Evaluating Methods for Analyzing Subpopulation Data with Single-Level and Multilevel Pseudo Maximum Likelihood Estimation

Natalie A. Koziol, Ph.D.
Houston F. Lester, M.A.
Jayden Nord, B.A.

Nebraska Center for Research on Children, Youth, Families & Schools
Nebraska Academy for Methodology, Analytics & Psychometrics

This work was completed utilizing the Holland Computing Center at the University of Nebraska, which receives support from the Nebraska Research Initiative.

Background

Subpopulation analysis

- Research, policies, and practices often target specific groups
- Complex probability sampling complicates subpopulation analyses
  - Design-based variance estimators define variation across all possible samples under the original sampling design
  - Subsetting the data ignores the randomness of the subpopulation sample size
    - Problematic when using linearization methods *and* number of first stage sampling units is altered
  - Multiple-group and zero-weight approaches are preferable

Background

Clustering

- Multilevel modeling
  - Incorporate random effects into the linear predictor (variation in G matrix)
  - Fit the conditional mean
  - Estimators target cluster-specific effects
  - Weighted modeling (e.g., MPML) requires multiple sets of weights and scaling corrections
- Single-level modeling
  - Specify a more complex R matrix / use empirical variance estimators
  - Fit the marginal mean
  - Estimators target population-averaged effects
  - Weighted modeling (e.g., PML) requires one set of weights and no scaling

Background

Combining Subpopulation and Clustering Considerations

- Subpopulation analysis literature limited to single-level modeling
  - Multiple-group and zero-weight approaches provide equivalent results
  - Subsetting the data only negatively impacts variance estimation
- Subpopulation analysis is more nuanced with multilevel modeling
  - Scaling corrections may additionally lead to differences in point estimation
  - Level 1 grouping variables may present complications
    - Only the multiple-group approach can account for correlated group-specific cluster effects
    - Subpopulation cluster sizes may be small (problematic for MPML)
  - No simulation studies have compared subpopulation methods with MPML

Present Study

Purpose

To investigate the interactive effect of subpopulation method and estimation method on the performance of fixed effect parameter and standard error estimators in the context of performing a subpopulation analysis.

Method

Study Conditions

Factor	Level
Subpopulation Method	Multiple-group Zero-weight Subset
Estimation Method	MPML PML
Design Informativeness	Informative Non-informative
Level of group assignment	Level 1 Level 2
Proportion of cases in target group	$\pi_1 = .10$ $\pi_1 = .15$ … $\pi_1 = .90$

Method

Data Generation

1) Generate finite population data

$$Y_{ij,g} = \gamma_{00,g} + e_{ij,g} + u_{0j,g}$$

$\gamma_{00,g} = -.4 + g_{ij} \times .8$ where $g_{ij} \sim Bernoulli(\pi_1)$

$e_{ij,g} \sim N(0, \sigma_g^2)$; $\sigma_0^2 = \sigma_1^2 = .7$

$u_{0j,g} \sim N(0, \tau_{00,g})$; $\tau_{00,0} = \tau_{00,1} = .3$; $\mathrm{Cor}(u_{0j,0}, u_{0j,1}) = .75$ (L1 grouping) or 0 (L2 grouping)

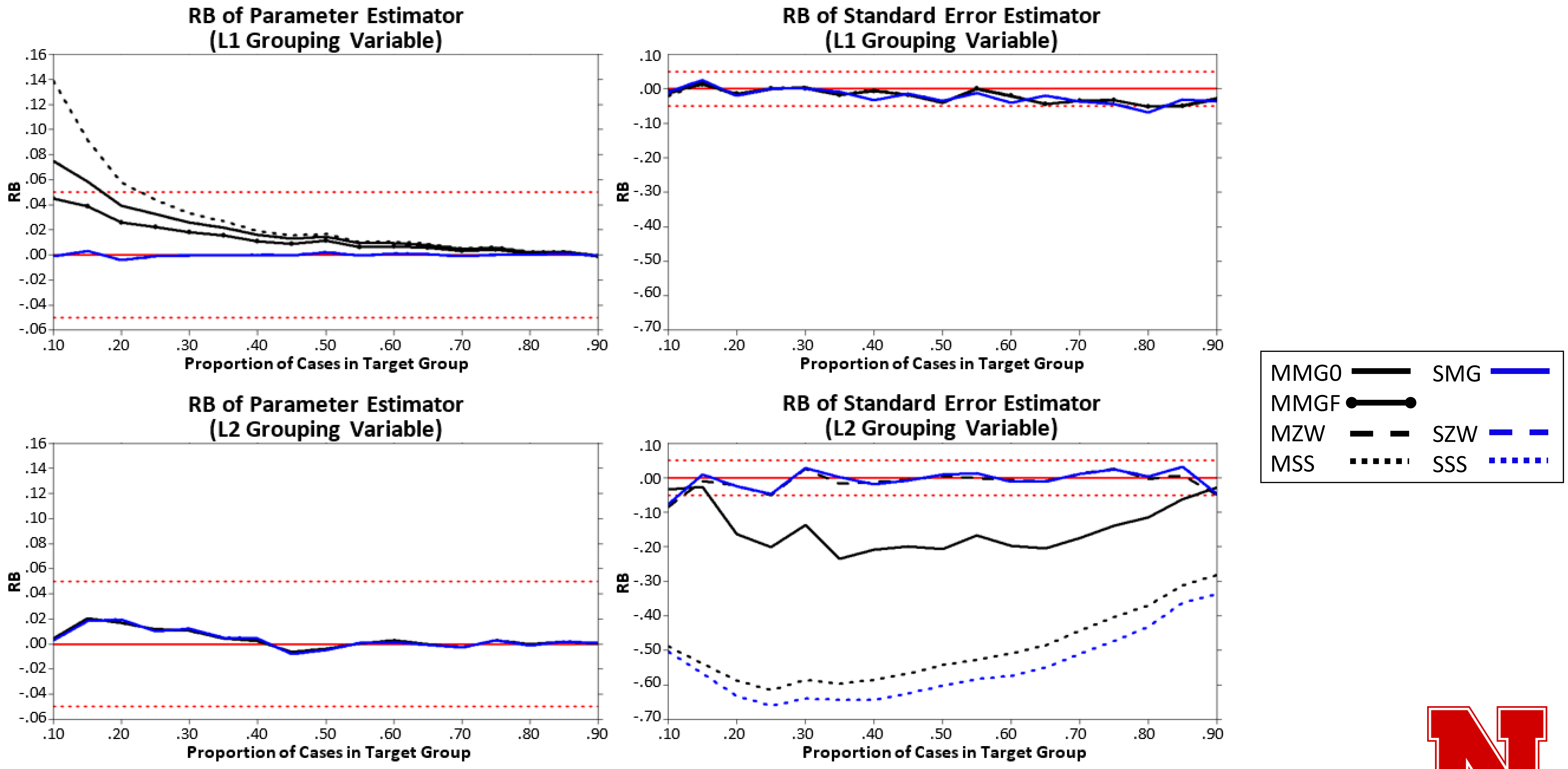
- Generate 20,000 clusters across ten L1 strata
- Generate ≈1,300,000 individual units across two L2 strata

2) Generate sample data

- Select 200 PSUs using stratified systematic PPS sampling
- Select ≈7,000 SSUs using stratified SRS

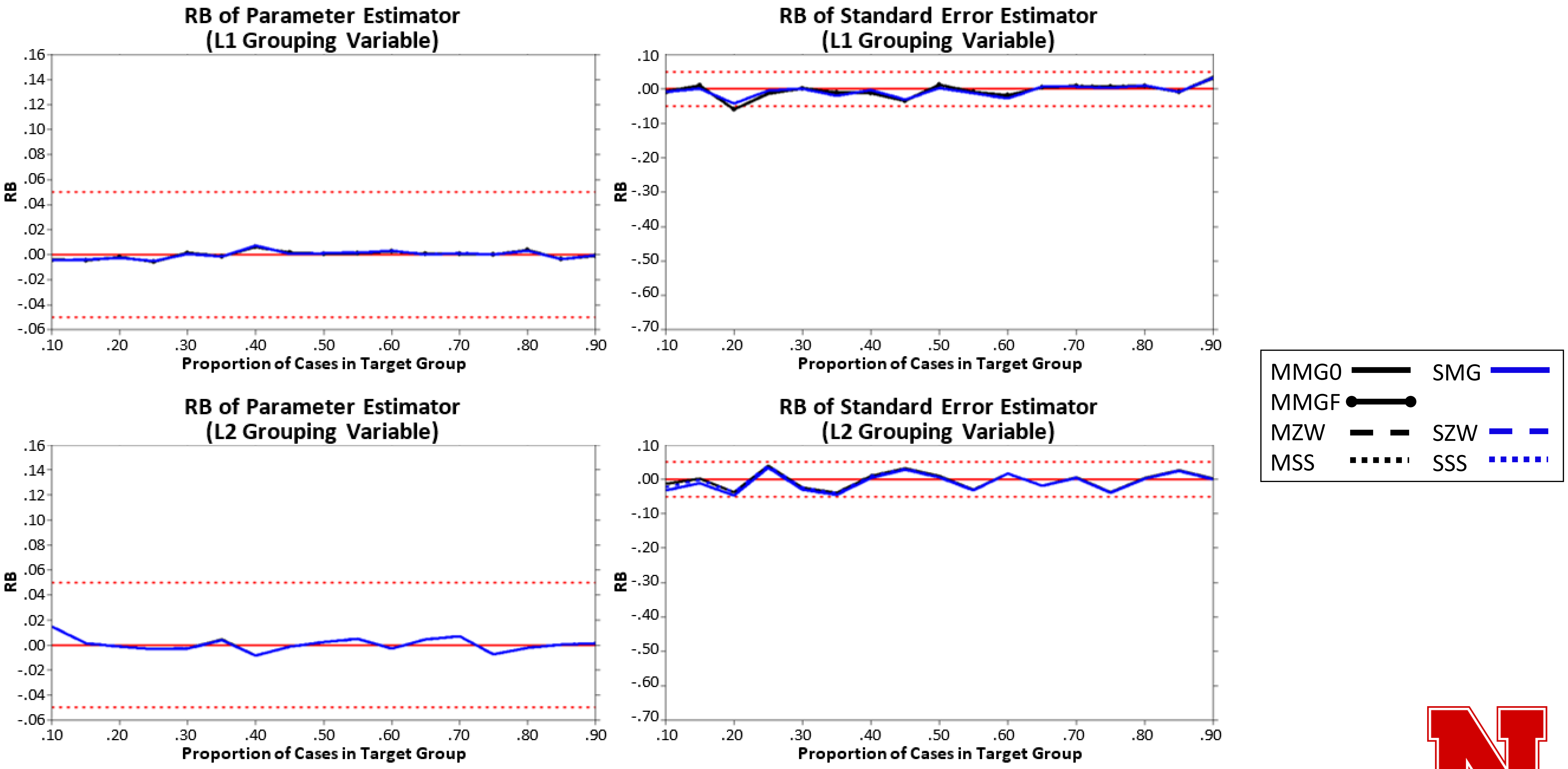
3) Repeat first two steps 1,000 times/condition

Results



Informative Design (weights)

Results



Non-Informative Design (no weights)

Discussion

Main Findings

Existing literature on subpopulation analysis cannot be blindly generalized to multilevel modeling

	PML	MPML
Differences between subsetting approach and other approaches	X	X
Differences between multiple-group and zero-weight approaches		X
Differences among approaches in variance estimation	X	X
Differences among approaches in point estimation		X
Differences among approaches when first stage design is altered	X	X
Differences among approaches when first stage design is unaltered		X
Sensitivity to cluster size		X

Discussion

Recommendations*

- Evaluate informativeness of design
  - Informative design (need sampling weights)
    - PML preferable to MPML when cluster sizes are small
    - For PML, multiple-group = zero-weight > subset
    - For MPML with L1 grouping, multiple-group > zero-weight > subset
    - For MPML with L2 grouping, zero-weight > multiple-group > subset
  - Non-informative design (omit sampling weights)
    - Single-level and multilevel methods both perform well
    - Differences among subpopulation approaches are trivial
- Compare approaches to evaluate robustness of conclusions

**Recommendations may not extend to conditions outside those examined in the present study.*
In particular, comparisons are more complex with non-Gaussian data.

References

Asparouhov, T., & Muthén, B. (2006). Multilevel modeling of complex survey data. *Proceedings of the Joint Statistical Meeting: ASA Section on Survey Research Methods*, 2718-2726.

Asparouhov, T., & Muthén, B. (2012). *Multiple group multilevel analysis* (Mplus Web Notes: No. 16). Los Angeles, CA: Muthén & Muthén.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics – Theory and Methods, 35*, 439-460.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*, 279-292.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Korn, E. L., & Graubard, B. I. (1999). *Analysis of health surveys*. New York, NY: John Wiley & Sons.

Koziol, N. A., Bovaird, J. A., & Suarez, S. (2017). A comparison of population-averaged and cluster-specific approaches in the context of unequal probabilities of selection. *Multivariate Behavioral Research*, 1-25 (advanced online publication).

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61*, 317-337.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York, NY: Springer-Verlag.

Scaling of Sampling Weights for Two Level Models in Mplus 4.2. (2008). Mplus Web Notes. Los Angeles, Muthén & Muthén.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 59-88). New York, NY: John Wiley & Sons.

Questions? Comments?

Corresponding author:
Natalie Koziol
nkoziol@unl.edu