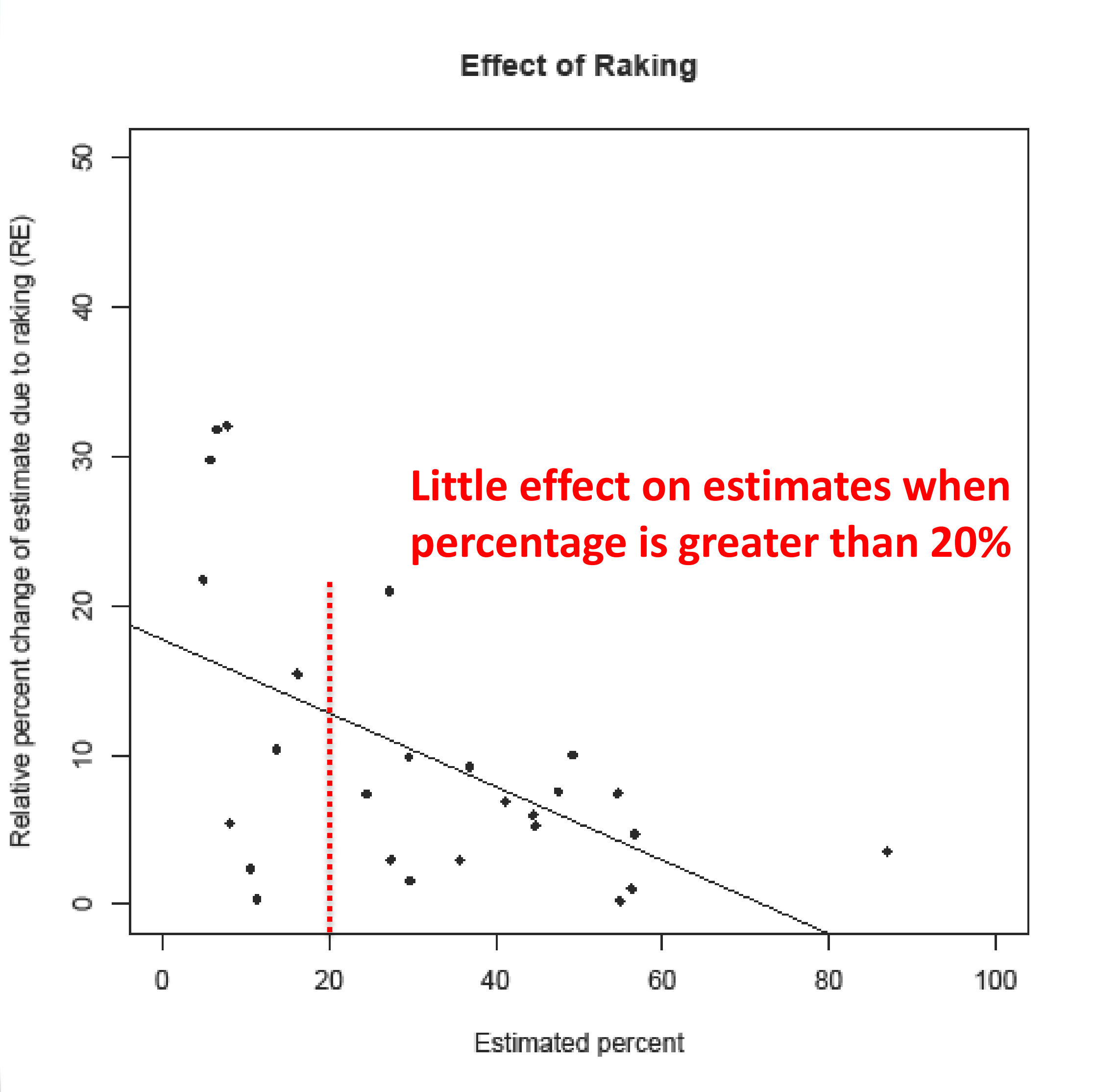
- Examine effects and assumptions
  - Begin with bias reduction due to raking
  - Assess propensity model fit and IPW adjustments
  - Assess outcome model for a particular estimate

# Effect of Poststratification or Raking on Bias

- The Relative Raking Effect (RE) is a measure of how much an estimate changes (relative to the IPW estimate) due to raking.
- Computed for substantive items in Web survey is a modification of the poststratified measure

$$RE(y) = 100\left( \frac{\sum_g N_g \hat{N}_{ipw,g}^{-1} \tilde{y}_g - \sum_g \tilde{y}_g}{\sum_g \tilde{y}_g} \right)$$

# Relative Raking Effect for Substantive Items

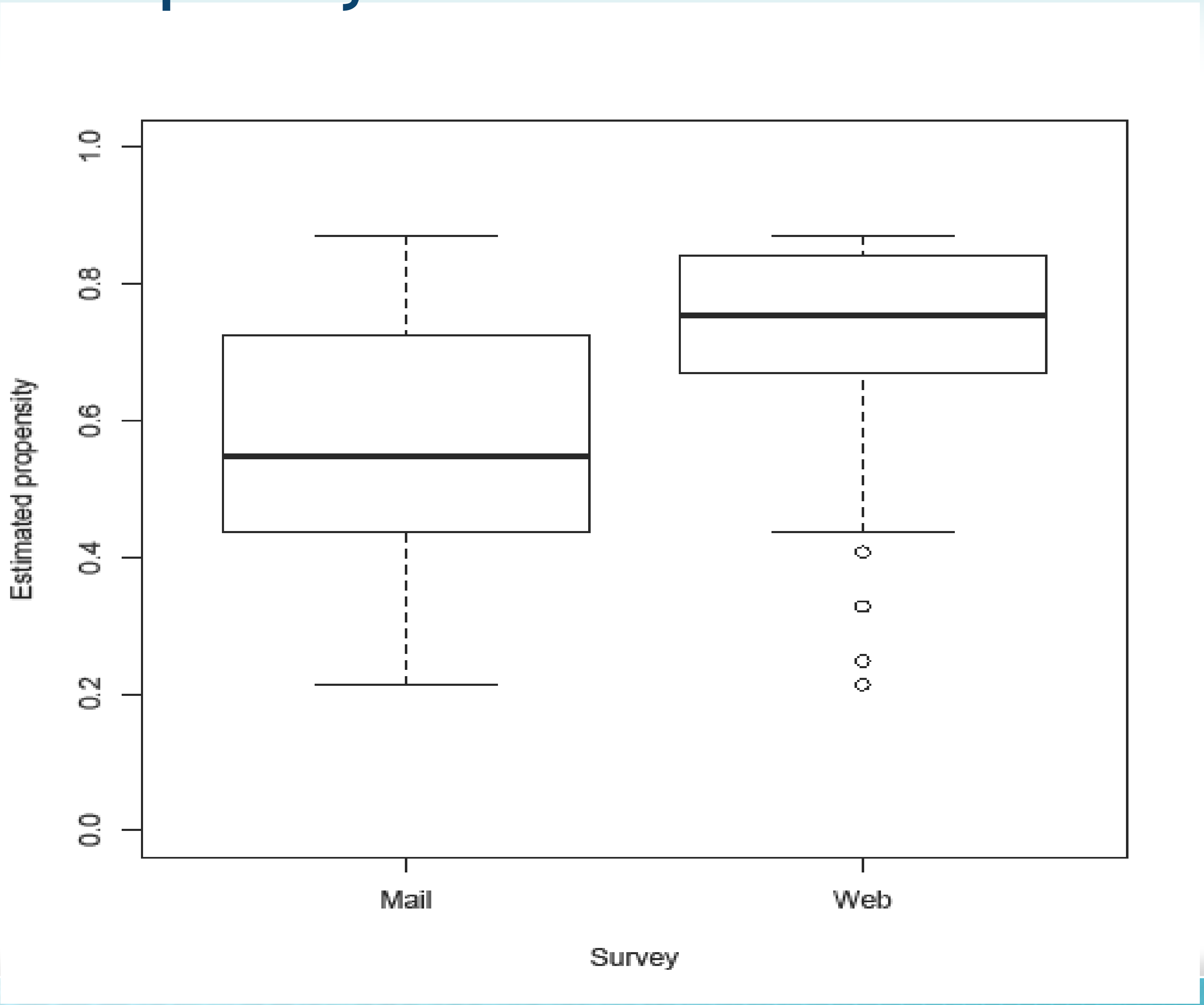


Westat

# Common Support Analysis

- IPW is intended to reduce selection bias
- Commonly used tool of causal analysis is examination of the propensity distributions of the control (in our case Mail PS survey) and treated sample (Web NPS survey)
  - Shown for the IPW-L propensities

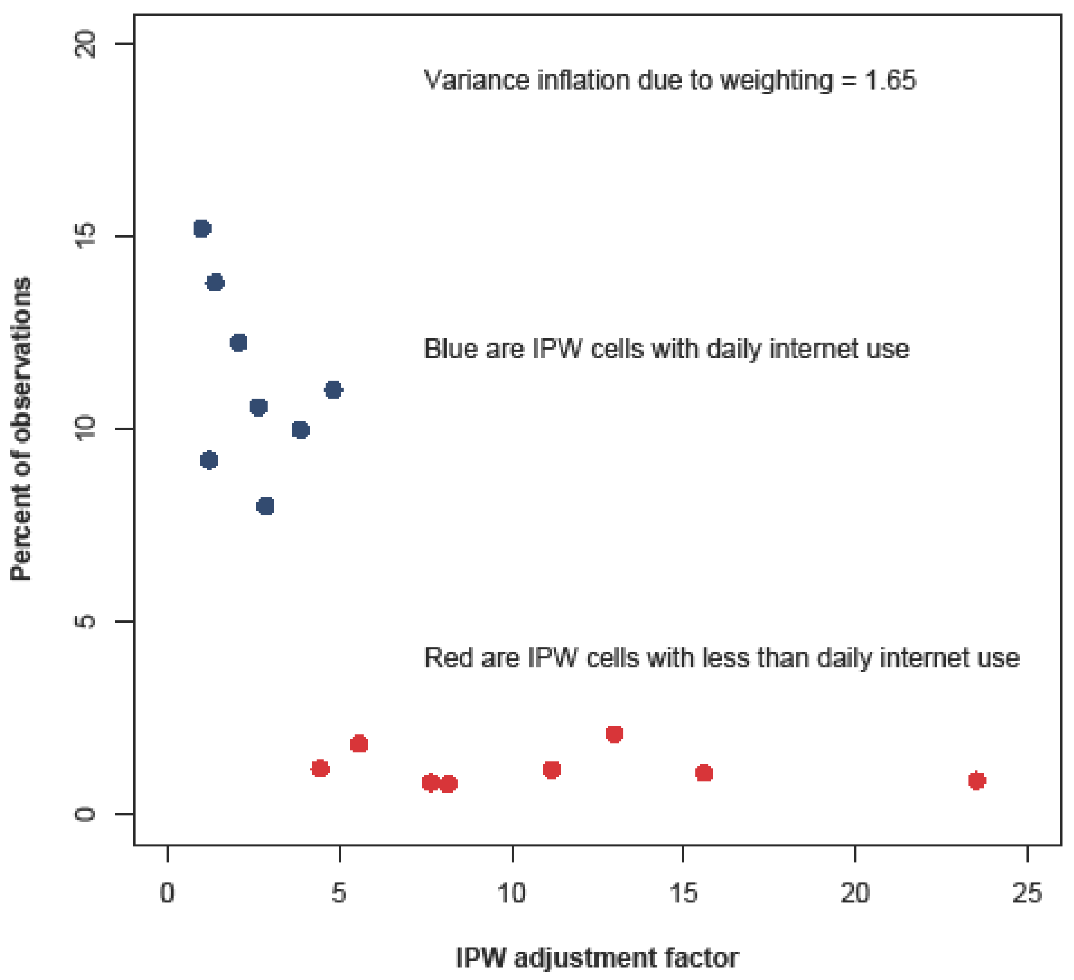
# IPW-L Propensity Distribution



# IPW Adjustment Factors

- The graph shows weak evidence for the common support assumption and raises concerns about the effectiveness and stability of the IPW adjustments
- Considerable range of weights and instability when using the logistic regression approach (IPW-L)

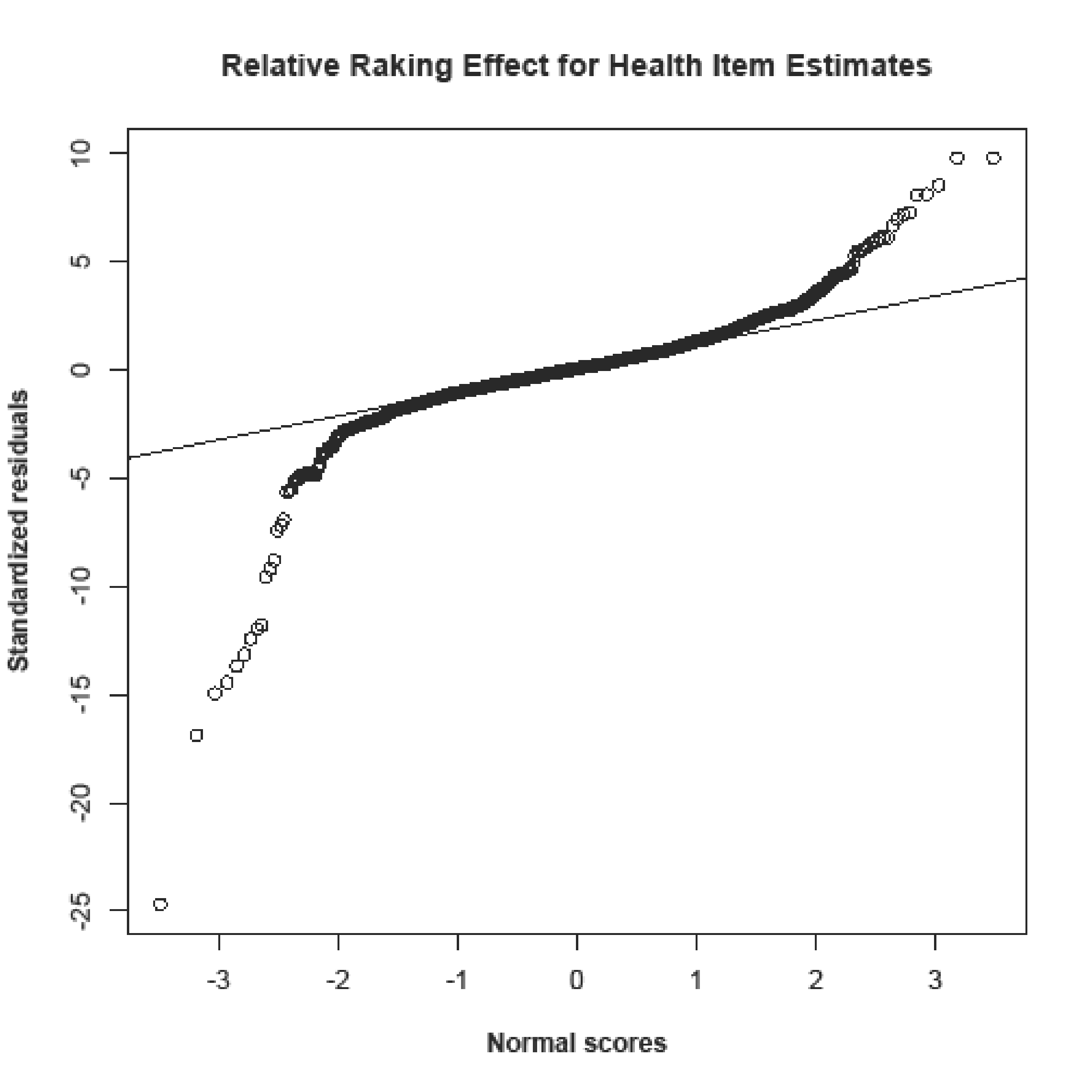
# IPW-N Relative Adjustment Factors



# Closer Look at Outcome Model

- Under model we would assume standardized differences from the "predicted" mean would be approximately N(0,1)
- Examine this for "how you rate your health" by computing residuals from raking dimension means across other raking dimensions

# QQ Plot of Residuals for Comp-N estimates



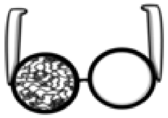
# Variance Estimation

- Estimated design effect (deff) is not simply the clustering and weight adjustment effect
- Median deff for Comp-L is 14.9 (mean 48.3)
  - Without replicating, median is 5.8/mean 6.2
  - Hugely unstable logistic model of propensity
- Median deff for Comp-N is 5.5 (mean 6.5)
  - No difference with replicating IPW-N
  - This means the effective sample size is closer to 1,000 than 5,000

# Discussion

- The formal structure helps in evaluating NPS
- Assumptions for unbiased estimation not well supported
- We need more evaluation tools
  - Especially tools for understanding when estimates from NPS may be more reliable are needed

# What About PS?



- Tools and more theory needed for PS since 10% response rates and low coverage rates are too far from assumptions of design-based theory

- Compositional model may be applicable
  – Current set of tools for evaluating effectiveness of weighting are very limited

# Thanks !!!

mikebrick@westat.com