Emerging Topics in Statistics and Biostatistics

Hon Keung Tony Ng
Daniel F. Heitjan *Editors*

Recent Advances on Sampling Methods and Educational Statistics

In Honor of S. Lynne Stokes

Springer

Emerging Topics in Statistics and Biostatistics

Hon Keung Tony Ng • Daniel F. Heitjan
Editors

Recent Advances on Sampling Methods and Educational Statistics

In Honor of S. Lynne Stokes

Springer

Editors

Hon Keung Tony Ng
Department of Mathematical Sciences
Bentley University
Waltham, MA, USA

Daniel F. Heitjan
Department of Statistical Science
Southern Methodist University
Dallas, TX, USA

ISSN 2524-7735 ISSN 2524-7743 (electronic)
Emerging Topics in Statistics and Biostatistics
ISBN 978-3-031-14524-7 ISBN 978-3-031-14525-4 (eBook)
https://doi.org/10.1007/978-3-031-14525-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

S. Lynne Stokes

Life and Works of S. Lynne Stokes

I was born on December 16, 1950, in Corsicana, the seat of Navarro County, Texas, where six generations of the Stokes family had lived. I was the second-born to a family of teachers. My dad taught mathematics and physics and coached the baseball team at Navarro Junior College, which had been established in 1946 as he and so many others were returning from WWII. My mother taught Spanish and agriculture, neither of which she had ever taken a course in, at the high school in her nearby hometown, Richland. In 1952, my parents decided to pull up roots and head to graduate school at Peabody College, now a part of Vanderbilt University, in Nashville, Tennessee. Their families were horrified that they would move so far away, and especially that a mom of two would take such an unconventional path. But the GI Bill had placed higher education within reach for many families who would now be called "first-generation," including mine. My parents went on to earn doctorates and have careers as college professors, he in mathematics and she in psychology. Their last and longest stint was at Austin Peay State University, where my dad chaired the Math Department for more than 20 years and my mom helped train a generation of school counselors in Clarksville, Tennessee. From this exposure and the joy they had in their careers, I decided at a young age that being a professor was my goal.

I studied mathematics at the University of the South in Suwanee, Tennessee. One of the faculty members, Mac Priestley, agreed to supervise me in an independent study out of Kemeny and Snell's book on Markov chains. From that experience, I decided that enrolling in a statistics PhD program was the right path for me, not realizing that it was actually probability I had been fascinated by. Luckily, I liked statistics even better, which I realized after joining the program at the University of North Carolina.

My years in Chapel Hill are among my fondest memories. My advisor, Norman Johnson, was endlessly encouraging and supportive. He asked me to read Dell and Clutter's 1972 ranked set sampling paper, then recently published, to see if I had any ideas on extensions for my dissertation work. Since that time, I have had the pleasure of discussing and collaborating with many on this topic, including several contributors to this volume.

My first job after school was in the Department of Mathematics at Vanderbilt, which was near my family home. I was one of only two statisticians in a large department. I soon decided I preferred real data and the company of other statisticians, and moved on to the Patuxent Wildlife Research Center in Laurel, Maryland. Patuxent was then a part of the US Fish and Wildlife Service and located in a 16,000-acre refuge of beautiful forest and wetlands in the midst of the Washington DC/Baltimore urban sprawl. There I learned from scratch about birds, and how to model bird-banding and capture-recapture data from the talented biometricians there, including Jim Nichols, a mentor and co-author. This is a skill I transferred from birds to people (at Census) and back to fish and the people who catch them (for NOAA) over the course of my career.

Patuxent changed my life in another way as well. There I met Dan Moulton, a biologist in the bird-banding lab, where he worked between field seasons on Laysan Island in Hawaii, where he was studying and banding Laysan ducks. During his second 6-month field season, we corresponded by letter and audio tape. These could be transported only by military plane or ship as they patrolled the Hawaiian archipelago. Soon after Dan returned from Laysan, we married.

While he was away, I left Patuxent for the US Census Bureau, which was just a short trip around the Beltway. Mary Mulry and I were hired into the Statistical Methods Division by Paul Biemer, whom we had first met at age 20 when all three of us were participants in an undergraduate NSF summer mathematics program at Texas A&M. Mary, Paul, and I have been colleagues, friends, and collaborators for 50 years, and we have NSF to thank for that.

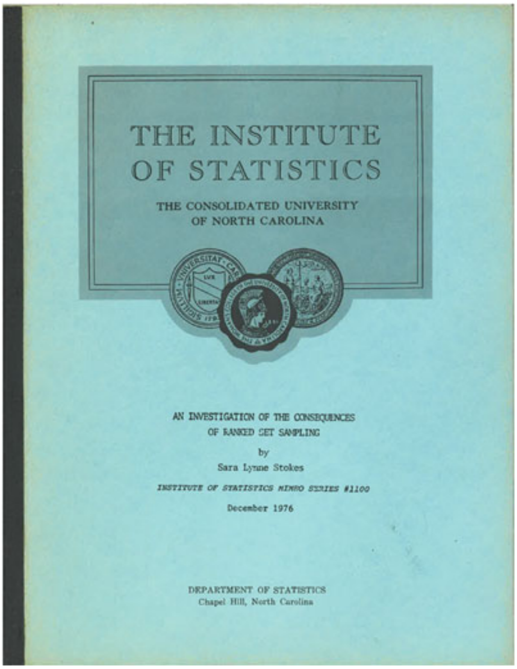
Paul had studied under H. O. Hartley, and he introduced me to sampling theory and measurement error methods. The Census Bureau provided an unlimited supply of real-life problems for non-sampling error research, which has remained a lifelong interest. Fortunately for my career, errors occur whenever data are collected. This allowed me to dabble in many fascinating application areas over the years. Two of these areas, fisheries and education surveys, are well represented in this volume (Brick, Andrews & Foster; Becker & Gozutok).

When Dan took a position at Texas Parks and Wildlife in Austin in 1983, I moved with him and worked remotely for Census, before that was common. This arrangement was facilitated with the help of Kent Marquis, my division chief at Census, and Carl Morris, then in the Mathematics Department at the University of Texas. Kent and Carl had known each other at Rand, proving once again that it helps to get lucky. Soon a faculty position opened for a statistician in the Management Science Department at UT's Business School, and I was again in the right place at the right time. In my 15 years at UT, I expanded the range of problems I worked on with colleagues in fields from finance to demography to operations research.

In 2001, I left UT for the Statistics Department at Southern Methodist University, after a convincing chat with my long-time acquaintance Bill Schucany. I had first met Bill at a Conference of Texas Statisticians meeting shortly after moving to Texas, and had received useful advice from him over the years. SMU was a perfect place for the last 20 years of my career, providing a helpful administration, supportive colleagues, and excellent graduate students. I chaired the department for one term, and then became the inaugural director of SMU's Data Science Institute in my last 2 years there. Several of the contributors to this volume are cherished colleagues and former students from SMU.

My path likely would not have been so straight and well-marked if it had not been for the opportunities that began to open up for women at just the right time for me. I also benefited from introductions provided by supportive male mentors, colleagues, and classmates. I entered the University of the South the first year they accepted women (1969). My entering cohort in the Statistics Department at UNC in 1972 were half women and half men, marking the first year that women who were not wives of students were admitted in significant numbers. I was the fourth woman to receive a PhD in statistics at UNC, three of whom were supervised by Norman Johnson, who may have been influenced by his wife Regina from the UNC Biostatistics Department. At Vanderbilt, I was the first woman to fill a tenure-track position in the Mathematics Department, and at SMU, the first woman chair of the Statistics Department. My network-building began in the NSF program I attended as a 20 year old, which I believe illustrates the value of promoting diversity in such programs for young scholars.

Dallas, TX, USA S. Lynne Stokes
May 2022



THE INSTITUTE OF STATISTICS

THE CONSOLIDATED UNIVERSITY
OF NORTH CAROLINA

AN INVESTIGATION OF THE CONSEQUENCES
OF RANKED SET SAMPLING

by

Sara Lynne Stokes

INSTITUTE OF STATISTICS MIMEO SERIES #1100

December 1976

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

S. Lynne Stokes's PhD thesis

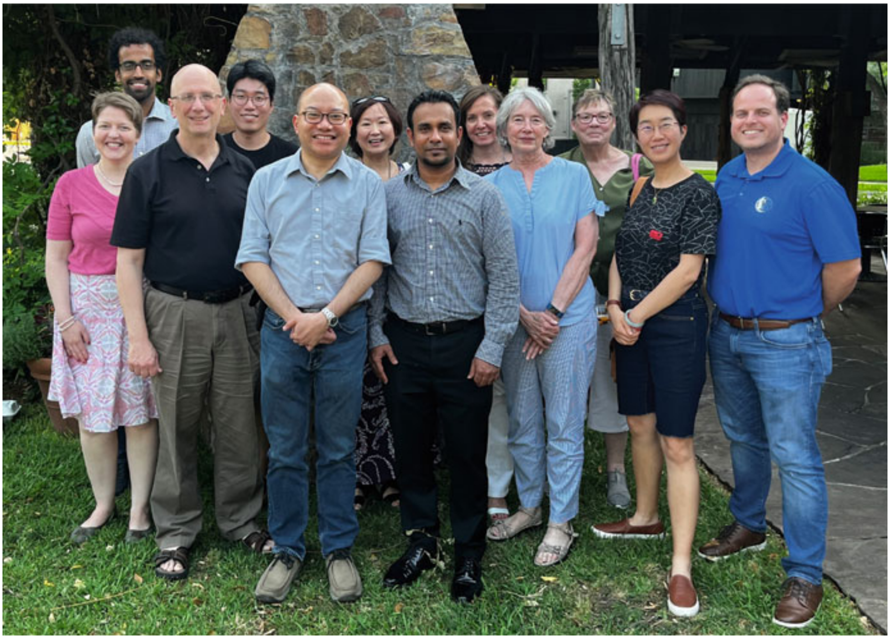
Lynne enjoying the snow at Patuxent Wildlife Research Center, circa 1980

Mary Mulry, Paul Biemer, and Lynne at an NSF program reunion circa 1980

Lynne enjoying Friday morning teatime at SMU in 2007

Celebrating Betsy Becker's election to Fellow at 2008 JSM with an Educational Statistics mentor for both of us, Ingram Olkin

Helena Jia, Lynne, and Bingchen Liu in downtown Princeton during a meeting at ETS in 2017



From left to right: Jessica Wickersham, Raanju R. Sundararajan, Daniel F. Heitjan, Chul Moon, Hon Keung Tony Ng, Xinlei (Sherry) Wang, Mahesh Fernando, Monnie McGee, S. Lynne Stokes, Sheila Crain, Jing Cao, and Charles South in Dallas, Texas, during a department faculty gathering in May 2022

Awards, Honors, and Publications of S. Lynne Stokes

Awards and Honors

- Caren Prothro Faculty Service Award, Southern Methodist University (2019)
- Founders Award, American Statistical Association (2013)
- Dedman Family Distinguished Professor, Southern Methodist University (2013)
- United Methodist Church University Scholar/Teacher of the Year Award (2011)
- Don Owen Award, American Statistical Association, San Antonio Chapter (2005)
- Fellow of the American Statistical Association (1998)
- Phi Beta Kappa
- Sigma Pi Sigma

Publications

Refereed Journals and Proceedings

1. “Investigating Record Linkage for Combining Voluntary Catch Reports with a Probability Sample,” (B. Williams, L. Stokes, and J. Foster), *Fisheries Research*, 251, 106301 (2022).
2. “Predictive modeling of maximum injury severity and potential economic cost in a car accident based on the General estimates system data,” (G. Alkan, R. Farrow, H. Liu, C. Moore, H.K.T. Ng, S. L. Stokes, Y. Xu, Z. Xu, Y. Yan, and Y. Zhang), *Computational Statistics*, 36, 1561–1575 (2021).
3. “The Impact of non-sampling errors on estimators of catch from electronic reporting Systems,” (L. Stokes, B. Williams, R. McShane, and S. Zalsha), *Journal of Survey Statistics and Methodology*, 9, 159–184 (2021).

4. “Prevalence of Sexual Victimization among Female and Male College Students: A Methodological Note with Data,” (Jouriles, E. N., Nguyen, J., Krauss, A., Stokes, S. L., and McDonald, R.), *Journal of Interpersonal Violence*, (2020).
5. “A method to correct for frame membership error in dual frame estimators,” (D. Lin, Z. Liu, and L. Stokes), *Survey Methodology*, 45, 543–565 (2019).
6. “Accumulating Evidence of the Impact of Voter ID Laws: Student Engagement in the Political Process,” (K. S. McConville, L. Stokes, and M. Gray), *Statistics and Public Policy*, 5, 1–8 (2018).
7. “Cross-Cultural Issues in Teaching Ethics in a Statistics Curriculum,” (A. Elliott, L. Stokes, and J. Cao) *The American Statistician*, 72, 359–367 (2018).
8. “Comparison of Different Ranking Methods in Wine Tasting,” (J. Cao and S.L. Stokes), *Journal of Wine Economics*, 12, 203–210 (2017).
9. “Estimation of total from a population of unknown size and application to estimating recreational red snapper catch in Texas,” (B. Liu, S.L. Stokes, T. Topping, and G. Stunz), *Journal of Survey Statistics and Methodology*, 5, 350–371 (2017).
10. “Just in time teaching in Statistics Classrooms,” (M. McGee, L. Stokes, and P. Nadolsky), *Journal of Statistics Education*, 24, 16–26 (2016).
11. “A power analysis for fidelity measurement sample size determination,” (L. Stokes and J. Allor) Psychological Methods, 21, 35–46 (2016).
12. Using Ranked Set Sampling with Cluster Randomized Designs for Improved Inference on Treatment Effects.” (X. Wang, J. Lim, and L. Stokes), *Journal of the American Statistical Association*, 111, 1576–1590 (2016).
13. “Analyses of Wine Tasting Data: A Tutorial,” (I. Olkin, Y. Lou, L. Stokes, and J. Cao), *Journal of Wine Economics*, 10, 4–30 (2015).
14. “The National Children’s Study 2014: Commentary on a Recent National Research Council/Institute of Medicine Report Academic Pediatrics,” *Academic Pediatrics*, 14, 545–546 (2014).
15. “Sample Size Calculation for a Hypothesis Test,” (L. Stokes), *Journal of the American Medical Association*, 312, 180–181 (2014).
16. “Kernel Density Estimator from Ranked Set Samples,” (X. Wang, J. Lim, M. Chen, and L. Stokes), *Communications in Statistics – - Theory and Methods*, 43, 2156–2168 (2014).
17. “Methods for Improving Response Rates in Two-Phase Mail Surveys,” (M. Brick, W. Andrews, P. Brick, H. King, N. Mathiowetz, and L. Stokes), *Survey Practice*, 5, 1–6. (2012).
18. “Stranger at the Gate: the Effect of the Plaintiff’s use of an Interpreter on Juror Decision-Making,” (D. Shuman, L. Stokes, and G. Martinez), *Behavioral Sciences and the Law*, 29, 499–512 (2011).
19. “Performance of Weighted Random Effects Model Estimators under Complex Sampling Designs,” (Y. Jia, L. Stokes, I. Harris, and Y. Wang), *Journal of Educational and Behavioral Statistics*, 36, 6–32 (2011).
20. “Evaluation of Wine Judge Performance through Three Characteristics: Bias, Discrimination, and Variation,” (J. Cao and L. Stokes), *Journal of Wine Economics*, 5, 1–11. (2010)

21. "A Bayesian Approach to Ranking and Rater Evaluation: an Application to Grant Reviews," (J. Cao, S. Zhang, and L. Stokes), *Journal of Educational and Behavioral Statistics*, 35, 194–215. (2010).
22. "Data Masking for Disclosure Limitation," (L. Stokes and G. Duncan), *Wiley Interdisciplinary Reviews: Computational Statistics*, 1, 1–10 (2009).
23. "Bayesian IRT guessing models for partial guessing behaviors," (J. Cao and L. Stokes), *Psychometrika*, 73, 209–230 (2008).
24. "A Nonparametric Mean Estimator for Judgment Post-Stratified Data," (X. Wang, J. Lim, and L. Stokes), *Biometrics*, 64, 355–363 (2008).
25. "Judgment Post-Stratification with Multiple Rankers," (L. Stokes, X. Wang, and M. Chen), *Journal of Statistical Theory and Applications*, 6, 344–359 (2007).
26. "Concomitants of multivariate order statistics with application to judgment post-stratification," (X. Wang, L. Stokes, J. Lim, and M. Chen), *Journal of the American Statistical Association*, 101, 1693–1704 (2006).
27. "Forming Post-Strata via Bayesian Treed Capture-Recapture Models," (X. Wang, J. Lim, and L. Stokes), *Biometrika*, 93, 861–876, (2006).
28. "An Estimator of Number of Species from Quadrat Sampling," (P. Haas, Y. Liu, and L. Stokes), *Biometrics*, 62, 135–141 (2006).
29. "Antecedents and consequences of residential choice and school transfer," (T. Falbo, R. Glover, L. Holcombe, and L. Stokes), *Education Policy Analysis Archives*, 13, 29 (2005).
30. "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding," (G. Duncan and L. Stokes), *Chance*, 17, 16–20 (2004).
31. "Using Spreadsheet Solvers in Sample Design" (L. Stokes and J. Plummer), *Computational Statistics and Data Analysis*, 44, 527–546 (2004).
32. "Using Auxiliary Information for Improving Estimation in the Number of Species Problem," *Statistica Sinica*, 13, 655–671 (2003).
33. "Comment on 'Can a Statistician Deliver?' " *Journal of Official Statistics*, 17, 103–106 (2001).
34. "Acceptance Sampling with Rectification when Classification Errors are Present," (M. Anderson, B. Greenberg, and L. Stokes), *Journal of Quality Technology*, 33, 493–505 (2001).
35. "Editorial: Special issue on Statistical Design and Analysis with Ranked Set Samples," (N. P. Ross and L. Stokes), *Environmental and Ecological Statistics*, 6, 1–6 (1999).
36. "Estimating the Number of Classes in a Finite Population" (P. Haas and L. Stokes), *Journal of the American Statistical Association*, 93, 1475–1487 (1998).
37. "Success rate with repeated cycles of in vitro fertilization-embryo transfer," (D. Meldrum, K. Silverberg, M. Bustillo, and L. Stokes), *Fertility and Sterility*, 69, 1005–1009 (1998).
38. "Do Product Warnings Increase Safe Behavior?: A Meta Analysis" (E. Cox, L. Stokes, E. Murff), *Journal of Public Policy and Marketing*, 25, 195–204 (1997).

39. "Estimation of the CDF of a Finite Population using a Calibration Sample" (M. Luo, L. Stokes, and T. Sager), *Environmental and Ecological Statistics*, 15, 346–352 (1997).
40. "Considerations of Cost Trade-Offs in Insurance Solvency Surveillance Policy" (J. Lamm-Tennant, L. Starks, L. Stokes), *Journal of Banking and Finance*, 20, 835–852 (1996).
41. "Repetitive Testing in the Presence of Inspection Errors," (B. Greenberg and L. Stokes), *Technometrics*, 37, 102–111 (1995).
42. "Sampling-Based Estimation of the Number of Distinct Values of an Attribute," (P. Haas, J. Naughton, S. Sehadri, and L. Stokes), *VLDB 95: Proceedings of the International Conference on Very large Databases* (U. Dayal, P. Gray, S. Nishio, Eds.), 311–322 (1995).
43. "Parametric Ranked Set Sampling," (L. Stokes), *Annals of the Institute of Statistical Mathematics*, 47, 465–482 (1995).
44. "Reliability of Coherence of Causal, Diagnostic, and Joint Subjective Probabilities," (K. Wright, L. Stokes and J. Dyer), *Decision Sciences*, 25, 691–709 (1994).
45. "Estimating Nonconformance Rate after Zero-Defect Sampling with Rectification," (B. Greenberg and L. Stokes), *Technometrics*, 34, 203–215 (1992).
46. "An Empirical Bayes Approach to Estimating Loss Ratios," (J. Lamm-Tennant, L. Starks, and L. Stokes), *Journal of Risk and Insurance*, 59, 426–442 (1992).
47. "Estimating the Size of a Subdomain: An Application in Auditing," (L. Stokes), *Journal of Business and Economic Statistics*, 8, 337–346 (1990).
48. "Developing an Optimal Call Scheduling Strategy for a Telephone Survey" (B. Greenberg and L. Stokes), *Journal of Official Statistics*, 6, 421–435 (1990).
49. "The Optimal Design of Quality Control Samples to Detect Interviewer Cheating," (P. Biemer and L. Stokes), *Journal of Official Statistics*, 5, 23–40 (1989).
50. "Combining Multiple Risk Assessments for Construction Risk Identification," (D. Ashley, L. Stokes, and Y.H. Peng), *Proceedings of the 7th International Conference on Offshore Mechanics and Arctic Engineering Conference*, American Society of Mechanical Engineers, 183–192 (1988).
51. "Characterization of a Ranked Set Sample with Application to Estimating Distribution Functions," (L. Stokes and T. Sager), *Journal of the American Statistical Association*, 83, 374–381 (1988).
52. "Estimation of Interviewer Effects for Categorical Items in a Random Digit Dial Telephone Survey," (S. L. Stokes), *Journal of the American Statistical Association*, 83, 623–630 (1988).
53. "Estimation of the Correlated Component of Response Variance for Categorical Variables," (L. Stokes and M. Mulry), *Journal of Official Statistics*, 3, 389–401 (1987).
54. "Optimal Design of Interviewer Variance Estimates in Complex Surveys," (P. Biemer and L. Stokes), *Journal of the American Statistical Association*, 80, 158–166 (1985).

55. “The Jolly-Seber Method Applied to Age-Stratified Populations,” (S. L. Stokes), Journal of Wildlife Management, 48, 1053–1059 (1984).
56. “Additional Comments on the Assumption of Homogeneous Survival Rates in Modern Bird Banding Estimation Models,” (J. Nichols, L. Stokes, J. Hines, and M. Conroy), Journal of Wildlife Management, 46, 953–962 (1982).
57. “Remarks on the Use of Mark-recapture Methodology in Estimating Avian Population Size” (J. Nichols, B. Noon, L. Stokes, and J. Hines), *Studies in Avian Biology*, 6, 121–136 (1981).
58. “Estimation of Variance Using Judgment Ordered Ranked Set Samples,” (S. L. Stokes), *Biometrics*, 36, 35–42 (1980).
59. “Inferences on the Correlation Coefficient in Bivariate Normal Populations from Ranked Set Samples,” (S. L. Stokes), *Journal of the American Statistical Association*, 75, 989–995 (1980).
60. “Ranked Set Sampling with Concomitant Variables,” (S. L. Stokes), *Communications in Statistics – Theory and Methods*, 6, 1207–1211 (1977).

Book Chapters

1. “Measuring treatment fidelity with reliability and validity across a program of intervention research: Practical and theoretical considerations,” (Allor, J. H. and Stokes, L.), In G. Roberts, S. Vaughn, S. N. Beretvas, and V. Wong (Eds.), *Measuring and Modeling Treatment Fidelity in Studies of Educational Intervention*, New York: Routledge Taylor & Francis Group (2017).
2. “Interviewer Effects,” in *Encyclopedia of Research Methods for the Social Sciences*, M. Lewis-Beck, A. Brayman and T.F. Liao, Editors, Sage Publications (2003).
3. “Identifying and Adjusting for Recall Error with Application to Fertility Surveys,” (T. Pullum and L. Stokes), Chapter 31 (pp. 711–732), *Survey Measurement and Process Quality*, John Wiley and Sons (1997).
4. “A Cost-Effective Approach for Regulating Insurance Company Solvency,” (J. Lamm-Tennant, L. Starks, and L. Stokes), in *The Financial Dynamics of the Insurance Industry*, 153–167, E.I Altman and I.T. Vanderhoof, Editors, Irwin Professional Publishing, New York (1995).
5. “Some Recent Results on the Modeling and Estimation of Measurement Errors in Surveys,” (with P. Biemer and L. Stokes), Chapter 24 (pp. 487–516) in *Measurement Errors in Surveys*, John Wiley & Sons (1991).
6. “A New Approach to Identifying Sources of Interviewer Effects in Telephone Surveys,” (L. Stokes and M. Yeh), Chapter 22 (pp. 357–373) in *Telephone Survey Methodology*, Robert M. Groves, Editor, John Wiley and Sons, Inc. (1988).
7. “Ranked Set Sampling,” in *Encyclopedia of Statistical Sciences*, N. Johnson and S. Kotz, Editors, John Wiley & Sons, 585–588 (1986).

Preface

When our colleague Lynne Stokes announced her intention to transition to emerita status at the end of the 2022 academic year, our initial reactions were dismay—at losing a valued colleague—and surprise—that she would walk away while still at the top of her game. How can you retire, Lynne; what will you do? And what will our department do without you?

After reconciling ourselves to the coming new reality, we decided that we should do something special to commemorate Lynne's remarkable career and recognize this momentous life change. A symposium, we thought—but Lynne said she did not want a symposium. Well then, a party hosting current and past colleagues and students. No, Lynne said, no party. Perhaps an intimate dinner with the faculty? No again. A Texas barbecue? A Lynne-themed Friday tea time? No and no. Well how about a *festschrift*?

And that is how this book came to be.

So we made the rounds of Lynne's many students, co-authors, and past and current colleagues, who were universally eager to contribute papers in areas where she has worked over the years. We express our sincere gratitude to all of them for writing chapters of such high quality on a tight deadline. Special thanks are also due to the referees, many of them authors as well, for their constructive reviews. And we acknowledge the team from Springer Nature Group—Laura Aileen Briskman, Kirthika Selvaraju, Faith Su, and Amelie von Zumbusch—who have gently guided the project from inception to production.

Most importantly, we are grateful to our colleague and friend Lynne Stokes for blessing this work and for supporting our efforts with her characteristic energy, generosity, and humility. It is our great pleasure to present her with this book on the occasion of her transition to the next phase of a most interesting and well-lived life.

Waltham, MA, USA — Hon Keung Tony Ng
Dallas, TX, USA — Daniel F. Heitjan
June 2022

Contents

Part III Educational and Behavioral Statistics

Contributors

Soohyun Ahn Department of Mathematics, Ajou University, Gyeonggi, Korea

Talha Alvi Department of Psychology, Southern Methodist University, Dallas, TX, USA

William R. Andrews NOAA Fisheries, Silver Spring, MD, USA

Betsy Jane Becker College of Education, Florida State University, Tallahassee, FL, USA

Paul P. Biemer RTI International, Research Triangle Park, NC, USA

J. Michael Brick Westat, Rockville, MD, USA

G. Gordon Brown SAS Institute, Cary, NC, USA

Jennifer Brown University of Canterbury, School of Mathematics and Statistics, Christchurch, New Zealand

Jing Cao Department of Statistical Science, Southern Methodist University, Dallas, TX, USA

John Foster NOAA Fisheries, Silver Spring, MD, USA

Ahmet Serhat Gözütok Measurement and Evaluation in Education Educational Sciences, Ereğli Faculty of Education, Zonguldak Bülent Ecevit University, Zonguldak, Turkey

Yue Jia Educational Testing Service, Princeton, NJ, USA

Jiae Kim Department of Statistics, Indiana University, Bloomington, IN, USA

Olena Kravchuk University of Adelaide, School of Agriculture, Food and Wine, Adelaide, SA, Australia

Yi-Hsuan Lee Educational Testing Service, Princeton, NJ, USA

Johan Lim Department of Statistics, Seoul National University, Seoul, Korea

Bingchen Liu Educational Testing Service, Princeton, NJ, USA

Steven N. MacEachern Department of Statistics, The Ohio State University, Columbus, OH, USA

John Mazzeo Educational Testing Service, Princeton, NJ, USA

Vincent T. Mule Jr. U.S. Census Bureau, Suitland, MD, USA

Mary H. Mulry U.S. Census Bureau, Suitland, MD, USA

James D. Nichols Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, FL, USA

Linh H. Nghiem School of Mathematics and Statistics, University of Sydney, Sydney, NSW, Australia
Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT, Australia

Omer Ozturk Department of Statistics, The Ohio State University, Columbus, OH, USA

Bivin Philip Sadler Master of Science in Data Science, Southern Methodist University, Dallas, TX, USA

S. Lynne Stokes Department of Statistical Science, Southern Methodist University, Dallas, TX, USA

Benjamin A. Tabak Department of Psychology, Southern Methodist University, Dallas, TX, USA

Zachary Wallmark Department of Musicology and Ethnomusicology, University of Oregon, Eugene, OR, USA

Xinlei Wang Department of Statistical Science, Southern Methodist University, Dallas, TX, USA

Christopher Wiesen The Odum Institute for Research in Social Sciences, University of North Carolina, Chapel Hill, NC, USA

Benjamin Williams Department of Business Information and Analytics, Daniels College of Business, University of Denver, Denver, CO, USA

Part I
Ranked-Set Sampling, Judgement Post-stratified Sampling, and Capture-Recapture Methods

Predictive Modelling and Judgement Post-stratification

Steven N. MacEachern and Jiae Kim

Abstract Predictive modelling has come to the forefront of statistics in recent years as interest in forecasting the results of experiments and interventions has increased. We now routinely see forecasts in the news media that include point predictions, an assessment of variation to accompany the prediction and even a full predictive distribution. In the area of ranked set sampling, Stokes and coauthors' work on the use of measured order statistics, and their concomitants provided a crucial step that allows one to pass from the subjective assessment of ranks of responses within a set to the use of covariates. The transition also allows one to make use of formal models for a response given measured covariates to improve upon the basic ranked set sampling estimators while retaining the robustness properties of the method. This chapter pursues the use of predictive distributions in the context of ranked set sampling. We find that the predictive viewpoint naturally leads us away from imposing a strict ranking on the units in a set to expressing a distribution over ranks for each unit in the set. In turn, this change suggests the use of judgement post-stratification rather than ranked set sampling. It also yields novel estimators which are shown to outperform the standard estimators.

1 Ranked Set Sampling and Judgement Post-stratification

Stokes' pioneering work (Stokes, 1977) brought measured covariates to ranked set sampling (RSS). Briefly restating her work and establishing notation, consider a set of nH units that are partitioned at random into n sets, each of size H. The units are presumed to form a random sample from some distribution. Within a given set, we

S. N. MacEachern (✉)
Department of Statistics, The Ohio State University, Columbus, OH, USA
e-mail: snm@stat.osu.edu

J. Kim
Department of Statistics, Indiana University, Bloomington, IN, USA
e-mail: jiaekim@iu.edu

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_1

begin with $(X_h, Y_h), h = 1, \ldots, H$. These units are ranked on the X_h, so that $X_{(r:H)}$ is the rth order statistic in the set. The measured response, $Y_{[r:H]}$, associated with this unit is its concomitant. To draw a RSS of size n from such a population, sample sizes n_h, $h = 1, \ldots, H$, are specified, with $\sum_{h=1}^{H} n_h = n$. One unit is drawn from each of the n sets; in n_h sets, the unit ranked h is selected. The resulting sample is a RSS.

The earliest description of RSS appears in McIntyre (1952) (republished as McIntyre, 2005). In McIntyre's description of the technique, ranking is based on the subjective judgement of an experimenter who examines each set of H units, specifying the ranks of the units in the set. Once the units in each set have been ranked, the sample is drawn as described above and the response of interest, Y, is measured on the n sampled units. Extending our notation to capture both set and rank within set, the mean of the nH units is

$$\bar{Y} = (nH)^{-1} \sum_{i=1}^{n} \sum_{h=1}^{H} Y_{ih} \ , \tag{1}$$

where Y_{ih} is the response of the unit with rank h in set i. Suppressing the notation for the rank, define Y_i to the be ith of the n sampled units. Provided $n_h > 1$ for all h,

$$\bar{Y}_{rss} = H^{-1} \sum_{h=1}^{H} \bar{Y}_h \ , \tag{2}$$

where $\bar{Y}_h$ is the sample mean of the n_h sampled units with rank h. The RSS estimator is unbiased: $E[\bar{Y}_{rss} \mid \bar{Y}] = \bar{Y}$ for any collection of nH units. Furthermore, when the units are a random sample from a distribution with mean $\mu = E[Y]$, $E[\bar{Y}_{rss}] = E[\bar{Y}] = \mu$. The goal of RSS is to estimate μ. Stokes and Sager (1988) cast estimation of a cumulative distribution function as estimation of a proportion (mean) for all cut points on the real line.

RSS with estimation following (2) is robust to variation in the specifics of how the ranks are created. When created subjectively, better ranking leads to greater separation of the means of the rank classes (or strata), in turn leading to greater reduction in variance relative to estimators based on a random sample from the population. When ranks arise from a measured covariate, the same holds. Sound experimental practice includes blinding the ranker to which units will be fully measured. When implemented, the estimator is unbiased for μ as long as the ranks can be determined before the responses of the selected units are measured. If the ranking process makes modest use of the measured responses, the bias is typically small.

Judgement post-stratification (JPS) is a common variant of ranked set sampling. To draw a JPS sample, a collection of n units is selected for full measurement (both X and Y) from the population as described above. For each fully measured unit, a set is filled out by independently drawing an additional $H - 1$ units. For these supplemental units, only X is measured. The end result is n sets of H units.

Within a set, X_h is measured for all H units, while Y_h is measured for a single unit. Upon ranking the units, conceptually, we have the pairs $(X_{(r:H)}, Y_{[r:H]})$, for $r = 1, \ldots, H$. In practice, most of the responses are missing and we have only one measured response, $Y_{[r:H]}$, for some r. When units are ranked on the basis of a measured covariate, the name *judgement* post-stratification is a misnomer. The name stems from the original work on the technique (MacEachern et al., 2004) where ranking was based on subjective judgement about the units.

An equivalent description of JPS exists. As in the RSS, we could form n sets, each consisting of H units. Instead of specifying the n_h, $h = 1, \ldots, H$, we select a single unit at random from each set for full measurement. With ranks based on the measured covariate, the n_h, $h = 1, \ldots, H$, are random variables. The vector $(n_1, \ldots, n_H)$ follows a multinomial distribution with n trials and parameter vector $(1/H, \ldots, 1/H)$.

Whichever description of JPS is used, the data that are used for estimation consist of n independent and identically distributed (IID) vectors (Y_i, R_i), where Y_i is the measured value and R_i is the rank of the unit within its set. For estimation, we parallel the technique of post-stratification from survey sampling. Conditioning on the observed n_h and using the estimator in Eq. (2), an estimator for μ can be obtained as

$$\hat{\mu}_{jps1} = H^{-1} \sum_{h=1}^{H} \bar{Y}_h = H^{-1} \frac{\sum_{i=1}^{n} Y_i I_{ih}}{\sum_{i=1}^{n} I_{ih}} , \tag{3}$$

where shorthand notation has $I_{ih} = I(R_i = h)$. The within-rank sample size is $n_h = \sum_{i=1}^{n} I_{ih}$, $h = 1, \ldots, H$. The resulting estimator is unbiased for μ, conditional on all of the $n_h > 0$. Various patches exist to define the estimator when one or more $n_h = 0$. Frey and Feeman (2012) and Frey (2016) developed methods to reduce the mean square error of $\hat{\mu}_{jps1}$ by allowing some conditional bias in the estimator.

To extend the technique to two rankers, the data used for estimation consist of the vectors (Y_i, R_{1i}, R_{2i}). The information from both rankers is used to form the estimator

$$\hat{\mu}_{jps2} = H^{-1} \sum_{h=1}^{H} \frac{\sum_{i=1}^{n} Y_i p_{ih}}{\sum_{i=1}^{n} p_{ih}} , \tag{4}$$

where $p_{ih} = [I(R_{1i} = h) + I(R_{2i} = h)]/2$. The notation p_{ih} reflects an empirical estimate of the probability that the ith fully measured unit has rank h. The method is easily extended to more than two rankers and to rankers of differing quality.

The move from RSS to JPS has several advantages. For one, it allows the experimenter to use a conventional design (based on a random sample from the population), with estimates improved by the use of covariates measured on additional units. A second advantage is that JPS can be used in situations where the units are not actually ranked. This may be due to disagreements between multiple rankers as in MacEachern et al. (2004), or it may be due to the presence of more than one informative covariate, as in Wang et al. (2006). Wolfe provided an insightful review of RSS, JPS and related techniques (Wolfe, 2012).

2 Multivariate Order Statistics and JPS

In Wang et al. (2006), Stokes and coauthors posed the intriguing question of how to use multiple covariates to convey information about the ranks of units for use in JPS. Their solution is to rank on each of the distinct covariates. In the case of a continuous bivariate covariate, (X_1, X_2), each of the units in the set would be assigned a pair of ranks—one for X_1 and the other for X_2. This pair of ranks defines the post-stratum (or rank class) of the unit. For a set of size H, there are H^2 post-strata. We denote these post-strata with $\mathbf{r} = (r_1, r_2)$, where $r_1, r_2 \in \{1, \ldots, H\}$. We focus on a bivariate covariate but note that the technique extends to covariates of greater dimension. Figure 1 illustrates the situation for a bivariate order statistic for set size $H = 5$.

The increase in the number of post-strata from H to H^2 necessitates reconsideration of the basic post-stratification estimator (3). Marginally, each covariate for the measured unit will have rank $r_i = h$ with probability $1/H$ for $i = 1, 2$ and $h = 1, \ldots, H$. The joint distribution of $\mathbf{R}$ leads to the stratum probability $\pi_{\mathbf{r}} = P(\mathbf{R} = \mathbf{r})$. In general, these probabilities can be found via numerical integration if the model for (X_1, X_2) is fully specified. Some of the $\pi_{\mathbf{r}}$ may be much smaller than H^{-2}, leading to a large probability that the estimator is undefined.

Wang et al. (2006) handled this issue by appealing to a parametric model as an aid to estimation. The authors defined $\mu_{[\mathbf{r}]} = E[Y \mid \mathbf{R} = \mathbf{r}]$. The value of $\mu_{[\mathbf{r}]}$ can be found by numerical integration over the conditional distribution of $Y \mid \mathbf{R}$. Once the

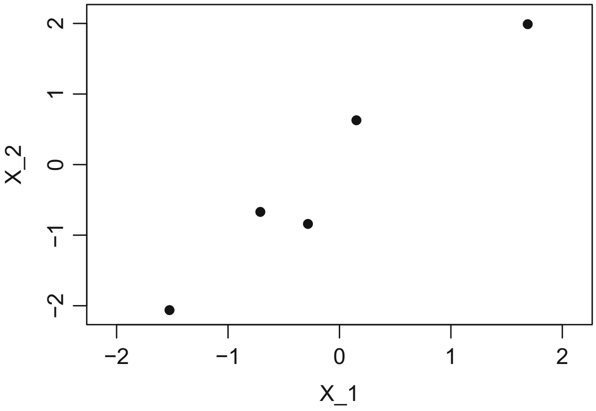


Fig. 1 Covariate pairs for a set of size $H = 5$. The bivariate rank vectors are (1, 1), (2, 3), (3, 2), (4, 4), and (5, 5). The ranks based on X_1 and X_2 agree for three of the five items and disagree for two. Extreme differences in the ranks may be very rare

stratum means are in place, they are connected to the mean of Y via the expression $\mu = \sum_{\mathbf{r}} \pi_{\mathbf{r}} \mu_{[\mathbf{r}]}$. It is helpful to introduce the difference between the stratum mean and the overall mean, $\delta_{[\mathbf{r}]} = \mu_{[\mathbf{r}]} - \mu$. The authors suggested estimation by ordinary least squares applied to a model for μ, with observations in stratum $\mathbf{r}$ offset by $\delta_{[\mathbf{r}]}$. The data are $(Y_i, \mathbf{r}_i)$, $i = 1, \ldots, n$, and the estimator is

$$\hat{\mu}_{oLS} = n^{-1} \sum_{i=1}^{n} (Y_i - \delta_{[\mathbf{r}_i]}) \ . \tag{5}$$

The estimator $\hat{\mu}_{oLS}$ can be viewed in two stages: In the first, each observation is bias-corrected by subtracting its $\delta_{[\mathbf{r}]}$; in the second, the sample mean of the bias-corrected observations is computed. Partitioning the sample into strata reduces the within-stratum variances. Removing bias and then using the sample mean ensures that each observation receives equal weight in the estimator. Together, these two stages lead to substantial variance reduction, especially for relatively large set sizes.

In a refinement, Wang et al. (2006) suggested consideration of a weighted least squares estimator that takes within-stratum variances into account. The within-stratum variances are computed on the basis of numerical integration. This estimator takes the form

$$\hat{\mu}_{wLS} = \frac{\sum_{i=1}^{n} \sigma_{\mathbf{r}_i}^{-2} (Y_i - \delta_{[\mathbf{r}_i]})}{\sum_{i=1}^{n} \sigma_{\mathbf{r}_i}^{-2}} \ . \tag{6}$$

In the event that the $\delta_{[\mathbf{r}]}$ and $\sigma_{\mathbf{r}_i}^{2}$ are estimated, we place hats over them to denote this. In the framework of bias-corrected estimators, $\hat{\mu}_{oLS}$ and $\hat{\mu}_{wLS}$ are excellent performers—the mean and an optimally weighted mean. Wang et al. (2006) demonstrated superior performance of these new estimators when the class of models (multivariate normal distributions) is correct and the parameters in the model are known or are estimated.

The theory developed in Wang et al. (2006) implies that the weighted average of the offsets is zero for every model for which μ exists. That is,

$$\sum_{\mathbf{r}} \pi_{\mathbf{r}} \delta_{[\mathbf{r}]} = 0 \ . \tag{7}$$

This is a delicate expression, as it is naturally satisfied when both the $\pi_{\mathbf{r}}$ and the $\delta_{[\mathbf{r}]}$ are correctly specified. Asymptotically, we expect the expression to hold if we replace these two quantities with consistent estimators of them. If not, one would expect the expression (7) to evaluate to something other than 0, leaving us with a Fisher-consistent estimator of a quantity near, but not exactly equal to, μ.[1]

[1] Huber (1981), in his study of robustness, found a need to redefine consistency when the distribution that generates the data might not lie in a tidy parametric family. His definition of Fisher consistency focuses on functionals of the empirical distribution converging to a well-defined population quantity. This quantity often differs from the nominal target of inference.

An open question is whether one can develop estimators that are nearly as stable as $\hat{\mu}_{oLS}$ and $\hat{\mu}_{wLS}$ and yet show more robustness to violations of the model that is implicit in their construction. In the sequel, we develop estimators that show greater robustness to departures from the joint model for $\mathbf{X}$ and from the model for $Y|\mathbf{X}$. In certain circumstances, our estimators show greater stability than do those of Wang et al. (2006).

3 Consistency of JPS Estimators

The literature on RSS and JPS demonstrates the consistency of the estimators $\bar{Y}_{rss}$ and $\hat{\mu}_{jps1}$ in (2) and (3), respectively, under minimal conditions. These traditional estimators borrow heavily from the design-based perspective of survey sampling, where (approximate) unbiasedness is prized. Small variance is the secondary consideration. Modern work with surveys adjusts the balance, relying more heavily on models, especially where missing data is a concern (Lohr, 2010). With this perspective, a bit more bias is allowed, provided it is accompanied by a substantial reduction in variance. Simulations are used to evaluate the estimators' performance when the model does not hold. Wang et al. (2006) pursued this path.

We work in the infinite population setting where we collect IID sets, observing a single member of each set. As such, we envision that the data come from some distribution which we refer to as the "true model". In addition, there is a model used to construct the estimator. We assume that μ exists under both models. Consistency concerns arise when the true model and that used for analysis differ.

To set the framework for our consideration of robustness, we split the models into two parts. The first is the conditional distribution of $Y \mid \mathbf{R}$. The second is the distribution of $\mathbf{R}$ for the unit that is to be fully measured. The true and analysis models may differ in one or both of these aspects. A given estimator may be robust to differences in one portion of the model but not to differences in the other portion of the model. We consider each of the estimators in turn, presenting a heuristic argument for or against consistency. Our statements are to be taken loosely; simulations appearing in a later section support our claims. Formal statements and proofs of these results await another venue.

We briefly note that the estimators $\hat{\mu}_{jps1}$ and $\hat{\mu}_{jps2}$ are consistent for μ. These estimators do not rely on a model, and so we need not consider the gap between the true and analysis models. Consistency was established in MacEachern et al. (2004).

The estimators based on parametric models, $\hat{\mu}_{oLS}$ and $\hat{\mu}_{wLS}$, may or may not be consistent. We begin with $\hat{\mu}_{oLS}$. For a given stratum $\mathbf{r}$, an offset observation, $Y - \delta_{[\mathbf{r}]} = Y - \mu_{[\mathbf{r}]} + \mu$, has mean μ—provided the true and analysis models agree for the distribution of $Y|(\mathbf{R} = \mathbf{r})$ so that $\mu_{[\mathbf{r}]}$ has the same value under the two models and the offset has been correctly specified (or will be estimated consistently). Averaging across the strata, we see that the estimator targets the quantity $\mu - \sum_{\mathbf{r}} \pi_{\mathbf{r}} \delta_{[\mathbf{r}]}$. The estimator will be consistent for μ if (7) holds so that the average offset is zero. It is clear that this will be the case when the distribution on $\mathbf{R}$ and the conditional mean of $Y \mid (\mathbf{R} = \mathbf{r})$ are correctly specified for each of the H^2 strata. The first ensures accuracy of the $\pi_{\mathbf{r}}$, while the second ensures accuracy of the $\delta_{[\mathbf{r}]}$. Together, these imply (7). While these conditions stop short of full agreement between the true and analysis models, they are nearly there.

We might suspect that these conditions are essentially necessary for consistency for μ. However, the alternative description of the estimator lends insight. Suppose only that the conditional means of $Y \mid (\mathbf{R} = \mathbf{r})$ are correctly specified. Then the $\delta_{[\mathbf{r}]}$ are correct. Each debiased observation, $Y - \delta_{[\mathbf{r}]}$, has mean μ. The estimator is the simple average of the n debiased observations and so is consistent for μ. Accuracy of the $\pi_{\mathbf{r}}$ is not needed.

Interestingly, there is a third path to consistency. Suppose that the distribution on $\mathbf{R}$ is correctly specified, leading to a set of $\pi_{\mathbf{r}}$ that are the same under true and analysis models. Since these probabilities agree, and since, by the very definition of $\delta_{[\mathbf{r}]}$, $\sum_{\mathbf{r}} \pi_{\mathbf{r}} \delta_{[\mathbf{r}]} = 0$ under all models, (7) holds under the analysis model. It also holds under the true model. The debiasing for individual observations is inaccurate if the conditional means are incorrectly specified, but the inaccuracies cancel in the sum. In practice, for a finite sample size, the estimator would be conditionally biased, given the $n_{\mathbf{r}}$. However, for large samples, the $n_{\mathbf{r}}$ will be approximately proportional to the $\pi_{\mathbf{r}}$ and the bias will be small. In the limit, the bias disappears. Thus, we see that $\hat{\mu}_{oLS}$ is doubly robust, needing only one of the two portions of the model to hold to obtain consistency.

We next turn to $\hat{\mu}_{wLS}$. This estimator targets the quantity

$$\mu + \frac{\sum_{\mathbf{r}} \pi_{\mathbf{r}} \sigma_{\mathbf{r}}^{-2} \delta_{[\mathbf{r}]}}{\sum_{\mathbf{s}} \pi_{\mathbf{s}} \sigma_{\mathbf{s}}^{-2}}. \tag{8}$$

As with the ordinary least squares version of the estimator, we consider the debiased observations, $Y - \delta_{[\mathbf{r}]}$. The estimator is the precision weighted average of these debiased observations, each of which has mean μ, provided the $\delta_{[\mathbf{r}]}$ are accurate. This ensures consistency under the condition that the conditional means of $Y \mid (\mathbf{R} = \mathbf{r})$ are correctly specified. Under this condition, the estimator is unbiased for μ, and it has minimum variance in the class of weighted averages of the debiased observations if the conditional variances are also correctly specified.

Unfortunately, the argument for consistency when only the $\pi_{\mathbf{r}}$ are accurately specified does not go through for $\hat{\mu}_{wLS}$. The presence of the $\sigma_{\mathbf{r}}^{-2}$ in (8) impacts the weighting of the various rank classes. While (7) holds,

$$\sum_{\mathbf{r}} \pi_{\mathbf{r}} \sigma_{\mathbf{r}}^{-2} \delta_{[\mathbf{r}]} = 0 \tag{9}$$

does not. The estimator is not doubly robust.

The arguments for consistency of the various estimators lend insight into their performance. In all cases, we expect a better model to lead to more accurate estimation. Having the right family of models for $Y \mid (\mathbf{R} = \mathbf{r})$ allows us to create a consistent estimator for the conditional distribution of response given ranks. This leads to consistency for all of the estimators we have discussed. But it is difficult to capture this relationship correctly. The conditional distribution is most naturally driven by a latent model for $Y \mid \mathbf{X}$. Beginning with a model of this sort, an integration over the distribution of $\mathbf{X} \mid \mathbf{R}$ is needed to obtain that of $Y \mid \mathbf{R}$. The conditional distribution of $\mathbf{X} \mid \mathbf{R}$ relies on the joint distribution for the covariates $\mathbf{X}$. Having the right joint distribution for the covariates would also lead to the correct stratum probabilities $\pi_{\mathbf{r}}$. Thus, the joint distribution of the covariates deserves attention when creating a model to aid in estimation.

There is one situation where it is easy to get the stratum probabilities correct. This is when there is a single (univariate) set of ranks. In this case, the construction of the JPS leads immediately to the probability $\pi_r = 1/H$ for each of the H strata. In turn, this leads to consistency of $\hat{\mu}_{oLS}$ based on this single set of ranks for μ. We note that using the method in Wang et al. (2006), of debiasing the Y_i after passing to a univariate summary of the covariates differs from the common practice of mapping the covariate vector $\mathbf{X}$ into a fitted value, ranking on the fitted values, and then using an estimator of the form (3).

4 Covariates or Ranks?

The use of the vector of measured covariates, $\mathbf{X}_i$, to induce the ranks opens up many possibilities. One might ask whether ranking on X_1 and X_2 is optimal, or whether there is a mapping to another set of variates that leads to a better estimator. One possibility stands out, especially when relying on a multivariate normal model for $(Y, \mathbf{X})$. The vector $\mathbf{X}$ can be mapped to the regression of Y on $\mathbf{X}$ and its orthogonal complement. Under the multivariate normal model, this corresponds to an affine transformation of the covariates, $\mathbf{X}$, to a new set of covariates, say $\mathbf{W} = A\mathbf{X}$. The first coordinate of $\mathbf{W}$ is $E[Y \mid \mathbf{X}]$. The second coordinate is independent of both the first coordinate and the response and can be dropped.

In practice, we do not expect to know the relationship between covariates and response. With this in mind, we might estimate the relationship by fitting a model for $Y \mid \mathbf{X}$ to our n fully observed cases. Having done so, the fitted values become the first coordinate of $\mathbf{W}$. Often, the fitted values are estimates of $E[Y \mid \mathbf{W}] = E[Y \mid \mathbf{X}]$. From here, a natural estimate of μ can be obtained by averaging the fitted values (estimated means) for all nH observations. Following this path, the ranks have disappeared, and we are no longer in the setting of RSS or JPS.

The “covariate” approach leads to a natural estimator in the regression setting. The model for $Y \mid \mathbf{X}$ is a constant variance linear regression model. The chain of algebra below yields the estimator when the covariance matrix for $\mathbf{X}$ and Y is known.

Define $\bar{Y}_{srs}$ and $\bar{\mathbf{X}}_{srs}$ to be the mean of the response and the covariates for the n fully measured units, respectively. Take $\bar{\mathbf{X}}$ (a vector) to be the mean of the covariates for all nH units. For the covariance matrix, with Y in position 1 followed by the vector **X** in the trailing positions, the matrix can be written in partitioned form. This leads to Σ_{12} and Σ_{22} for the covariance of Y and the vector **X** and the variance matrix for the vector **X**, respectively. Then

$$\begin{aligned}\hat{\mu}_{reg} &= \frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\hat{E}[Y_{ih}|\mathbf{X}_{ih}]\\ &= \frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\hat{\mu}_{Y} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_{ih} - \hat{\mu}_{X})\\ &= \frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\bar{Y}_{srs} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_{ih} - \bar{\mathbf{X}}_{srs})\\ &= \bar{Y}_{srs} + \Sigma_{12}\Sigma_{22}^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{X}}_{srs}).\end{aligned} \tag{10}$$

This estimator is constructed by replacing the unknown parameters with estimates from the n fully measured units. In the event that the covariance matrix was not known, it would be replaced by the estimated covariance from the fully measured units. If the covariance matrix is unknown, estimates can be plugged in for the unknown quantities.

Why would one choose to pass from the covariate **X** to the coarser summary of its rank? The advantage of working with the rank-based estimators is their ability to handle deficiencies in the assumed model for (**X**, Y). A well-chosen estimator either will be consistent or will be Fisher consistent for a value very near the truth. (Parenthetically, estimators based directly on (**X**, Y) may also be consistent.) The rank-based estimators also seem to be better able to handle poorer quality covariates, including those whose distribution is not fully stable from one set to another. They also lead to methods with enhanced robustness for data sets with missing covariate values and imperfect models for the missing covariates given the observed covariates.

5 The Predictive Rank Distribution

Ranks lie at the heart of JPS, and indeed all of RSS. Focusing on a single set, we can describe the ranks of the H units in terms of a matrix P. Each row of the matrix corresponds to a unit in the set, each column to the rank of the unit in the set. A perfectly ranked sample corresponds to a permutation matrix where the row for unit h, if having rank j, is the H-vector with a 1 in position j and 0 in all other positions. We use the notation p_h to represent row h of the matrix P.

JPS and RSS rely on the rank matrix P but do not rely on an assumption of perfect ranking. Whether the ranks come from subjective judgement or from measured covariates, they yield a permutation matrix P, provided there are no ties in the ranking. In the event that there are ties, perhaps due to a pair of rankers (or measured covariates) providing different ranking matrices, P_1 and P_2, MacEachern et al. (2004) suggested use of the average $\bar{P} = 0.5P_1 + 0.5P_2$. This is appropriate when there is no reason to prefer one ranking over the other. Replacement of the permutation matrix P with the average necessitates replacement of the estimator (3) with one that allows non-indicator vectors p_h. Relying on the extensive body of work on ratio estimation in survey sampling, MacEachern et al. (2004) suggested the estimator in (4). This estimator effectively prorates the response across the strata to which it may belong.

The replacement of an $H \times H$ permutation matrix P with a convex combination over permutation matrices has been used productively in RSS by a number of authors, primarily when concerned with creating models for imperfect rankings (e.g. Bohn and Wolfe, 1994; Frey, 2007, while Dell and Clutter, 1972 and Fligner and MacEachern, 2006 developed models for imperfect ranking of differing form). The permutation matrices represent the extreme points of the set of doubly stochastic matrices—matrices with non-negative entries whose row sums and column sums total one. As a consequence, all other doubly stochastic matrices may be represented as an average of permutation matrices.

The use of measured covariates for JPS allows one to build a model for the response Y as a function of the measured covariates, $\mathbf{X}$. The model may be constructed from the data at hand, or it may have been developed in previous studies. With more than one covariate, a regression model for Y on $\mathbf{X}$ effectively transforms the vector of covariates into a single covariate while capturing much of the information connecting covariate to response. If the units in a set are ranked on the fitted value from the model when the covariate distribution is continuous, there will be no ties among the covariate values, ranking will be unambiguous, and the ranking matrix P will be a permutation matrix. Chen et al. (2005) took this approach to form a logistic regression model for a binary response.

A second approach to predictive modelling seeks to provide a full predictive distribution for Y given $\mathbf{X}$. There are a variety of ways to produce such a distribution, including Bayesian methods. Here, we consider a simple plug-in approach. Having specified a model for $Y \mid \mathbf{X}$, consider a set of H units. The predictive distribution of the rank of Y, given the observed $\mathbf{X}_h$, $h = 1, \ldots, H$, is computed. This predictive distribution yields the rows of the predictive rank matrix, P.

For the upcoming simulation study, we rely on a multivariate normal model to obtain the predictive rank distribution. When $H = 2$, calculation can be done in closed form. The predicted means for the cases in a set are $\mathbf{x}_h^\top \beta$ for $h = 1, 2$, and the predicted variances are $\sigma_y^2(1 - \rho^2)$ where σ_y^2 is the (marginal) variance of Y and ρ^2 is the coefficient of determination. This leads to the probability that unit 1 is ranked smallest:

$$P(Y_1 < Y_2 \mid \mathbf{x}_1, \mathbf{x}_2) = 1 - \Phi\left(\frac{\mathbf{x}_1^\top \beta - \mathbf{x}_2^\top \beta}{\sigma_y\sqrt{2(1-\rho^2)}}\right) = \Phi\left(\frac{\mathbf{x}_2^\top \beta - \mathbf{x}_1^\top \beta}{\sigma_y\sqrt{2(1-\rho^2)}}\right), \qquad (11)$$

where $\Phi(\cdot)$ represents the standard normal distribution function. A corresponding expression provides the probability that unit 1 is ranked largest. This leads to the equation

$$\mathbf{p}_1 = \left(\Phi\left(\frac{\mathbf{x}_2^\top \beta - \mathbf{x}_1^\top \beta}{\sigma_y \sqrt{2(1-\rho^2)}} \right), \Phi\left(\frac{\mathbf{x}_1^\top \beta - \mathbf{x}_2^\top \beta}{\sigma_y \sqrt{2(1-\rho^2)}} \right) \right) . \tag{12}$$

Similar calculations can be performed for unit 2 producing

$$\mathbf{p}_2 = \left(\Phi\left(\frac{\mathbf{x}_1^\top \beta - \mathbf{x}_2^\top \beta}{\sigma_y \sqrt{2(1-\rho^2)}} \right), \Phi\left(\frac{\mathbf{x}_2^\top \beta - \mathbf{x}_1^\top \beta}{\sigma_y \sqrt{2(1-\rho^2)}} \right) \right) . \tag{13}$$

When $H > 2$, the rank probabilities result from the integral of a multivariate normal distribution over a region defined by (hyper) planes. There is no closed-form expression for this integral, but simulation or numerical integration techniques allow us to approximate the integral. For our implementation, we only need the vector of rank probabilities for the fully measured unit in the set.

To approximate the rank probabilities, we use a simple technique, described for the case when unit 1 is sampled. We first generate Y_h, $h = 2, \ldots, H$, independently from normal distributions with means $\mathbf{x}_h \beta^\top$ and common variance $\sigma_y^2(1-\rho^2)$. These values are taken to be the responses for the $H-1$ unsampled units in the set. We next turn to the measured unit. We ignore the observed response, Y_1, and use the model to compute the H rank probabilities from the normal distribution with mean $\mathbf{x}_h \beta^\top$ and variance $\sigma_y^2(1-\rho^2)$ and cut-offs from the drawn values of $Y_2, \ldots, Y_H$. This gives us model-based rank probabilities for the measured unit, conditional on $Y_2, \ldots, Y_H$. We then repeat this process and average the results to provide $\hat{\mathbf{p}}_1$, a Monte Carlo approximation to the desired row of the permutation matrix, $\mathbf{p}_1$. For the upcoming simulations, we used a Monte Carlo sample size of 100 repetitions. A similar process is used if a different unit in the set is measured.

A formal description of our estimator requires a little more notation. Let $\mathbf{Y}$ denote the n-vector of measured responses. Let $\mathbf{1}_H$ denote the H-vector all of whose entries are 1 and $\mathbf{1}_n$ denote a similar vector of length n. Stack the row vectors for the n measured units' predictive rank probabilities in an $n \times H$ matrix, Q. Making use of our notation, the estimator (4) takes the form

$$\hat{\mu}_{jps2} = H^{-1} \left(\frac{\mathbf{Y}^\top Q}{\mathbf{1}_n^\top Q} \right)^\top \mathbf{1}_H \ , \tag{14}$$

where the ratio of H-vectors in parentheses is to be interpreted as elementwise division. For a given rank, h, the contribution to the estimator (4) or (14) is the ratio of two unbiased estimators—one for μ_h in the numerator and one for π_h in the denominator. While the π_h are known to equal $1/H$, the use of an empirical estimate of this quantity tends to improve the estimator, as it does in survey sampling Lohr (2010).

Furthering the parallel to survey sampling, we consider a “regression estimator” based on the measured Y_i and using the corresponding predictive rank probability vectors as covariates. This yields the estimator

$$\hat{\mu}_{jps3} = H^{-1}\mathbf{1}_H^\top(Q^\top Q)^{-1}Q^\top\mathbf{Y}\,. \tag{15}$$

The regression provides estimates of the means of the H rank classes which are then averaged to form the estimate of μ.

The combination of ranking on a single dimension and regression based on Q makes use of our knowledge of least squares regression. One of the basic properties of the least squares regression surface is that it passes through the “point of averages” given by the mean of the covariates (here, rank classes) and the mean of the response. This is true both for the population level regression relationship and regression based on a sample of data. Here, the population point of averages is known for the predictive rank distribution—from the construction of the JPS, it is simply H^{-1} for each of the H rank classes. The estimand is the mean of Y. The estimator comes from first estimating the “slope” of the regression surface and then adjusting from the sample point of averages for the predictive rank distribution to the population point of averages.

6 Simulation Study

This section presents the results of simulation studies comparing the performance of the various estimators of the mean based on a JPS sample. The findings for existing estimators are in line with the results in Wang et al. (2006). They also highlight the value that the predictive rank probabilities bring to estimation, particularly for the new estimator in (15).

The first study investigates the performance of eight estimators when the model that generates the data is fully known and is exactly right. This allows us to look at the potential performance of the estimators, exclusive of uncertainty about the model. Large sample sizes let us compare the asymptotic performance of the estimators.

The eight estimators are JPS1 from (3), a plug-in estimator based on the rank of $E[Y \mid X_1, X_2]$ (LS), OLS and WLS from Wang et al. (2006), TRs from (4), JPS2 and JPS3 from (14) and (15) and REG from (10). JPS2 and JPS3 make use of the predictive rank distribution. The estimator TRs has the same form as JPS2 but, as in MacEachern et al. (2004), uses the two ranks from the concomitants instead of the model-based predictive rank distribution. The REG estimator makes direct use of the covariates.

The model is the following. There are n sets, each consisting of H units. There are two covariates and a single response of interest. The covariates are measured on all nH units, while the response is measured for a single unit in each set. The vector (X_1, X_2, Y) follows a multivariate normal distribution with standard normal marginal distributions and covariances (correlations) specified in Tables 1, 2, and 3. The varied correlations range from a strong relationship between the concomitants and Y to a relatively weak relationship between them. Sample sizes $n = 20$, 50 and 100 are investigated for set size $H = 2$. For larger set sizes, results are presented only $n = 50$ and 100. For these set sizes, some of the estimators did not exist for some replicates. For the simulation, 10,000 replicates were used.

The tables present the relative accuracy of the various estimators to the sample mean based on a SRS. The entries are the ratio of MSEs for the SRS relative to the estimator in question. A number greater than 1 indicates smaller MSE for the estimator than for SRS.

We begin with a comparison of the new estimators JPS2 and JPS3. The overall pattern is that JPS3 is more accurate than JPS2, sometimes noticeably so. There are exceptions, particularly for smaller sample sizes and when the coefficient of determination is large (i.e. when the variance of $Y \mid (X_1, X_2)$ is small). There is some indication that, for a small sample size relative to set size, JPS3 may become numerically unstable, particularly when the coefficient of determination is large (this instability not visible in the tables).

Table 1 Simulated performance of estimators for set size H=2 relative to SRS. Entries are ratios of MSE

$(\rho_{1y}, \rho_{2y}, \rho_{12})$	(0.9,0.9,0.65)			(0.8,0.8,0.5)			(0.5,0.5,0.5)			(0.5,0.5,0.8)		
n	20	50	100	20	50	100	20	50	100	20	50	100
JPS1	1.28	1.35	1.32	1.21	1.24	1.22	1.04	1.08	1.08	1.09	1.15	1.14
LS	1.40	1.43	1.43	1.31	1.35	1.33	1.08	1.12	1.12	1.10	1.15	1.15
OLS	1.58	1.58	1.53	1.45	1.44	1.42	1.15	1.15	1.14	1.18	1.20	1.18
WLS	1.58	1.59	1.54	1.45	1.45	1.42	1.16	1.15	1.14	1.18	1.20	1.18
TRs	1.46	1.50	1.46	1.35	1.37	1.34	1.11	1.12	1.12	1.13	1.18	1.17
JPS2	1.49	1.51	1.50	1.43	1.45	1.42	1.09	1.08	1.07	1.11	1.14	1.11
JPS3	1.48	1.51	1.51	1.48	1.51	1.52	1.14	1.17	1.19	1.19	1.26	1.26
REG	1.99	1.96	1.95	1.77	1.73	1.71	1.22	1.20	1.21	1.27	1.30	1.28

Table 2 Simulated performance of estimators for set size H=3 relative to SRS. Entries are ratios of MSE

$(\rho_{1y}, \rho_{2y}, \rho_{12})$	(0.9,0.9,0.65)		(0.8,0.8,0.5)		(0.5,0.5,0.5)		(0.5,0.5,0.8)	
n	50	100	50	100	50	100	50	100
JPS1	1.56	1.62	1.38	1.40	1.08	1.12	1.20	1.22
LS	1.79	1.82	1.60	1.65	1.13	1.16	1.21	1.24
OLS	2.06	2.08	1.79	1.79	1.20	1.22	1.31	1.29
WLS	2.08	2.10	1.79	1.80	1.20	1.22	1.30	1.29
TRs	1.87	1.91	1.62	1.64	1.16	1.17	1.26	1.27
JPS2	1.93	1.97	1.78	1.78	1.10	1.10	1.19	1.19
JPS3	1.92	1.97	1.87	1.92	1.21	1.24	1.33	1.37
REG	2.83	2.84	2.28	2.25	1.27	1.28	1.41	1.42

Table 3 Simulated performance of estimators for set size H=4 relative to SRS. Entries are ratios of MSE

$(\rho_{1y}, \rho_{2y}, \rho_{12})$	(0.9,0.9,0.65)		(0.8,0.8,0.5)		(0.5,0.5,0.5)		(0.5,0.5,0.8)	
n	50	100	50	100	50	100	50	100
JPS1	1.72	1.82	1.47	1.53	1.08	1.15	1.20	1.26
LS	2.13	2.20	1.82	1.87	1.15	1.21	1.23	1.28
OLS	2.53	2.57	2.10	2.10	1.25	1.28	1.35	1.35
WLS	2.58	2.62	2.12	2.11	1.25	1.29	1.36	1.36
TRs	2.24	2.31	1.84	1.88	1.19	1.23	1.29	1.32
JPS2	2.38	2.42	2.06	2.10	1.10	1.14	1.22	1.21
JPS3	2.34	2.40	2.16	2.24	1.24	1.31	1.36	1.42
REG	3.75	3.80	2.77	2.77	1.32	1.36	1.48	1.48

With JPS3 generally outperforming JPS2, we turn to a comparison of JPS3 to OLS and WLS of Wang et al. (2006). For JPS3, we see a pattern of increasing efficiency relative to SRS as sample size increases. This comes from variation in the observed predictive rank distribution—for the measured units, the distribution is not uniform on the H rank classes. With a larger sample size, the distribution tends to be closer to uniform. This effect is larger for larger set sizes. For large sample size, JPS3 outperforms both OLS and WLS in all settings except the high correlation setting. We attribute this to the effective use of the predictive rank distribution in a context where the predictive rank distribution is fairly close to uniform.

The assumptions underlying the REG estimator are exactly right in this simulation. As we would expect, making full use of this model produces a very accurate estimator. In all cases covered by the simulation, the REG estimator has smaller MSE than any of the JPS style estimators.

Based on this simulation study, we make these recommendations: When one believes that a specific regression model is correct, use REG. Among the JPS style estimators, when n is small relative to H and one has confidence in the model upon which OLS and WLS are based, use either OLS or WLS; when n is large relative to H and the correlation is not extremely strong, use JPS3.

The second study investigates robustness of the estimators. In all cases, the model for $Y \mid (X_1, X_2)$ is incorrectly specified. Following Wang et al. (2006), we generate data from a multivariate normal model for (W_1, W_2, Y), and then compute $X_i = \exp(W_i)$ for $i = 1, 2$. The multivariate normal has mean vector $\mathbf{0}$ and correlations that match those in the first simulation study. The correlations between the X_i and between the X_i and Y are reported in Table 4. The correlations were obtained from a massive, 2 billion observation simulation from the joint distribution of (X_1, X_2, Y). Note that, in this simulation, the bivariate rank probabilities for (X_1, X_2) match those for (W_1, W_2). This portion of the model is correct.

Table 5 presents the results of our robustness simulation for set size $H = 2$. Focusing on the comparison of JPS3 to OLS and WLS, we find that JPS3 nearly

Table 4 Simulated correlations for lognormal/normal models for (X_1, X_2, Y)

$(\log X_1, \log X_2, Y)$			(X_1, X_2, Y)		
$(\rho_{1y}, \rho_{2y}, \rho_{12})$			$(\tilde{\rho}_{1y}, \tilde{\rho}_{2y}, \tilde{\rho}_{12})$		
0.9	0.9	0.65	0.6866	0.6866	0.5329
0.8	0.8	0.5	0.6103	0.6103	0.3776
0.5	0.5	0.5	0.3815	0.3815	0.3776
0.65	0.65	0.9	0.4959	0.4959	0.8495

Table 5 Simulated performance of the estimators for set size H=2 relative to SRS when model does not hold. Entries are ratios of MSE. The table includes simulated correlations for (X_1, X_2, Y) from Table 4

$(\tilde{\rho}_{1y}, \tilde{\rho}_{2y}, \tilde{\rho}_{12})$	(0.6866, 0.6866, 0.5329)					(0.6103, 0.6103, 0.3776)				
n	50	200	800	3200	12800	50	200	800	3200	12800
JPS1	1.31	1.36	1.38	1.36	1.32	1.26	1.26	1.29	1.23	1.25
LS	1.42	1.45	1.47	1.47	1.44	1.34	1.36	1.40	1.35	1.37
OLS	1.52	1.54	1.55	1.52	1.51	1.45	1.42	1.46	1.40	1.42
WLS	1.52	1.54	1.55	1.53	1.52	1.45	1.43	1.46	1.40	1.42
TRs	1.46	1.50	1.51	1.49	1.47	1.39	1.37	1.41	1.35	1.37
JPS2	1.21	1.24	1.24	1.24	1.22	1.19	1.18	1.20	1.15	1.17
JPS3	1.63	1.70	1.71	1.72	1.71	1.51	1.53	1.59	1.52	1.56
REG	1.44	1.46	1.45	1.47	1.43	1.42	1.37	1.42	1.35	1.37
$(\tilde{\rho}_{1y}, \tilde{\rho}_{2y}, \tilde{\rho}_{12})$	(0.3815, 0.3815, 0.3776)					(0.4959,0.4959,0.8495)				
n	50	200	800	3200	12800	50	200	800	3200	12800
JPS1	1.07	1.09	1.08	1.09	1.07	1.14	1.14	1.15	1.15	1.17
LS	1.10	1.12	1.11	1.12	1.11	1.15	1.14	1.15	1.16	1.17
OLS	1.13	1.14	1.13	1.13	1.12	1.18	1.16	1.17	1.17	1.18
WLS	1.13	1.14	1.13	1.14	1.12	1.19	1.16	1.17	1.17	1.18
TRs	1.11	1.12	1.12	1.12	1.11	1.17	1.16	1.17	1.17	1.18
JPS2	1.03	1.03	1.02	1.03	1.02	1.03	1.00	1.03	1.02	1.04
JPS3	1.11	1.15	1.14	1.15	1.15	1.17	1.19	1.20	1.20	1.22
REG	1.11	1.13	1.12	1.12	1.13	1.16	1.14	1.15	1.13	1.16

always outperforms OLS and WLS. The only instance where OLS and WLS do better is for $n = 50$ with weak correlations among (Y, X_1, X_2) where OLS and WLS are slightly better. REG performs poorly relative to JPS3, OLS and WLS.

Table 6 presents the MSE of a SRS relative to the various estimators for set size $H = 4$. We note that, for this larger set size, OLS and WLS perform relatively better than for $H = 2$. JPS3 performs better for strong correlations and larger sample sizes, while OLS and WLS outperform for weak correlations and smaller sample sizes. REG follows the pattern of OLS and WLS but does not perform as well as these estimators.

Table 6 Simulated performance of the estimators for set size H=4 relative to SRS when model does not hold. Entries are ratios of MSE. The table includes simulated correlations for (X_1, X_2, Y) from Table 4

$(\tilde{\rho}_{1y}, \tilde{\rho}_{2y}, \tilde{\rho}_{12})$	(0.6103, 0.6103, 0.3776)			(0.6103, 0.6103, 0.3776)		
n	50	100	200	50	100	200
JPS1	1.76	1.79	1.85	1.48	1.50	1.57
LS	2.11	2.17	2.23	1.78	1.83	1.89
OLS	2.49	2.42	2.48	2.05	2.01	2.05
WLS	2.53	2.47	2.53	2.07	2.03	2.06
TRs	2.30	2.28	2.36	1.87	1.85	1.90
JPS2	1.51	1.47	1.51	1.36	1.31	1.34
JPS3	2.38	2.60	2.72	1.98	2.04	2.15
REG	1.85	1.83	1.91	1.69	1.61	1.66
$(\tilde{\rho}_{1y}, \tilde{\rho}_{2y}, \tilde{\rho}_{12})$	(0.3815, 0.3815, 0.3776)			(0.4959,0.4959,0.8495)		
n	50	100	200	50	100	200
JPS1	1.08	1.14	1.15	1.25	1.27	1.30
LS	1.13	1.21	1.21	1.26	1.29	1.31
OLS	1.23	1.27	1.25	1.37	1.34	1.35
WLS	1.23	1.27	1.25	1.37	1.35	1.35
TRs	1.18	1.23	1.22	1.33	1.33	1.35
JPS2	1.03	1.06	1.05	1.08	1.06	1.07
JPS3	1.15	1.22	1.23	1.30	1.32	1.35
REG	1.17	1.21	1.19	1.28	1.25	1.25

7 Discussion

For us, one of the most intriguing aspects of our exploration of the estimators developed by Stokes and colleagues in Wang et al. (2006) is the double robustness of their OLS estimator. Their clever use of debiasing followed by a simple average (or, for their WLS estimator, a weighted average) stabilizes the weights of individual cases in the estimator. This stabilization is especially important for smaller sample sizes and when there are multiple covariates, where large differences in the weights are common. The combination of stability of the estimator (small variance) with minimal bias in a fashion that is robust to violations of the presumed model for $Y \mid \mathbf{X}$ has proven to be extremely effective. We suspect that this robustness is an important factor in the practical success of the method.

The estimators that we develop in this work pursue the path laid out by Stokes and colleagues. Our new estimators focus on stability while maintaining small bias. To control the bias, we resort to a form of dimension reduction, passing from multiple covariates to a one-dimensional ranking. To enhance stability, we pass from a single observed rank to the predictive rank distribution of the measured unit. Together, these adjustments produce estimators that can be more accurate than previously developed estimators for RSS and JPS data. In some circumstances, the gains are striking.

The new estimators make use of the covariate values for all units in a set to create the predictive rank distribution. In contrast, the estimators of Wang et al. (2006) make less use the observed covariate values. Their estimators use the covariate to create the ranks within a set and, implicitly, to estimate the covariance matrix for covariates and response, leading to the offsets $\delta_{[\mathbf{r}]}$. This can be accomplished with information from only the fully measured units without reference to the unmeasured units in a set. One can imagine that, in some circumstances, one could collect sets that are ranked on covariates and yet observe numerical covariate values only for the fully measured units. In such a setting, the Wang et al. (2006) estimators could be used, while the new estimators could not be computed. We believe that these situations would be relatively rare.

The success of all of these estimators leads us to alter our view on RSS and JPS estimation. Most of the literature on these methods takes one of two forms. It either makes very minimal assumptions about the mechanism that gives rise to the data and is essentially nonparametric in nature, or it makes very strong assumptions about this mechanism. The latter approach has generated papers that make heavy use of strong parametric assumptions and that presume that rankings are perfect. In nearly all cases, the explicit goal is to find the rank of a unit within a set. In contrast, along the lines of MacEachern et al. (2004), we find that there is value in allowing for a distribution over ranks. With the availability of covariates, this suggests that we should devote considerable effort to building covariate-driven models for the rank of the measured unit. Accepting the uncertainty that comes with such models improves estimation when compared to effectively selecting a rank at random from the predictive rank distribution and using this to create the estimator. We believe that shifting the perspective from the creation of estimators to building sound models for the data will, in the end, result in better estimators.

Our development of novel estimators suggests specific directions for further research. One is to sharpen the heuristic arguments for consistency and double robustness of the estimators in Wang et al. (2006) and to formally establish these results. A second is to combine the debiasing that is implicit in the estimators of Wang et al. (2006) with our predictive rank techniques. A third is to more completely explore the impact of developing and fitting a model to the data to which it will be applied. The use of split-sample techniques such as the jackknife (dropping a set at a time) or half-sample methods may help control the bias that arises from multiple uses of the same data.

There are many ways to obtain a predictive rank distribution. We have used a simple calculation under the assumption that the model is fully known. When fitting a class of models to data, we might turn to a plug-in estimator. Such estimators are commonly derived from maximum likelihood, maximum penalized likelihood or generalized estimating equations. Alternatively, we could use Bayesian methods to account for uncertainty in the fit of a fitted model and also to account for uncertainty in which model should be fit (c.f. Meeden and Lee, 2014).

In addition to the relatively simple setting considered here, we note that similar techniques can be developed for many other data structures. Following Stokes and Sager (1988), we could look at $P(Y \in A)$ for some set A, including the multivariate setting. We could look at a multivariate mean, or a measure of dependence between variables. We could look at survival data where censoring is a concern. We could look at data that lie in non-Euclidean spaces, and so on. When the following features are present, we have a clear route on which to proceed: A measured covariate that can be ranked (to yield a concomitant) to play the role of $\mathbf{X}$, a target phrased as an expectation to play the role of $\mu = E[Y]$ and a RSS or JPS to produce the data. Stokes and her colleagues have paved the route.

Acknowledgments The authors thank Sherry Wang for providing code used to fit the estimators in Wang, Stokes, Lim and Chen. This work was supported in part by NSF grant numbers SES-1921523 and DMS-2015552.

References

Bohn, L. L., & Wolfe, D. A. (1994). The effect of imperfect judgment rankings on properties of procedures based on the ranked-set samples analog of the Mann-Whitney-Wilcoxon statistic. *Journal of the American Statistical Association, 89*, 168–176.

Chen, H., Stasny, E. A., & Wolfe, D. A. (2005). Ranked set sampling for efficient estimation of a population proportion. *Statistics in Medicine, 24*, 3319–3329.

Dell, T. R., & Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics, 28*, 545.

Fligner, M. A., & MacEachern, S. N. (2006). Nonparametric two-sample methods for ranked-set sample data. *Journal of the American Statistical Association, 101*, 1107–1118.

Frey, J. C. (2007). New imperfect rankings models for ranked set sampling. *Journal of Statistical Planning and Inference, 137*, 1433–1445.

Frey, J. (2016). A more efficient mean estimator for judgement post-stratification. *Journal of Statistical Computation and Simulation, 86*, 1404–1414.

Frey, J., & Feeman, T. G. (2012). An improved mean estimator for judgment post-stratification. *Computational Statistics & Data Analysis, 56*, 418–426.

Huber, P. J. (1981). *Robust statistics*. Wiley.

Lohr, S. L. (2010), *Sampling: Design and analysis* (2nd ed.). Brooks Cole.

MacEachern, S. N., Stasny, E. A., & Wolfe, D. A. (2004). Judgement post-stratification with imprecise rankings. *Biometrics, 60*(1), 207–215.

McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets. *Crop & Pasture Science, 3*, 385–390.

McIntyre, G. A. (2005). A method for unbiased selective sampling, using ranked sets. *The American Statistician, 59*, 230–232.

Meeden, G., & Lee, B. (2014). More efficient inferences using ranking information obtained from judgment sampling. *Journal of Survey Statistics and Methodology, 2*, 38–57.

Stokes, S. L. (1977). Ranked set sampling with concomitant variables. *Communications in Statistics-Theory and Methods, 6*, 1207–1211.

Stokes, S. L., & Sager, T. (1988). Characterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association, 83*, 374–381.

Wang, X., Stokes, L., Lim, J., & Chen, M. (2006). Concomitants of multivariate order statistics with application to judgment poststratification. *Journal of the American Statistical Association, 101*, 1693–1704.

Wolfe, D. A. (2012). Ranked set sampling: Its relevance and impact on statistical inference. *International Scholarly Research Notices, 2012*, 1–32.

Judgment Post-stratified Sampling with Multiple Ranking: A Comparison with Ranked Set Sampling

Omer Ozturk, Jennifer Brown, and Olena Kravchuk

Abstract Ranked set sampling and judgment post-stratified sampling designs form groups among sample units using their relative positions (ranks) in small comparison sets. This rank information governs the decision on whether to include units in a final ranked set sample (RSS), but only supplements the primary selection of units in a judgment post-stratifed sample (JPS). If the position information in the comparison sets is accurate, for both designs, the samples represent the population better than a simple random sample (SRS) of the same size. The RSS design uses the ranking information in a more direct way. However, the RSS design induces a strong structure in a sample, and the data so collected may not be suitable for studies where a multipurpose analysis is desired. The JPS design is slightly less efficient, but more flexible and enables multipurpose analyses. This paper explores the benefits of the JPS over the RSS design of the same sample size. We show that the efficiency loss in the JPS design can be reduced by using ranks from multiple comparison sets. The paper presents results from an extensive simulation study to demonstrate the benefit of the JPS design over the SRS and RSS designs when the JPS is constructed using multiple ranking methods.

O. Ozturk (✉)
The Ohio State University, Department of Statistics, Columbus, OH, USA
e-mail: ozturk.4@osu.edu

J. Brown
University of Canterbury, School of Mathematics and Statistics, Christchurch, New Zealand
e-mail: jennifer.brown@canterbury.ac.nz

O. Kravchuk
University of Adelaide, School of Agriculture, Food and Wine, Adelaide, SA, Australia
e-mail: olena.kravchuk@adelaide.edu.au

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_2

1 Introduction

In field sampling and social science research, creating samples that are representative of the population is important. This can be achieved by using stratified sampling, cluster sampling, or post-stratified sampling designs. In certain cases, the stratification variable may be subjective, rough, and imprecise, but can still provide valuable information about the relative position of a sample unit in a small set. Such stratification variable can be used to reduce the sampling variation, and cost in ranked set and judgment post-stratified sampling designs. These designs stratify the sample into groups of homogeneous observations using sample units' relative positions (ranks) in small comparison sets.

For a ranked set sample (RSS) of size n, one first determines a set size H and then selects nH units at random from the population. These units are divided into n comparison sets, each of size H. Units in the comparison sets are then ranked from the smallest to the largest, without measurement. Ranking can be performed on either the variable of interest assessed on a less elaborate scale or an auxiliary variable. The unit judged to be the h-th smallest $(Y_{[h]j})$ is measured in n_h comparison sets for $j = 1, \ldots, n_h$, $\sum_{h=1}^{H} n_h = n$. The measured observations $Y_{[h]j}$, $j = 1, \ldots, n_h$; $h = 1, \ldots, H$ are called a ranked set sample. If $n_h = d$ for all $h = 1, \ldots, H$ so that $n = dH$, the RSS is called balanced, and d is called the cycle size. If there is no ranking error, the square brackets are replaced with round parentheses, and the $Y_{(h)j}$ becomes the h-th order statistic in a sample of size H.

Ranked set sampling design was introduced by McIntyre (1952, 2005). The main motivation in McIntyre's work was to enable field researchers to conduct pasture yield (and similar) field assessments in an objective and efficient way. Takahasi and Wakimoto (1968) developed the theoretical foundation of the ranked set sampling design and showed that the RSS mean is always better than a sample mean of a simple random sample (SRS). Dell and Clutter (1972) showed that even with some ranking errors, the RSS mean is as good as, or better than, the SRS mean depending on the quality of ranking information. Research activities in RSS designs then expanded in different directions, including parametric and nonparametric settings. In the parametric setting, a few representative publications are Stokes (1995), Chen and Bai (2000), Arslan and Ozturk (2013), Hatefi et al. (2014), and Hatefi et al. (2015). In the nonparametric setting, readers are referred to Bohn and Wolfe (1992, 1994), Hettmansperger (1995), Koti and Babu (1996), Ozturk (1999), and Fligner and MacEachern (2006). Two books have been published on ranked set sampling design, Chen et al. (2003), and Bouza and Al-Omari (2019). A comprehensive list of references can be found in these publications.

The RSS research activities also considered the finite population setting. Patil et al. (1995) constructed an RSS using sampling without replacement selection procedure. Deshpande et al. (2006) expanded the RSS design to three different schemes of sampling without replacement. Frey (2011), Ozturk and Jafari Jozani (2014), and Jafari Jozani and Johnson (2011) used probability sampling and constructed Horvitz-Thompson-type estimators. Ozturk and Bayramoglu Kavlak (2018) constructed inference using a superpopulation model in ranked set sampling.

MacEachern et al. (2004) introduced the judgment post-stratification design to provide the flexibility for a multipurpose analysis of sample data. For a judgment post-stratified sample (JPS), one first selects and measures an SRS of size n, Y_i, $i = 1, \ldots, n$. For each measured unit Y_i, one then selects additional $H - 1$ units from the population, without directly measuring them, to form a comparison set of size H. The units in the comparison set are ranked from the smallest to the largest, and the rank of Y_i, R_i is recorded. The pairs (Y_i, R_i), $i = 1, \ldots, n$, constitute a JPS.

In recent years, the JPS design in an infinite population setting has generated extensive research interest. Ozturk (2014) considered the estimation of the population quantile and variance from a JPS. Wang et al. (2006) used the concomitant order statistics to estimate the population mean. Frey and Feeman (2012, 2013) constructed estimators for the population mean and variance by conditioning on the judgment group sample sizes. These new estimators improve the unconditional JPS estimators. Chen et al. (2014), Frey and Ozturk (2011), Wang et al. (2012), Wang et al. (2008), and Stokes et al. (2007) constructed constrained estimators using stochastic ordering among judgment ranking groups. The main idea in the constraint estimators is to minimize the impact of ranking error by forcing judgment class means to follow the stochastic order among ranking groups. In a different direction, Ozturk (2017) constructed conditional ranks in smaller comparison sets of size $K < H$ given the original ranks in a larger comparison set of size H. The impact of any ranking error on the estimator in this case was relatively small, and less than for the estimator based on the large comparison set of size H. Ozturk (2013) and Ozturk and Demirel (2016) used a multi-ranking approach to reduce the impact of ranking error in judgment post-stratified and ranked set samples.

In the finite population settings, Ozturk (2016a, 2016b, 2019) constructed estimators for the population mean and total for the JPS design. A JPS can be constructed by sampling with or without replacement. It is shown that the variance estimator of the sample mean requires a finite population correction factor when sampling without replacement. Ozturk and Bayramoglu Kavlak (2018, 2019, 2020) developed inference to predict the population mean and total using a superpopulation model.

In the JPS design, the ranks are constructed post-experimentally after an SRS is chosen. Hence, it is possible to have more than one rank for each measured unit in the SRS by permuting the $n(H - 1)$ unmeasured units used in the construction of comparison sets in the first created JPS. Each permutation creates n comparison sets, each of size H, containing the measured unit. The units in the sets are ranked again, without measurement, and the ranks of the measured units in the comparison sets are determined. This permutation procedure can be done many times and each permutation creates a new set of ranks for the same measured values. Ranks from different permutations are conditionally independent given the original SRS. One may then combine all these ranks using the Rao-Blackwell theorem by conditioning on the original SRS.

A similar idea can be used in the RSS design, but the extension to multiple ranks is not as trivial as in the JPS design. In the RSS design, the measured observations, $Y_{[h]j}$, are not identically distributed. Hence, the units in the comparison set constructed after the permutation of $n(H-1)$ units are not iid since each comparison set contains one of the $y_{[h]j}$ from the original ranked set sample and this will have a different distribution from the other units in the set. Even though the comparison sets will be different after each permutation, the rank of $y_{[h]j}$ will depend on the original rank h. Hence, the idea of multiple ranks in the judgment post-stratified sampling may not be easily extended to ranked set sampling.

There are a few other differences between the RSS and JPS designs. One of the major differences is whether the ranking is done before or after the units are measured for the variable of interest. In RSS, the ranking is performed before one measures the units, and the ranks guide the measurement decision. The rank and the measurement of a unit cannot be separated. Hence, an RSS cannot be reduced to an SRS, unless it is unusual situation where the ranking variable is not correlated with the measurement variable. In a JPS, ranking is performed after one measures the units in the SRS. The ranks are not the essential part of the measured units; they are the ranks of the variable of interest measured on a quicker scale (e.g., visual inspection) after the construction of an SRS. Since the auxiliary (ranking) variable is only post-associated with the response measurements, it can be ignored and a JPS can be reduced to an SRS if desired.

Another major difference is the distributional properties of the ranks. The ranks in RSS are pre-determined nonrandom constants. Hence, the ranking group sizes n_h, $h = 1, \ldots, H$, are nonrandom integers. In a JPS, the rank R_i is a discrete uniform random variable with the support on integers $1, \ldots, H$. Hence, the judgment group sample size vector $(n_1, \ldots, n_H)$ has a multinomial distribution with the sample size n and the success probability vector $(1/H, \ldots, 1/H)$.

One may look at the RSS and JPS designs in terms of the trade-off between the efficiency gain of the RSS and the adaptability of JPS for multipurpose studies. To our knowledge, this trade-off has not yet been posed and investigated. In this paper, we provide a comprehensive study to compare the RSS and JPS designs for their efficiencies and multiple ranking properties. In Sect. 2, we provide a detailed description of multi-ranking in RSS and JPS designs. In Sect. 3, we review the distributional properties of the RSS and JPS means. In Sect. 4, we present empirical results to compare the RSS and JPS designs. In Sect. 5, we illustrate the use of RSS and JPS designs with an agricultural application example. Section 6 provides concluding remarks.

2 Sampling Designs with Multiple Ranking Methods

We consider a finite population of size N. The population values of the variable Y are denoted as $y_1, \ldots, y_n$. The mean and variance of the population are given by

$$\bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad S_N^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2.$$

From this finite population, we construct RSS and JPS with multiple ranks. The samples are constructed using the sampling *with* and *without* replacement selection procedures. Unless stated otherwise, we always consider a finite population setting in this paper.

RSS with Multiple Ranks We first consider an RSS selected using the sampling with replacement (SWR) selection procedure. For cycle j and rank h, we construct a comparison set of size H using a sampling without replacement (SWOR) scheme. The units in the comparison set are ranked by the best ranking method available. The unit judged to be the h-th smallest, $Y_{[h]j}$, is measured. For the observation $Y_{[h]j}$, additional $K-1$ ranks can be constructed in two ways. If there are $K-1$ ($K>1$) rankers or ranking variables available, the rank of $Y_{[h]j}$, $R_{k|j,h}$, among the units in the comparison set is determined for each method k, $k=2,\ldots,K$. After these ranks are determined, all units in the comparison set are returned to the population before constructing the next comparison set. Hence, the same unit may appear in the final sample more than once, and all the observations are independent. We note that units within a comparison set are selected using the SWOR procedure to minimize the ranking error. The ranks using the first ranking method ($k=1$) are predetermined (nonrandom constants) to have a balanced ranked set sample, $n_h=d$, for $h=1,\ldots,H$. The remaining $K-1$ ranks are random and may not necessarily be balanced.

Even if there is only one ranker or one auxiliary variable to rank the units, we can still construct an RSS with multiple ranks. For given values of h and j, $Y_{[h]j}$ is measured in a comparison set. Next, we form $K-1$ different comparison sets by selecting $H-1$ additional units at random from the population without measurement, $V_{k|h,j}=\left\{Y_{[h]j},Y_{k,1},\ldots,Y_{j,H-1}\right\}$, $k=2,\ldots,K$, and determine the rank of $Y_{[h]j}$, $R_{k|j,h}$, in each set for $k=2,\ldots,K$. The RSS with multiple ranks can be written as

$$\left\{(Y_{[h]j},R_{k|j,h}),h=1,\ldots,H,j=1,\ldots,d;k=1,\ldots,K\right\},$$

where $R_{k|j,h}$ is the conditional rank assigned by ranking method k given that the observation $Y_{[h]j}$ is assigned rank h. We note that $P(R_{1|j,h}=h)=1$. The ranks assigned by another ranking method are random variables, but their distributions depend on the ranks assigned by the first (best) ranking method.

An RSS with multiple ranks using a SWOR selection scheme can be constructed in a similar fashion. The only difference is that after determining the rank of $Y_{[h]j}$, all H units in the comparison set are removed from the population before constructing the next comparison set. Hence, for each ranking method, all comparison sets are disjoint.

The final sample cannot have repeated observations and the observations are not independent. If the population size N is large with respect to sample size n, ranked set samples constructed using SWR or SWOR selection procedures become approximately equivalent.

Table 1 Illustration of RSS multi-ranker sampling with set size $H = 3$, cycle size $d = 2$, and the number of ranking methods $K = 3$

Cycle (j)	h	Balanced RSS	Ranks from K methods	Ranked set sample
1	1	$\{\boldsymbol{Y_{[1]1}}, Y_{[2]1}, Y_{[3]1}\}$	$\{1, R_{2\|1,1}, R_{3\|1,1}\}$	$\{Y_{[1]1}, 1, R_{2\|1,1}, R_{3\|1,1}\}$
1	2	$\{Y_{[1]1}, \boldsymbol{Y_{[2]1}}, Y_{[3]1}\}$	$\{2, R_{2\|1,2}, R_{3\|1,2}\}$	$\{Y_{[2]1}, 2, R_{2\|1,2}, R_{3\|1,2}\}$
1	3	$\{Y_{[1]1}, Y_{[2]1}, \boldsymbol{Y_{[3]1}}\}$	$\{3, R_{2\|1,3}, R_{3\|1,3}\}$	$\{Y_{[3]1}, 3, R_{2\|1,3}, R_{3\|1,3}\}$
2	1	$\{\boldsymbol{Y_{[1]2}}, Y_{[2]2}, Y_{[3]2}\}$	$\{1, R_{2\|2,1}, R_{3\|2,1}\}$	$\{Y_{[1]2}, 1, R_{2\|2,1}, R_{3\|2,1}\}$
2	2	$\{Y_{[1]2}, \boldsymbol{Y_{[2]2}}, Y_{[3]2}\}$	$\{2, R_{2\|2,2}, R_{3\|2,2}\}$	$\{Y_{[2]2}, 2, R_{2\|2,2}, R_{3\|2,2}\}$
2	3	$\{Y_{[1]2}, Y_{[2]2}, \boldsymbol{Y_{[3]2}}\}$	$\{3, R_{2\|2,3}, R_{3\|2,3}\}$	$\{Y_{[3]2}, 3, R_{2\|2,3}, R_{3\|2,3}\}$

The construction of ranked set samples using multiple ranking methods is illustrated in Table 1. In this table, the third column presents the comparison sets in which a balanced ranked set sample is constructed with the first ranking method. It highlights that the units are ranked using ranking method 1; the bold-faced values are measured. The fourth column lists the ranks obtained from all K ($K = 3$) different ranking methods for the bold-faced values in column 3. The last column gives the ranked set sample of size 6. In this example, each entry has three ranks generated by three ranking methods.

JPS with Multiple Ranks We first construct a simple random sample of size n using the SWR selection procedure and measure all n units, $Y_1, \ldots, Y_n$. For each Y_i, we then select additional $H - 1$ units under SWOR selection from the population to form a comparison set $V_i = \{Y_i, Y_1, \ldots, Y_{H-1}\}$. We rank these units from smallest to largest without measuring Y, using K different ranking methods, and identify the rank of Y_i, $R_{k|i}$, for each ranking method k, $k = 1, \ldots, K$, where $R_{k|i}$ is the rank of Y_i assigned by ranking method k. All units in the comparison set, including the one we measured, are returned to the population before the construction of the next comparison set. Hence, a JPS may have repeated observations and all Y_i, $i = 1, \ldots, n$, are independent. This process creates the sample

$$\{Y_i, R_{k|i}\};\ i = 1, \ldots, n, k = 1, \ldots, K.$$

If only one ranking method is available, for each Y_i, one can create K different comparison sets, $V_{k|i} = \{Y_i, Y_{1,k}, \ldots, Y_{H-1,k}\}$ for $k = 1, \ldots, K$, where $Y_{h,k} \neq Y_i$ is the additional unit selected from the population to construct the k-th comparison set. These sets are ranked using the ranking method and the ranks of Y_i, $R_{k|i}$, are determined in $V_{k|i}$, for $k = 1, \ldots, K$.

A JPS under the SWOR selection procedure is constructed in a similar fashion. The only difference here is that all comparison sets for each ranking method are disjoint, and hence, the JPS cannot have repeated observations. For small population sizes N, observations Y_i, $i = 1, \ldots, n$, in the sample are negatively correlated since sample units are selected as an SRS without replacement.

The construction of a JPS with multiple ranking methods and under the SWOR selection scheme is illustrated in Table 2. In this example, the sample and set

Table 2 Illustration of multi-ranker JPS under sampling without replacement selection with set size $H = 3$, sample size $n = 6$, and the number of ranking methods $K = 3$

(j)	SRS	Comparison sets	JPS
1	Y_1	$\{\boldsymbol{Y_1}, Y_{7,1}, Y_{8,1}\}, \{\boldsymbol{Y_1}, Y_{9,2}, Y_{19,2}\}, \{\boldsymbol{Y_1}, Y_{26,3}, Y_{12,3}\}$	$\{Y_1, R_{1\|1}, R_{2\|1}, R_{3\|1}\}$
2	Y_2	$\{\boldsymbol{Y_2}, Y_{9,1}, Y_{10,1}\}, \{\boldsymbol{Y_2}, Y_{20,2}, Y_{8,2}\}, \{\boldsymbol{Y_1}, Y_{17,3}, Y_{27,3}\}$	$\{Y_2, R_{1\|2}, R_{2\|2}, R_{3\|2}\}$
3	Y_3	$\{\boldsymbol{Y_3}, Y_{11,1}, Y_{12,1}\}, \{\boldsymbol{Y_3}, Y_{21,2}, Y_{22,2}\}, \{\boldsymbol{Y_3}, Y_{9,3}, Y_{10,3}\}$	$\{Y_3, R_{1\|3}, R_{2\|3}, R_{3\|3}\}$
4	Y_4	$\{\boldsymbol{Y_4}, Y_{13,1}, Y_{14,1}\}, \{\boldsymbol{Y_4}, Y_{15,2}, Y_{23,2}\}, \{\boldsymbol{Y_4}, Y_{14,3}, Y_{28,3}\}$	$\{Y_4, R_{1\|4}, R_{2\|4}, R_{3\|4}\}$
5	Y_5	$\{\boldsymbol{Y_5}, Y_{15,1}, Y_{16,1}\}, \{\boldsymbol{Y_5}, Y_{18,2}, Y_{24,2}\}, \{\boldsymbol{Y_5}, Y_{29,2}, Y_{13,2}\}$	$\{Y_5, R_{1\|5}, R_{2\|5}, R_{3\|5}\}$
6	Y_6	$\{\boldsymbol{Y_6}, Y_{17,1}, Y_{18,1}\}, \{\boldsymbol{Y_6}, Y_{12,2}, Y_{25,2}\}, \{\boldsymbol{Y_6}, Y_{8,3}, Y_{30,3}\}$	$\{Y_6, R_{1\|6}, R_{2\|6}, R_{3\|6}\}$

sizes are 6 and 3, respectively. For each measured unit, three ranks are constructed ($K = 3$). The second column presents a simple random sample of size $n = 6$. The third column presents three comparison sets, $V_{k|i}$, for each Y_i, one for each ranking method. The fourth column presents the JPS with three ranks. The comparison sets of each ranking method in Table 2, sets in block 1, 2, or 3 in column 3, are disjoint and cannot have repeated observations. Comparison sets for the different ranking methods (sets in different blocks) are not necessarily disjoint because the same ranking unit can appear in more than one set in different ranking methods. Sampling is without replacement and thus the comparison sets in different rows for the same ranking method are disjoint. We note that the sample units will not be independent if the population size N is small in relation to the sample size n.

3 Statistical Inference Using RSS and JPS

In this section, we provide a brief overview of statistical inference using the RSS and JPS designs. We first assume $K = 1$. The estimators for the population mean are given as the sample mean of the RSS and JPS:

$$\bar{Y}_{RSS} = \frac{1}{dH}\sum_{h=1}^{H}\sum_{j=1}^{d} Y_{[h]j}, \quad \bar{Y}_{JPS} = \frac{1}{d_n}\sum_{h=1}^{H} J_h I_h \sum_{j=1}^{n} Y_j I(R_j = h),$$

where $I(a)$ is 1 if a is true, $I_h = I(n_h > 0)$, $d_n = \sum_{h=1}^{H} I_h$, and $J_h = 1/n_h$ if $n_h > 0$ and zero otherwise. Both of these estimators are unbiased for the population mean $\bar{y}_N$ regardless of the ranking quality as long as a consistent ranking method is used. If all units in the comparison sets are ranked with the same ranking methods, the ranking procedure is called consistent. The following theorem provides variances of the sample means under SWR and SWOR selection schemes using a consistent ranking method.

Theorem 1 *Let $Y_{[h]j}$, $h = 1, \cdots, H$, $j = 1, \ldots, d$ and (Y_j, R_j), $j = 1, \ldots, n$ be RSS and JPS constructed using a consistent ranking methods, respectively.*

(i) If the samples are constructed with replacement, the variances of $\bar{Y}_{RSS}$ and $\bar{Y}_{JPS}$ are given by

$$\sigma^2_{RSS} = \frac{1}{dH^2} \sum_{h=1}^{H} S^2_{[h]} \quad \sigma^2_{JPS} = \frac{H}{H-1} Var\left(\frac{I_1}{d_n}\right) \sum_{h=1}^{H} (\bar{y}_{[h]} - \bar{y}_N)^2 + E\left(\frac{I_1^2}{d_n^2 n_1}\right) \sum_{h=1}^{H} S^2_{[h]},$$

where $S^2_{[h]} = Var(Y_{[h]1})$, $\bar{y}_{[h]} = E(Y_{[h]1})$, $Var(I_1/d_n) = \frac{1}{H^2} \sum_{k=1}^{H-1} (\frac{k}{H})^{n-1}$ and

$$E\left(\frac{I_1^2}{n_1 d_n^2}\right) = \frac{1}{H^n} \left(\frac{1}{n} + \sum_{k=2}^{H} \sum_{j=1}^{k-1} \sum_{t=1}^{n-k+1} \frac{(-1)^{j-1}}{k^2 t} \binom{H-1}{k-1} \binom{k-1}{j-1} \binom{n}{t} (k-j)^{n-t} \right).$$

(ii) If the samples are constructed without replacement, the variances of $\bar{Y}_{RSS}$ and $\bar{Y}_{JPS}$ are given by

$$\sigma^2_{RSS} = \frac{N-1-n}{n(N-1)} S^2_N - \frac{1}{nH} \sum_{h=1}^{h} \left(\bar{y}_{[h]} - \bar{y}_N\right)^2 - \frac{1}{nH} \sum_{h=1}^{H} S_{[h,h]}$$

$$\sigma^2_{JPS} = C_1(n, H) \left\{ \sum_{h=1}^{H} S^2_{[h]} - \sum_{h=1}^{H} S_{[h,h]} \right\} + C_2(n, H, N) \frac{H^2 S^2_N}{H-1},$$

where $S_{[h,h]} = Cov(Y_{[h]1}, Y_{[h]2})$,

$$C_1(n, H) = \left\{ \frac{1}{H(H-1)} + E\left(\frac{I_1^2}{d_n^2 n_1}\right) - \frac{H}{H-1} E\left(\frac{I_1^2}{d_n^2}\right) \right\}$$

$$C_2(n, H, N) = \left\{ Var\left(\frac{I_1}{d_n}\right) - \frac{1}{N-1} \left\{ \frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right) \right\} \right\}.$$

The proofs of σ^2_{JPS} in (i) and (ii) are given in Ozturk (2016a). The proof of σ^2_{RSS} in (ii) is given in Patil et al. (1995). It is clear that the variance of the JPS mean involves expected values and variances of the functions of judgment group indicator function (I_1), sample sizes (n_1), and the number of non-empty judgment groups (d_n). These quantities account for the variation due to the random sample sizes in judgment post-stratified samples. Ozturk (2016b) shows that as the sample size n becomes large, σ^2_{JPS} approaches from above σ^2_{RSS}.

We now introduce unbiased estimators for σ^2_{JPS} and σ^2_{RSS}. We first define the following quantities:

$$U_1 = \frac{1}{E\left(\frac{I_1 I_2}{d_n^2}\right)} \sum_{h=1}^{H} \sum_{h \neq h'}^{H} \frac{I_h I_{h'} J_h J_{h'}}{d_n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (Y_i - Y_j)^2 I(R_i = h) I(R_j = h'),$$

$$U_2 = \sum_{h=1}^{H} \frac{H I_h^* J_h J_h^*}{d_n^*} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (Y_i - Y_j)^2 I(R_i = h) I(R_j = h),$$

$$U_1^* = \frac{1}{2d^2 H^2} \sum_{h=1}^{H} \sum_{h' \neq h}^{H} \sum_{i=1}^{d} \sum_{j=1}^{d} (Y_{[h]i} - Y_{[h']j})^2$$

$$U_2^* = \frac{1}{2d(d-1)H^2} \sum_{h=1}^{H} \sum_{i=1}^{d} \sum_{j \neq i}^{d} (Y_{[h]i} - Y_{[h]j})^2,$$

where $d_n^* = \sum_{h=1}^{H} I(n_h > 1)$, and $J_h^* = 1/(n_h - 1)$ if $n_h > 1$ and zero otherwise.

Theorem 2 *Let $Y_{[h]j}$, $h = 1, \cdots, H$, $j = 1, \ldots, d$ and (Y_j, R_j), $j = 1, \ldots, n$ be RSS and JPS constructed using a consistent ranking method, respectively.*

(i) If the samples are constructed with replacement, $d > 1$ and at least one judgment group in a JPS has at least two observations, the unbiased variance estimators for $\bar{Y}_{RSS}$ and $\bar{Y}_{JPS}$ are given by

$$\hat{\sigma}^2_{JPS} = \frac{Var\,(I_1/d_n)}{2(H-1)} U_1 + \left\{ E\left(\frac{I_1^2}{d_n^2 n_1}\right) - Var\left(\frac{I_1}{d_n}\right) \right\} \frac{U_2}{2}$$

$$\hat{\sigma}^2_{RSS} = \frac{U_2^*}{d}.$$

(ii) If the samples are constructed without replacement, $d > 1$ and at least one judgment group in a JPS has at least two observations, the unbiased variance estimators for $\hat{Y}_{RSS}$ and $\bar{Y}_{JPS}$ are given by

$$\hat{\sigma}^2_{JPS} = C_1(n, H) U_2/2 + C_2(n, H, N) \frac{(N-1)(U_1 + U_{2,2})}{2N(H-1)},$$

$$\hat{\sigma}^2_{RSS} = \frac{U_2^*}{d} - \frac{U_1^* + U_2^*}{N}.$$

Theorem 2 provides unbiased estimators for the variance of the RSS and JPS means for an arbitrary but consistent ranking scheme when $K = 1$. An approximate $(1 - \alpha)100\%$ confidence interval for the population mean can be constructed using the normal approximation:

$$\bar{Y}_{RSS} \pm t_{1-\alpha/2,n-H}\hat{\sigma}_{RSS}$$
$$\bar{Y}_{JPS} \pm t_{1-\alpha/2,n-H}\hat{\sigma}_{JPS},$$

where $t_{a,df}$ is the a-th upper quantile of the t-distribution with degrees of freedom df. The degrees of freedom $df = n - H$ is suggested to account for the heterogeneity among ranking groups.

There are different ways to combine the ranking information in multi-ranking RSS and JPS designs. Ozturk and Kravchuk (2021a, 2021b) provided detailed developments of these procedures. In this paper, we only consider one of the approaches, in which each observation is weighted based on the agreement scores of the K ranking methods. Let $w_{h',i}$ be the proportion of K ranking methods which assign rank h' to the i-th observation in the sample:

$$w_{h'|i,h} = \frac{1}{K}\sum_{k=1}^{K} I(R_{k|i,h} = h')/K, \ h' = 1, \ldots, H, \ \text{for the RSS}$$

and

$$w_{h'|i} = \frac{1}{K}\sum_{k=1}^{K} I(R_{k|i} = h')/K, \ h' = 1, \ldots, H, \ \text{for the JPS.}$$

We estimate the population mean by allocating each observation into ranking group h' based on how strong the agreement is among the K ranking methods to assign the observation to judgment group h':

$$\bar{Y}_{RSS,w} = \sum_{h'=1}^{H} \frac{J_{w,h'}}{d_w} \sum_{h=1}^{H}\sum_{i=1}^{d} Y_{[h]i}\, w_{h'|i,h}, \quad J_{w,h'} = \begin{cases} \frac{1}{n_{w,h'}} & \text{if } n_{w,h'} > 0 \\ 0 & \text{otherwise} \end{cases}, \quad n_{w,h'} = \sum_{i=1}^{d}\sum_{h=1}^{H} w_{h'|i,h}.$$

$$\bar{Y}_{JPS,w} = \sum_{h'=1}^{H} \frac{J_{w,h'}}{d_w} \sum_{h=1}^{H}\sum_{i=1}^{d} Y_i\, w_{h'|i}, \quad J_{w,h'} = \begin{cases} \frac{1}{n_{w,h'}} & \text{if } n_{w,h'} > 0 \\ 0 & \text{otherwise} \end{cases}, \quad n_{w,h'} = \sum_{i=1}^{d}\sum_{h=1}^{H} w_{h'|i},$$

where $d_w = \sum_{h'=1}^{H} I(n_{w,h'} > 0)$. In the expressions above, $n_{w,h'}$ can be considered as the effective sample size for judgment group h'. The asymptotic distribution of $\bar{Y}_{JPS,w}$ is considered in MacEachern et al. (2004) and Ozturk and Kravchuk (2021a). The asymptotic distribution of $\bar{Y}_{RSS,w}$ is given in Ozturk and Kravchuk (2021b).

In this paper, we only consider the jackknife variance estimates of these estimators. Let $\bar{Y}_{RSS,w}^{(-[h]i)}$ ($\bar{Y}_{JPS,w}^{(-i)}$) be the RSS (JPS) estimator after the observation $Y_{[h]i}$ (Y_i) and all ranks associated with it are removed from the sample. The jackknife variance estimates are given by

$$\hat{\sigma}^2_{RSS,J} = fpc\frac{(n-1)^2}{n^2}\sum_{h=1}\sum_{i=1}\left(\bar{Y}^{(-[h]i)}_{RSS,w} - \bar{Y}^{-([.].)}_{RSS,w}\right)^2$$

$$\hat{\sigma}^2_{JPS,J} = fpc\frac{(n-1)^2}{n^2}\sum_{i=1}^{n}(\bar{Y}^{(-i)}_{JPS,w}, -\bar{Y}^{(.)}_{JPS,w})^2$$

$$fpc = \begin{cases} 1-\frac{n}{N-1} & \text{SWOR selection} \\ 1 & \text{SWR selection.} \end{cases}$$

where fpc is the finite population correction factor, $\bar{Y}^{-([.].)}_{RSS,w} = \frac{1}{dH}\sum_{h=1}^{H}\sum_{i=1}^{d} \bar{Y}^{-([h]i)}_{RSS,w}$ and $\bar{Y}^{(.)}_{JPS,w} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}^{(-i)}_{JPS,w}$. In the jackknife variance estimates, we used the coefficient $(n-1)^2/n^2$ since this coefficient provides smaller bias than the usual coefficient $(n-1)/n$, (Ozturk and Kravchuk, 2021a,b).

An approximate $(1-\alpha)100\%$ confidence interval for multi-ranking RSS and JPS designs can be constructed using the jackknife variance estimates:

$$\bar{Y}_{RSS,w} \pm t_{1-\alpha/2,n-H}\hat{\sigma}_{RSS,J}$$
$$\bar{Y}_{JPS,w} \pm t_{1-\alpha/2,n-H}\hat{\sigma}_{JPS,J}.$$

In the next section, we compare the RSS and JPS estimators in terms of their efficiencies and coverage probabilities for a varying degree of ranking quality and different set sizes.

4 Comparison of RSS and JPS Designs

We performed a simulation study to investigate the contrasting features of RSS and JPS estimators. In the simulation study, samples were generated from two finite populations with large population size $N = n + 1000$ and small population sizes $N = nH + 50$. We considered a normal, $N(\mu = 50, \sigma = 5)$, and a lognormal, $LN(\mu = 0, \sigma = 1)$, distribution. The population values of the response variable Y were generated using the quantile functions:

$$y_i = F_N^{-1}(i/(N+1), \mu = 50, \sigma = 5), \text{and } y_i = F_{LN}^{-1}(i/(N+1), \mu = 0, \sigma = 1), i = 1, \ldots, N,$$

where $F_N^{-1}(y, \mu, \sigma)$ and $F_{LN}^{-1}(y, \mu, \sigma)$ are the inverse cumulative distribution functions of a normal distribution with location parameter μ and scale parameter σ and lognormal distribution with scale parameter $exp(\mu)$ and shape parameter σ, respectively. The samples were generated using SWR and SWOR selection procedures for both population sizes $N = n + 1000$ and $N = nH + 50$. The quality of ranking was modeled using a ranking variable X, such that $X = Y + \tau\epsilon$, where ϵ has a normal distribution with mean zero and variance 1 and independent of Y. The correlation coefficient between X and Y is given by $\rho = \frac{1}{\sqrt{1+\tau^2/\sigma^2}}$. The values of ρ were selected to be $0.01, 0.25, 0.5, 0.75, 0.9, 1$ where values less than 1 will result in imperfect ranking. For the normal distribution, we fixed the sample size at $n = 36$ and varied the set sizes as $H = 2, 3, 4, 6, 12$ to explore the impact of different set sizes on the RSS and JPS designs. We purposely selected a smaller sample size $n = 36$ to evaluate the approximation of the coverage probabilities of the confidence intervals to the nominal coverage probability 0.95. The simulation size is taken to be 5000. An R-package *RankedSetSampling* (Ozturk et al., 2021) is used to compute the estimators and construct confidence intervals. The package is available to download at https://biometryhub.github.io/RankedSetSampling.

We first investigate the efficiencies of the RSS and JPS estimators. The relative efficiencies are defined as the ratio of the mean square errors of the RSS and JPS estimators:

$$RE = \frac{MSE(JPS)}{MSE(RSS)}.$$

A value of RE greater than 1 indicates that the RSS estimator is more efficient than the JPS estimator. Figure 1 presents the relative efficiencies for the population size $N = n + 1000$ when samples were generated using the SWR selection procedure. The set sizes and the number of ranking methods are indicated in the legend on each panel. The first panel shows the relative efficiency curves when both RSS and JPS were generated with just one ranking method $K = 1$. It is clear in this case that the RSS estimator is more efficient. The efficiency gain is minimal for $H = 2, 3$, moderate for $H = 4, 6$, and substantial for $H = 12$. This intuitively makes sense since large set sizes lead to many judgment groups having no measured observations in a JPS. Empty ranking groups inflate the variance of the JPS estimator. The RE values are similar to each other for all ρ values when $H = 2, 3, 4, 6$, except for ρ when $H = 12$ where it increases.

Figure 1 also presents the relative efficiencies in three different panels when different number of ranking methods (K=2, 5,10) is used. Comparing these panels with panel 1, one can see that the gain in RE values decreases with the number of ranking methods K. For example, the RE values in panel 1 ($K = 1$) are around 1.4 when $\rho < 0.75$, and it reduces essentially to 1 in panel 4 (K=10). Similar observation can be made in panel 2 ($K = 2$) and panel 3 ($K = 5$). Under perfect ranking, RSS is still superior to the JPS for all set sizes.

Figure 2 presents the efficiency curves for the small population size, $N = nH + 50$. In this part of the simulation, both the RSS and JPS were generated under the SWOR selection procedures. The efficiency results are similar to those in Fig. 1, with the key difference being that the RE curves are higher (lower) in Fig. 2 than in Fig. 1 when $K = 1$ and $K = 2$ ($K = 5$ and $K = 10$).

These efficiency results indicate that RSS estimator is more efficient than the JPS estimator when the number of ranking methods is small ($K = 1, 2$). For the larger

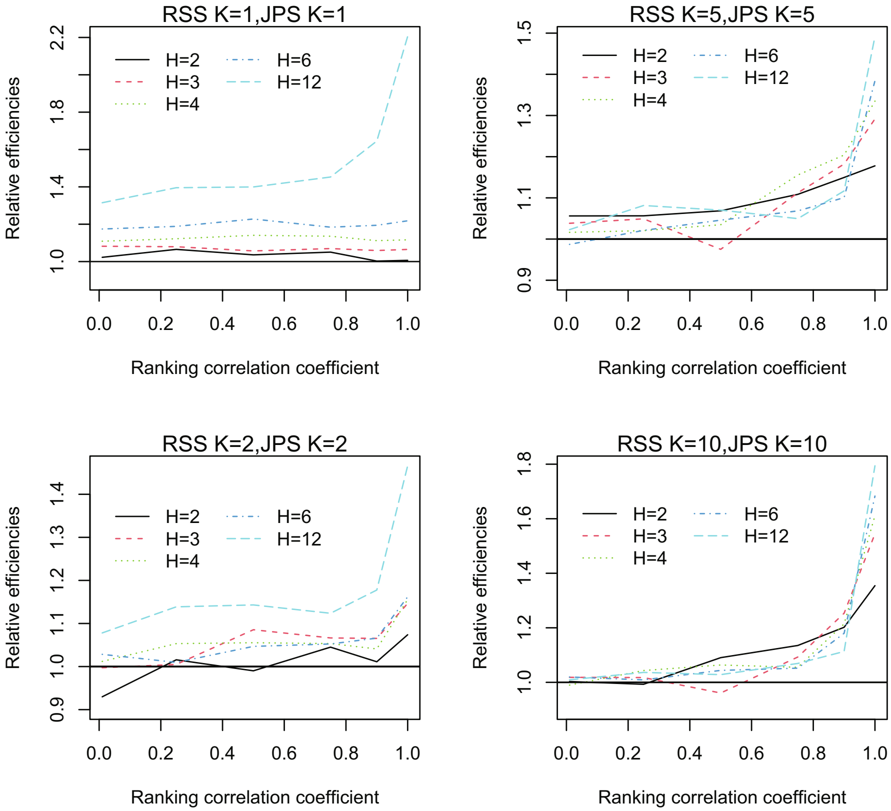


Fig. 1 Efficiency comparison of RSS and JPS designs under SWR selection for large-sized normal distribution population

number of ranking methods ($K = 5, 10$), difference between the efficiency gain of RSS and JPS estimators diminishes.

We also investigated the coverage probabilities of the confidence intervals for the population mean. Figure 3 presents the coverage probabilities for the samples constructed with replacement from the population of size $N = n + 1000$. We note that confidence intervals are constructed using unbiased variance estimates when $K = 1$. For $K \neq 1$, we used the jackknife variance estimates. The panels in the first and second columns of Fig. 3 present the coverage probabilities of RSS and JPS confidence intervals for $K = 1, 2, 5, 10$, respectively. The coverage probabilities of RSS confidence intervals can be seen to be reasonably close to the nominal coverage probability of 0.95 when $\rho \leq 0.75$ and $K = 2, 5, 10$, but they are slightly larger when $\rho > 0.75$ and $K = 2, 5, 10$. The coverage probabilities in the second column of Fig. 3 are reasonably close to the nominal coverage probability 0.95 for all ρ and K values in the simulation study.

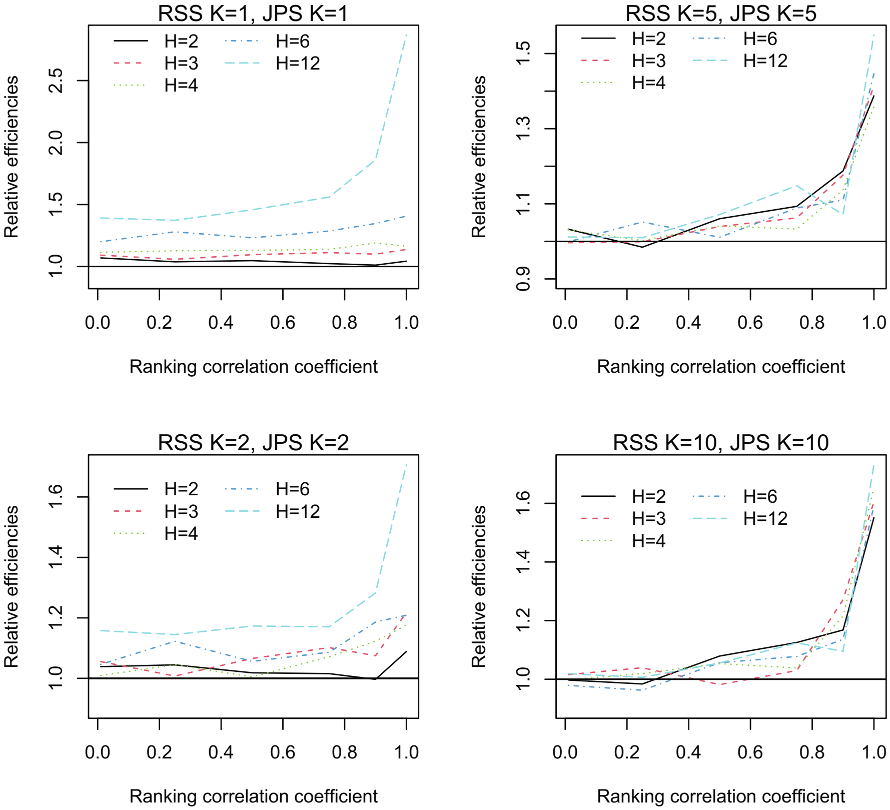


Fig. 2 Efficiency comparison of RSS and JPS designs under SWOR selection for the small-sized normal distribution population

Figure 4 presents the coverage probabilities for the population size $N = nH+50$. In this case, coverage probabilities are again close to nominal coverage probability of 0.95 under imperfect ranking ($\rho \leq 0.75$) for both RSS and JPS and $K = 1, 2, 5, 10$. Unlike Fig. 3, coverage probabilities are slightly inflated for both RSS and JPS confidence intervals when $\rho > 0.75$ and $K = 2, 5, 10$. Under perfect ranking ($\rho = 1$), jackknife variance estimator overestimates the variances of the RSS and JPS estimators and leads to a larger coverage probability than the nominal coverage probability of 0.95

In the second part of the simulation study, we generated samples from the lognormal distribution with the scale parameter $exp(\mu)(\mu = 0)$ and the shape parameter $\sigma = 1$. The sample and set sizes were as previously $n = 48$ and $H = 2, 3, 4, 6, 12$. All the other simulation parameters remained the same. The lognormal distribution is strongly positively skewed. For this reason, we increased the sample size from 36 to 48. Figures 5 and 6 present the relative efficiencies of the RSS and JPS estimators for large and small population sizes, respectively. The pattern

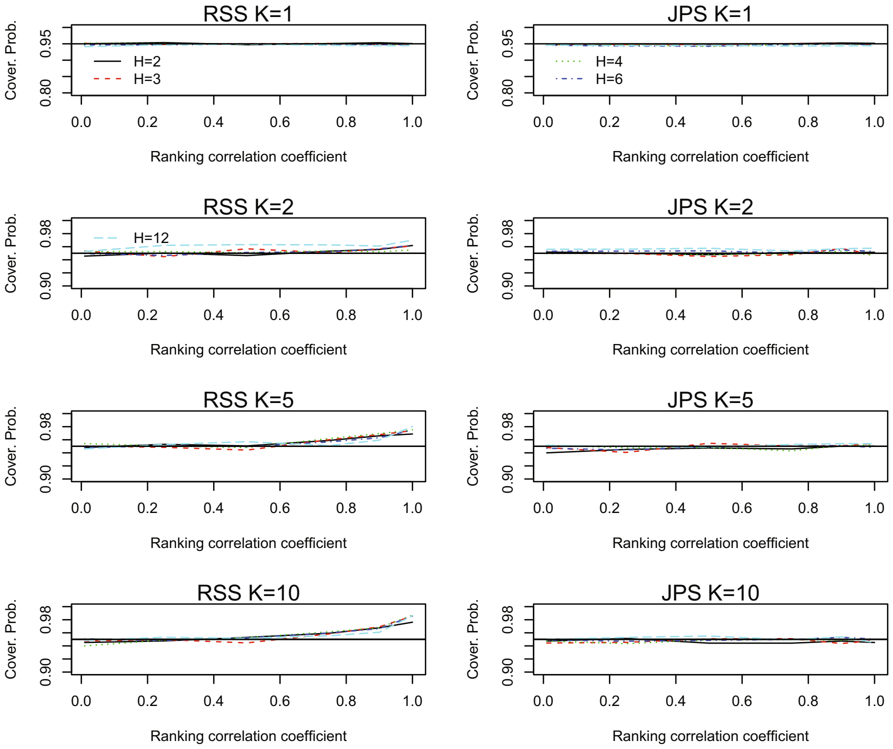


Fig. 3 Coverage probabilities of the jackknife confidence intervals under SWR selection for normal distribution

of the efficiency curves is very similar to that for the normal population. The main difference is in the magnitude of the efficiency gain. The efficiency curves reach to higher values for the normal distribution. This result is consistent with the efficiency results of ranked set samples in (McIntyre, 1952, 2005). McIntyre reported that the efficiencies are higher for symmetric distributions (highest for the uniform distribution) and decrease with skewness. Since the lognormal distribution has strong skewness, the efficiencies are slightly lower than for the normal distribution.

Figures 7 and 8 present the coverage probabilities of the jackknife confidence intervals of the population mean for the SWR and SWOR designs, respectively. It is clear that the coverage probabilities for the lognormal distribution are lower than the nominal coverage probability 0.95. The SWOR selection provides a better coverage probability than the SWR selection. Since a jackknife confidence interval relies on the normal approximation, the sample size $n = 48$ is not large enough for a good approximation when the underlying population is strongly skewed.

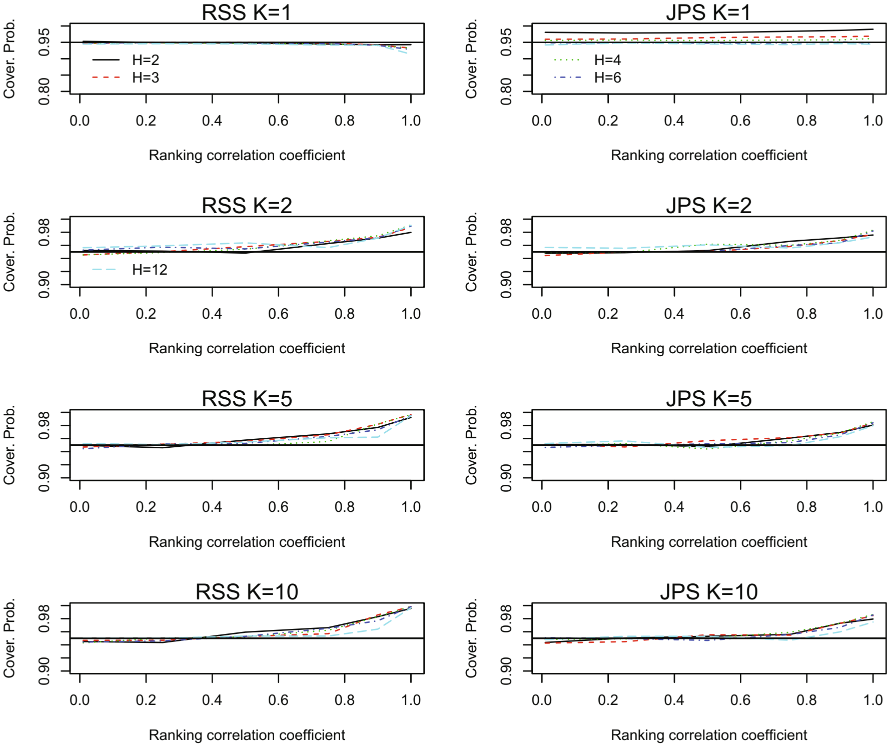


Fig. 4 Coverage probabilities of the jackknife confidence intervals under SWOR selection for normal distribution

5 Application

In this section, we use a real-life finite population example to compare the JPS and RSS estimators. The population consisted of 350 grapevine plants at Coombe vineyard at the University of Adelaide, Waite campus, Australia. The vineyard is used as a research and teaching facility. There are eight different rootstocks originally planted, on which Shiraz is grafted. These rootstocks are popular commercial choices in South Australia. The standard vineyard management of this population requires the monitoring and measuring of certain characteristics of vine plants. In this paper, we consider seven characteristics; X_1, trunk circumference (cm) in 2018; X_2, trunk circumference (cm) in 2019; X_3, shoot counts; X_4, total shoots; X_5, pruning weight (kg); X_6, cordon length (cm); and X_7, total bunch numbers and Y, nett fruit weight in 2019 (kg). Our interest was in the estimation of the mean nett fruit yield of this population of grapevines in 2019. The variables X_i, $i = 1, \ldots, 7$, were used as ranking variables in comparison sets, and hence, the number of ranking methods is $K = 7$. There were missing values on some vines, and after removing the plants having missing observations, the population size was reduced to $N = 309$. In

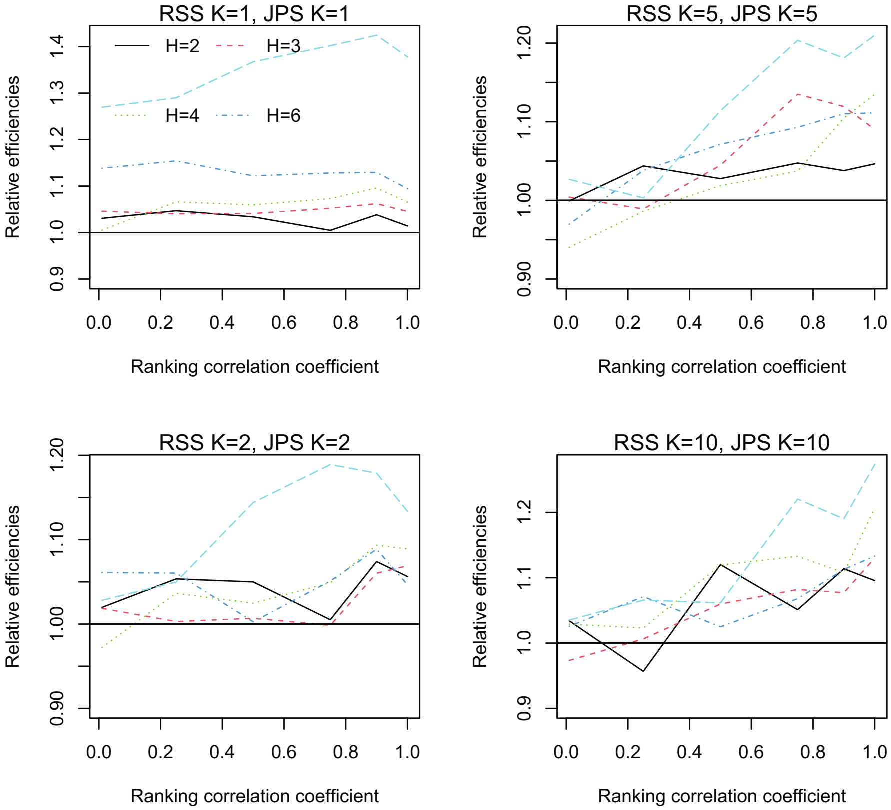


Fig. 5 Efficiency comparison of RSS and JPS designs under SWR selection for the large-sized lognormal distribution population

this population, the correlation coefficients between Y and X_i, $\rho_i = cor(Y, X_i)$ are $\rho_1 = 0.240$, $\rho_2 = 0.191$, $\rho_3 = 0.310$, $\rho_4 = 0.321$, $\rho_5 = 0.172$, $\rho_6 = 0.274$, and $\rho_7 = 0.713$. The mean and standard deviation of the Y variable are 10.558 kg and 3.855 kg, respectively.

We performed another simulation study using these 309 vine plants. In each replication of the simulation study, we generated the single-ranking judgment post-stratified and ranked set samples with the ranking variable X_7 ($K = 1$), the multi-ranking judgment post-stratified and ranked set samples with X_k, $k = 1, \ldots, 7$ ($K = 7$), and a simple random sample. The sample sizes were selected to be $n = 30$ and 48. For the sample size $n = 30$, the set sizes were chosen $H = 3, 5, 6, 10$. For the sample sizes $n = 48$, the set sizes were $H = 3, 4, 6$. Samples were generated using the SWR and SWOR selection procedures. The simulation size was 5000.

Table 3 presents the relative efficiency of the multi-ranking RSS estimator ($K = 7$) with respect to the other four estimators: the JPS estimator with $K = 7$ and

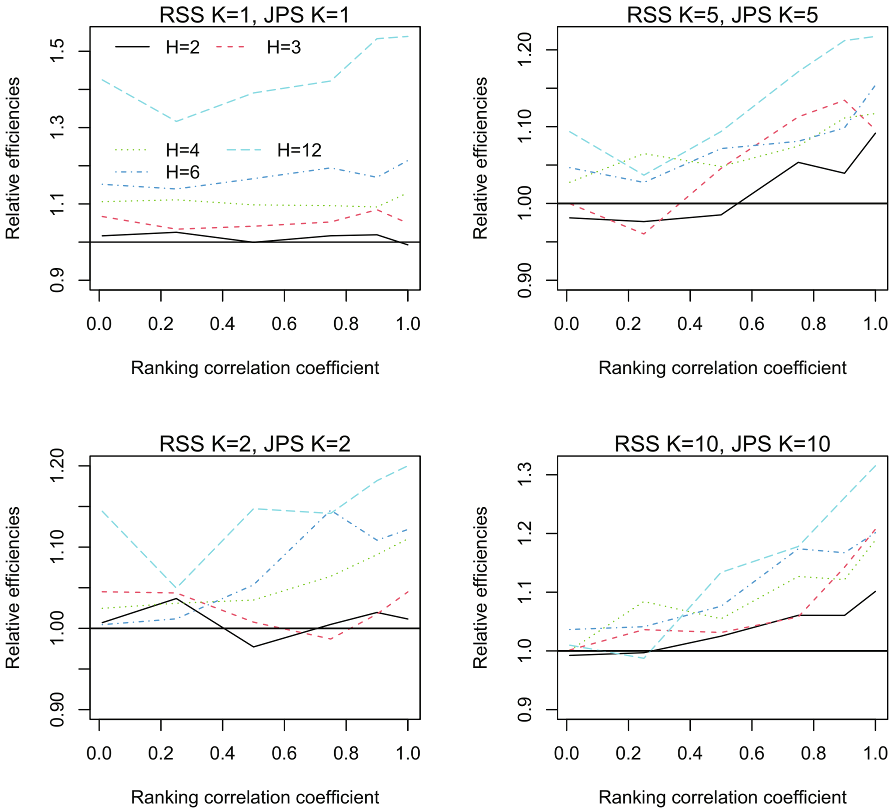


Fig. 6 Efficiency comparison of RSS and JPS designs under SWOR selection for the small-sized lognormal distribution population

$K = 1$, the SRS estimator, and the RSS estimator with $K = 1$. When the entries in Table 3 are greater than one, the multi-ranking RSS estimator with $K = 7$ was superior. The other efficiency results can be obtained by taking the ratio of any two efficiency columns in Table 3. For example, the efficiency of the JPS estimator with $K = 1$ relative to the SRS estimator can be obtained by taking the ratio of column 6 and column 5. When $n = 30$, $H = 3$, and the replacement is true, this efficiency is calculated $1.246(1.321/1.060 = 1.246)$. The other relative efficiencies can be computed in a similar fashion.

All entries in Table 3 are greater than one which indicates that the RSS multi-ranking estimator with $K = 7$ is more efficient than JPS and SRS estimators. The efficiencies of RSS estimator with $K = 7$ with respect to JPS and SRS estimators increase with set sizes, but remain relatively constant with RSS estimator wit $K = 1$ (column 7). The reason for this is that the correlation coefficient between ranking variable X_7 and response is 0.729, while the other correlation coefficients are all less than 0.321. Hence, the improvement of ranking quality due to ranking variables with

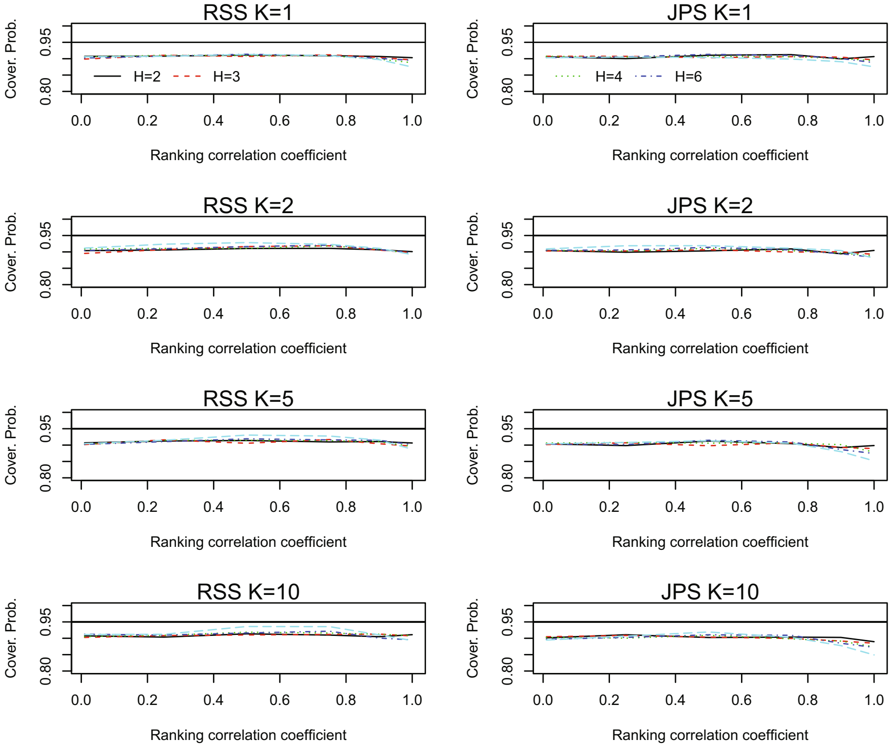


Fig. 7 Coverage probabilities of the jackknife confidence intervals under SWR selection for large-sized lognormal distribution population

low correlation coefficients is minimal, and the relative efficiency for multi-ranker estimator remains relatively constant. For this particular population and ranking methods, the JPS estimators are more efficient than the SRS estimator and less efficient than multi-ranker RSS estimator.

We also computed the coverage probabilities of the confidence intervals based on the judgment post-stratified, simple random, and ranked set samples for the population mean. All coverage probabilities were reasonably close to the nominal coverage probability of 0.95. Due to space considerations, these empirical coverage probabilities are not reported here.

6 Concluding Remarks

Field research is expensive and time-consuming, particularly in natural environments where variables are difficult to control. If auxiliary variables are available, they can be used to account in the analysis for the inherent variation among sampling

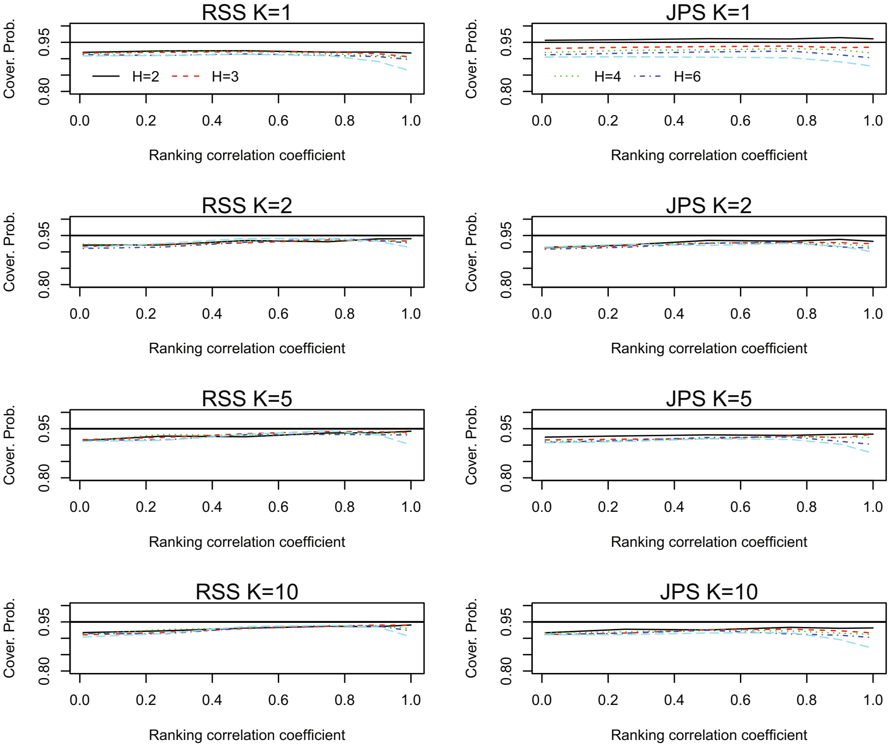


Fig. 8 Coverage probabilities of the jackknife confidence intervals under SWOR selection for small-sized lognormal distribution population

units. These auxiliary variables can be used as blocking variables if they can be evaluated in an objective manner. In certain settings, auxiliary variables may not be assessed accurately. Their assessment may be rough, imprecise, and subjective, but still helpful for ordering the units in a small set independently of knowing the actual values of the variable of interest.

Ranked set and judgment post-stratified sampling designs use this ordering information to construct samples that are more likely to span the full range of values in the population. It has been established in the literature that a ranked set sample is generally more efficient than a judgment post-stratified sample. However, RSS designs induce a strong structure in the sample. Hence, an RSS cannot be analyzed with the inferential procedures developed for an SRS design.

The JPS design may be less efficient than the RSS design, but the sample constructed can be reduced to a simple random sample, allowing the flexibility to perform multiple analyses of various responses on the same data set. This becomes useful if the data set is needed for a multipurpose study. In this paper, we show how to reduce the efficiency loss of a JPS with respect to an RSS by constructing multiple ranks for the response variable on each measured unit. Hence, the JPS

Table 3 Relative efficiency of the weighted RSS estimators with $K = 7$ for mean fruit yield of vine plants in Coombe vineyard. Entries greater than one indicate that the RSS estimator with $K = 7$ is more efficient

			JPS			RSS
Replace	n	H	$K = 7$	$K = 1$	SRS	$K = 1$
True	30	3	1.209	1.060	1.321	1.018
	30	5	1.283	1.203	1.433	1.020
	30	6	1.335	1.261	1.500	1.016
	30	10	1.508	1.473	1.724	1.011
	48	3	1.227	1.042	1.352	1.022
	48	4	1.270	1.042	1.419	1.026
	48	6	1.403	1.155	1.603	1.030
False	30	3	1.287	1.116	1.407	1.026
	30	5	1.390	1.224	1.560	1.025
	30	6	1.399	1.313	1.575	1.024
	30	10	1.633	1.619	1.860	1.003
	48	3	1.287	1.075	1.426	1.028
	48	4	1.317	1.098	1.475	1.028
	48	6	1.533	1.261	1.757	1.030

design provides the flexibility for multipurpose analysis at the expense of little efficiency loss with respect to a balanced ranked set sample. Another advantage of the JPS design is that it is relatively straightforward to construct a multi-ranking JPS even when there are no additional auxiliary ranking variables, and this can be done by permuting the units selected to form comparison sets. This idea is not easily extended to a ranked set sampling. We would recommend that the JPS design should be considered in field sampling, especially for multipurpose studies.

Acknowledgments This research is supported by the Australian Grains Research and Development Corporation (GRDC) as part of the Statistics for the Australian Grains Industry project (UA00164). The data was collected by Bachelor of Agricultural Science students of the School of Agriculture, Food and Wine, University of Adelaide, in 2019 under the supervision of Mr. Peter Kasprzak. The data was kindly made available for our purposes and other publications, as well as to improve the management and research decisions in the Coombe vineyard.

References

Arslan, G., & Ozturk, O. (2013). Parametric inference based on partially rank ordered set samples. *Indian Journal of Statistics, 51*, 1–24.

Bohn, L. L., & Wolfe, D. A. (1992). Nonparametric two-sample procedures for ranked set samples data. *Journal of the American Statistical Association, 87*(418), 552–561.

Bohn, L. L., & Wolfe, D. A. (1994). The effect of imperfect judgment rankings on properties of procedures based on the ranked set samples analog of the Mann-Whitney-Wilcoxon statistic. *Journal of the American Statistical Association, 89*(425), 168–176.

Bouza, C., & Al-Omari, A. I. (2019). *Ranked set sampling: 65 years improving the accuracy in data gathering*. Elsevier.

Chen, Z., & Bai, Z. (2000). The optimal ranked-set sampling scheme for parametric families. *Sankhyā: The Indian Journal of Statistics, Series A, 62*(2), 178–192.

Chen, Z., Bai, Z., & Sinha, B. K. (2003). *Ranked set sampling*. Springer.

Chen, M., Ahn. S., Wang, X., & Lim, J. (2014). Generalized isotonized mean estimators for judgment post-stratification with multiple rankers. *Journal of Agricultural, Biological, and Environmental Statistics, 19*, 405–418.

Dell, T. R., & Clutter, J. L. (1972). Ranked-set sampling theory with order statistics background. *Biometrics, 28*, 545–555.

Deshpande, J. V., Frey, J., & Ozturk, O. (2006). Nonparametric ranked set-sampling confidence intervals for a finite population. *Environmental and Ecological Statistics, 13*, 25–40.

Fligner, M. A., & MacEachern, S. N. (2006). Nonparametric two-sample methods for ranked set sample data. *Journal of the American Statistical Association, 101*(475), 1107–1118.

Frey, J. (2011). Recursive computation of inclusion probabilities in ranked set sampling. *Journal of Statistical Planning and Inference, 141*, 3632–3639.

Frey, J., & Feeman, T. G. (2012). An improved mean estimator for judgment post-stratification. *Computational Statistics & Data Analysis, 56*, 418–426.

Frey, J., & Feeman, T. G. (2013). Variance estimation using judgment post-stratification. *Annals of the Institute of Statistical Mathematics, 5*, 551–569

Frey, J., & Ozturk, O. (2011). Constrained estimation using judgment post-stratification. *Annals of the Institute of Statistical Mathematics, 3*, 769–789.

Hatefi, A., Jafari Jozani, M., & Ziou, D. (2014). Estimation and classification for finite mixture models under ranked set sampling. *Statistica Sinica, 24*, 675–698.

Hatefi, A., Jafari Jozani, M., & Ozturk, O. (2015). Mixture model analysis of partially rank-ordered set samples: Age groups of fish from length-frequency data. *Scandinavian Journal of Statistics, 42*, 848–871.

Hettmansperger, T. P. (1995). The ranked-set sampling sign test. *Nonparametric Statistics, 4*, 263–270.

Jafari Jozani, M., & Johnson, B. C. (2011). Design based estimation for ranked set sampling in finite populations. *Environmental and Ecological Statistics, 18*, 663–685.

Koti K. M., & Babu, J. G. (1996). Sign test for ranked-set sampling. *Communications in Statistics - Theory and Methods, 25*(7), 1617–1630.

MacEachern, S. N., Stasny, E. A., & Wolfe, D. A. (2004). Judgment post-stratification with imprecise rankings. *Biometrics, 60*, 207–215.

McIntyre, G. (1952). A method for unbiased selective sampling using ranked set sampling. *Australian Journal of Agriculture Research, 3*, 385–390.

McIntyre, G. A. (2005). A method of unbiased selective sampling using ranked-sets. *The American Statistician, 59*, 230–232.

Ozturk, O. (1999). Two-sample inference based on one-sample ranked set sample sign statistics. *Nonparametric Statistics, 10*, 197–212.

Ozturk, O. (2013). Combining multi-observer information in partially rank-ordered judgment post-stratified and ranked set samples. *The Canadian Journal of Statistics, 41*, 304–324.

Ozturk, O. (2014). Statistical inference for population quantiles and variance in judgment post-stratified samples. *Computational Statistics and Data Analysis, 77*, 188–205.

Ozturk, O. (2016a). Statistical inference based on judgment post-stratified samples in finite population. *Survey Methodology, 42*, 239–262.

Ozturk, O. (2016b). Estimation of finite population mean and total using population ranks of sample units. *Journal of Agricultural, Biological, and Environmental Statistics, 21*, 181–202.

Ozturk, O. (2017). Statistical inference with empty strata in judgment post stratifed samples. *Annals of the Institute of Statistical Mathematics, 69*, 1029–1057.

Ozturk, O. (2019). Statistical inference using rank based post-stratified samples in a finite population. *Test, 28*, 1113–1143.

Ozturk, O., & Bayramoglu Kavlak, K. (2018). Model based inference using ranked set samples. *Survey Methodology, 44*(1), 1–16, Catalogue No. 12-001-X.

Ozturk, O., & Bayramoglu Kavlak, K. (2019). Statistical inference using stratified ranked set samples from finite populations. In: C. Bouza & A. I. Al-Omari (Eds.), *Ranked set sampling: 65 years improving the accuracy in data gathering* (pp 157–170). Elsevier.

Ozturk, O., & Bayramoglu Kavlak, K. (2020). Statistical inference using stratified judgment post-stratified samples from finite populations. *Environmental and Ecological Statistics, 27*, 73–94.

Ozturk, O., & Demirel, N. (2016). Estimation of population variance from multi-ranker ranked set sampling designs. *Communications in Statistics-Simulation and Computation, 45*(10), 3568–3583.

Ozturk, O., & Jafari Jozani, M. (2014). Inclusion probabilities in partially rank ordered set sampling. *Computational Statistics and Data Analysis, 69*, 122–132.

Ozturk, O., & Kravchuk, O. (2021a) Combining ranking information from different sources in ranked set samples. *Canadian Journal of Statistics*. https://doi.org/10.1002/cjs.11656

Ozturk, O., & Kravchuk, O. (2021b). Judgment post-stratified assessment combining ranking information from multiple sources, with a field phenotyping example. *Journal of Agricultural, Biological and Environmental Statistics*. https://doi.org/10.1007/s13253-021-00439-1

Ozturk, O., Rogers, S., Kravchuk, O., & Kasprzak, P. (2021). RankedSetSampling: Easing the application of ranked set sampling in practice. R package version 0.0.1. https://biometryhub.github.io/RankedSetSampling/

Patil, G. P., Sinha, A. K., & Taillie, C. (1995). Finite population corrections for ranked set sampling. *Annals of the Institute of Statistical Mathematics, 47*, 621–636.

Stokes, L. (1995). Parametric ranked set sampling. *Annals of the Institute of Statistical Mathematics, 47*, 465–482.

Stokes, S. L., Wang, X., & Chen, M. (2007). Judgment post-stratification with multiple rankers. *Journal of Statistical Theory and Applications, 6*, 344–359.

Takahasi, K., & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics, 20*, 1–31.

Wang, X., Stokes, S. L., Lim, J., & Chen, M. (2006). Concomitants of multivariate order statistics with application to judgment post-stratification. *Journal of the American Statistical Association, 101*, 1693–1704.

Wang, X., Lim, J., & Stokes, S. L. (2008). A nonparametric mean estimator for judgment post-stratified data. *Biometrics, 64*, 355–363.

Wang, X., Wang, K., & Lim, J. (2012). Isotonized CDF estimation from judgment poststratification data with empty strata. *Biometrics, 68*, 194–202.

Efficient Sample Allocation by Local Adjustment for Unbalanced Ranked Set Sampling

Soohyun Ahn, Xinlei Wang, and Johan Lim

Abstract In applications that demand cost efficiency, balanced ranked set sampling (BRSS) is a well-established alternative to simple random sampling (SRS), which is proved to be more efficient in estimating the population mean than its SRS counterpart. The efficiency of BRSS can be further improved by considering unbalanced RSS (URSS) with appropriate unequal allocation. However, with a poor sample allocation scheme, URSS can have even worse performance than SRS. Conditions that render a URSS design more efficient than its BRSS counterpart have been rarely studied in the literature. For a fixed total sample size n and a fixed set size H, we characterize a sufficient set of allocation schemes in which estimation of the population mean from URSS is guaranteed to be more efficient than the BRSS counterpart. We illustrate this set using a simplex diagram based on $H = 3$ and compute theoretical relative efficiency over SRS under distributions with either heavy tails or skewness. We further consider two adjustment procedures of a URSS design that is less efficient than its BRSS counterpart. We numerically investigate their performance under various simulation settings and apply them to redesign less efficient URSS in realistic scenarios where BRSS is initially planned but unequal sample sizes in rank strata are caused by the prevalent issue of missing data.

S. Ahn
Department of Mathematics, Ajou University, Gyeonggi, Korea
e-mail: shahn@ajou.ac.kr

X. Wang
Department of Statistical Science, Southern Methodist University, Dallas, TX, USA
e-mail: swang@smu.edu

J. Lim (✉)
Department of Statistics, Seoul National University, Seoul, Korea
e-mail: johanlim@snu.ac.kr

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_3

1 Introduction

Randomized experiments play a vital role in modern scientific discovery. In general, they rely on simple random sampling (SRS) to recruit units, in which the efficiency of such experiments can be improved by simply increasing the sample size. However, in situations with constrained resources, ranked set sampling (RSS) can serve as a cost-effective alternative to SRS. RSS is a type of stratified sampling method, which uses auxiliary rank information to form strata (Stokes and Sager, 1988). If the experiment design considers the same number of replicates for each rank stratum, it is called balanced RSS (BRSS); otherwise, unbalanced RSS (URSS). The efficiency of RSS has been studied with a rich history (Chen et al., 2006 and references therein). It is well-known that BRSS offers more precise estimation than its SRS counterpart (i.e., SRS with the same sample size). The efficiency of BRSS can be further improved by implementing URSS with appropriate unequal allocation. Especially, when the underlying distribution is highly skewed, an URSS estimator can be (much) more efficient relative to its BRSS and SRS counterparts (Ahn et al., 2017; Chen & Bai, 2000; Bocci et al., 2010; Ozturk & Wolfe, 2004; Wang et al., 2017). However, if the number of replicates for each stratum in URSS is not properly assigned, its performance can be even worse than that of SRS. Conditions that render a URSS design more efficient than its BRSS counterpart (i.e., the balanced design with the same set size and sample size) remain largely unexplored.

Some proper allocation rules for RSS have been suggested in the literature to achieve better efficiency than the default BRSS (Bhoj & Chandra, 2019; Chen & Bai, 2000; McIntyre, 1952; Wang et al., 2004). Among them, the Neyman allocation, which allocates sample units into rank strata in proportion to the standard deviation of each stratum, is the most popular due to the optimality that it has the smallest variance in estimating the population mean. For this reason, most of the existing literature on URSS has focused on the Neyman allocation (Chen et al., 2006; Takahasi & Wakimoto, 1968; Wang et al., 2017). However, the Neyman allocation is "one of many" allocation schemes which are more efficient in estimating the population mean than their BRSS counterparts. In addition, it globally depends on the variances of all rank strata and so lacks flexibility in practice. In other words, suppose the current sampling scheme is not the Neyman optimal and further is less efficient than the balanced design, due to various complications and limitations in implementation. One may want to make it become the Neyman optimal or at least more efficient than its BRSS counterpart by some (small local) adjustment (e.g., adding a few more samples to a few rank strata). For the Neyman allocation, the sample size of "one" stratum depends on the variances of all other strata, and thus, if we add more samples, we have to do so for most of the rank strata, and this is often costly in practice.

The main purpose of this paper is to define a sufficient set of allocation schemes, in which estimation of the population mean from URSS is guaranteed to be more efficient than the BRSS counterpart. This sufficient set is characterized by the sample sizes of neighboring strata so that when an allocation scheme is not in the set, we can easily fix it. We further propose a local adjustment procedure based on the sufficient set that renders the resulting design more efficient than its BRSS counterpart when it is originally not. For comparison, we also consider a naive procedure that intends to achieve proximity to the optimal Neyman without discarding existing observations.

This paper is organized as follows. The sufficient set of sample allocation schemes that are more efficient than BRSS counterparts, denoted by $\mathscr{N}$, is proposed in Sect. 2. We illustrate the set $\mathscr{N}$ using a simplex diagram and compute theoretical relative efficiency (RE) over SRS for a fixed set size under various distributions (with either heavy tails or skewness) in Sect. 3. In Sect. 4, we consider two local adjustment procedures to modify a design that is less efficient than BRSS, one of which is based on $\mathscr{N}$, while the other is not. We compare the two procedures in terms of the number of added samples (the cost of reallocation) and the efficiency gain per an additional sample. In Sect. 5, we apply two methods to an educational data example. Finally, we conclude the paper with a brief summary in Sect. 6.

2 More Efficient URSS than BRSS

Suppose that we have RSS data with a set size H and a total sample size $n = \sum_{h=1}^{H} n_h$, where n_h is the number of measured units with rank h. Note that for BRSS, $n_h \equiv n/H$ for $h = 1, \cdots, H$ and let $m = n/H$. Here, we find a condition for the sample allocation $\mathbf{n} = \big(n_1, n_2, \ldots, n_H\big)$ that makes the URSS with $\mathbf{n}$ more efficient than its balanced counterpart in estimating the population mean. Let $\widehat{\mu}_{\text{RSS}}$ denote the RSS mean estimator, where $\widehat{\mu}_{\text{RSS}} = \frac{1}{H}\sum_{h=1}^{H} \bar{Y}_h$ and $\bar{Y}_h$ is the sample mean in the h-th stratum that contains all measured units with rank h. The variance of $\widehat{\mu}_{\text{RSS}}$ is

$$\text{V.rss}(\mathbf{n}) := \text{Var}\big(\widehat{\mu}_{\text{RSS}}\big)(\mathbf{n}) = \frac{1}{H^2}\sum_{h=1}^{H}\frac{\sigma^2_{[h]}}{n_h}$$

where $\sigma^2_{[h]}$ is the variance of the h-th rank stratum. For BRSS, it is known that the variance of the mean estimator is smaller than that of SRS with the same sample size n. On the other hand, for some unequal allocation $\mathbf{n}$, the URSS mean estimator has a larger variance than BRSS or sometimes SRS estimators with the same sample size, i.e.,

$$\text{V.rss}(\mathbf{m}) = \frac{1}{H^2}\sum_{h=1}^{H}\frac{\sigma^2_{[h]}}{m} \leq \frac{\sigma^2}{n} = \text{Var}(\hat{\mu}_{\text{SRS}}) \leq \text{V.rss}\big(\mathbf{n}\big)$$

where $\mathbf{m} = (m, m, \cdots, m)$ is the equal allocation for BRSS and $n = mH$; σ^2 is the population variance.

The optimal RSS design in estimating the population mean has been studied in Takahasi and Wakimoto (1968), which adopts the Neyman allocation and has the smallest variance. We denote the Neyman allocation by $\tilde{\mathbf{n}} = (\tilde{n}_1, \tilde{n}_2, \cdots, \tilde{n}_H)$ where

$$\tilde{n}_h = \frac{\sigma_{[h]}}{\sum_{l=1}^{H} \sigma_l} \cdot n$$

and

$$\text{V.rss}\big(\tilde{\mathbf{n}}\big) \leq \text{V.rss}(\mathbf{m}) \leq \text{Var}(\hat{\mu}_{\text{SRS}}).$$

As mentioned in the introduction, the Neyman allocation is not the only allocation scheme whose mean estimator is more efficient than that of BRSS. For simplicity, we relabel the strata to have monotonicity in stratum variances: $\sigma_1^2 \leq \sigma_2^2 \leq \cdots \leq \sigma_H^2$ and let n_h be the number of units for the corresponding h-th stratum. Note that after relabeling, units in the h-th stratum no longer have rank h and so we use σ_h^2 instead of $\sigma_{[h]}^2$. For a fixed total sample size n and a fixed set size H, the set $\mathscr{N}_0$, defined by

$$\mathscr{N}_0 = \left\{ \mathbf{n} = \big(n_1, n_2, \ldots, n_H\big) \middle| \text{V.rss}(\mathbf{n}) = \frac{1}{H^2} \sum_{h=1}^{H} \frac{\sigma_h^2}{n_h} \leq \frac{1}{H^2} \sum_{h=1}^{H} \frac{\sigma_h^2}{m} = \text{V.rss}(\mathbf{m}) \right\},$$

is the collection of all sample allocation schemes that is more efficient than the BRSS with $m = n/H$ in estimating the population mean.

We proceed to consider the sample allocation set

$$\mathscr{N} = \left\{ \mathbf{n} = \big(n_1, n_2, \ldots, n_H\big) \middle| 1 \leq \frac{n_{h+1}}{n_h} \leq \frac{\sigma_{h+1}^2}{\sigma_h^2}, h = 1, 2, \ldots, H-1 \right\},$$

which is a subset of $\mathscr{N}_0$, as will be shown in Theorem 1. That is, the condition in the set $\mathscr{N}$ is sufficient to make an URSS design more efficient than the BRSS design, but it is not a necessary condition. Note that the Neyman allocation $\tilde{\mathbf{n}}$ is included in $\mathscr{N}$ because $1 \leq \tilde{n}_{h+1}/\tilde{n}_h = \sigma_{h+1}/\sigma_h \leq \sigma_{h+1}^2/\sigma_h^2$.

Theorem 1 *(a) If* $\mathbf{n} \in \mathscr{N}$*, we have* $\text{V.rss}(\mathbf{n}) \leq \text{V.rss}(\mathbf{m})$. *(b) There exists a sample allocation* $\mathbf{n} \notin \mathscr{N}$ *such as* $\text{V.rss}(\mathbf{n}) \leq \text{V.rss}(\mathbf{m})$.

Proof

(a) Without loss of generality, we assume $\sigma_1^2 \leq \sigma_2^2 \leq \cdots \leq \sigma_H^2$ and n_hs are positive real numbers.

We first prove the claim for the case $H = 2$. Let

$$f(n_1) = \frac{1}{H^2}\left\{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - \frac{2}{n}\big(\sigma_1^2 + \sigma_2^2\big)\right\} = \text{V.rss}(\mathbf{n}) - \text{V.rss}(\mathbf{m}), \tag{1}$$

where $n_2 = n - n_1$. The function is convex in n_1 and has zero values when

$$\big(n_1, n_2\big) = \left(\frac{n}{2}, \frac{n}{2}\right) \text{ or } \left(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}n, \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}n\right).$$

Thus, for every $n_1 \in \big(n\sigma_1^2/(\sigma_1^2 + \sigma_2^2), \frac{n}{2}\big)$ or equivalently $1 \leq n_2/n_1 \leq \sigma_2^2/\sigma_1^2$, we have $f\big(n_1\big) \leq 0$ that means V.rss($\mathbf{n}$) $\leq$ V.rss($\mathbf{m}$).

We now prove the claim for a general H. Let

$$f(\mathbf{n}) = \text{V.rss}(\mathbf{n}) - \text{V.rss}(\mathbf{m}) = \frac{1}{H^2}\sum_{h=1}^{H}\frac{\sigma_h^2}{n_h} - \frac{1}{H^2}\sum_{h=1}^{H}\frac{\sigma_h^2}{m} := \frac{1}{H^2}\sum_{h=1}^{H} f_h(n_h) \tag{2}$$

where

$$f_h(n_h) = \frac{\sigma_h^2}{n_h} - \frac{\sigma_h^2}{m}.$$

A simple algebra shows that $f(\mathbf{n})$ is a convex function of $n_1, n_2, \ldots, n_{H-1}$.

For $h = 1, 2, \ldots, H-1$, define the set

$$A_h = \left\{\big(n_h, n_{h+1}\big)\Big| 1 \leq \frac{n_{h+1}}{n_h} \leq \frac{\sigma_{h+1}^2}{\sigma_h^2}\right\} \tag{3}$$

and so $\mathcal{N} = \bigcap_{h=1}^{H-1} A_h$. We show that for every $\mathbf{n} = \big(n_1, n_2, \ldots, n_H\big) \in \mathcal{N}$, $f\big(\mathbf{n}\big) \leq 0$. Given $\mathbf{n} = \big(n_1, n_2, \ldots, n_H\big) \in \mathcal{N}$, we consider a sequence of allocations $\mathbf{m}^a$, $a = 0, 1, 2, \ldots$ which starts with $\mathbf{m}^0 = \mathbf{n}$ and converges to $\big(1/m, 1/m, \ldots, 1/m\big)$. For $a = q(H-1) + h$, $q = 0, 1, 2, \ldots$, we update $(m_h^{a-1}, m_{h+1}^{a-1})$ in $\mathbf{m}^{a-1} = \big(m_1^{a-1}, m_2^{a-1}, \ldots, m_H^{a-1}\big)$; the updated (m_h^a, m_{h+1}^a) is

$$\big(m_h^a, m_{h+1}^a\big) = \left(\frac{m_h^{a-1} + m_{h+1}^{a-1}}{2}, \frac{m_h^{a-1} + m_{h+1}^{a-1}}{2}\right) \in A_h$$

and let $m_j^a = m_j^{a-1}$ for $j \neq h, h+1$; and $\mathbf{m}^a = \big(m_1^a, m_2^a, \ldots, m_H^a\big) \in \mathcal{N}$. We know that

$$\lim_{a\to\infty} \mathbf{m}^a = \left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right)$$

with $m = n/H$.

Now, to show the claim for general H, it suffices to show that, for every $a \geq 1$,

$$f\big(\mathbf{m}^{a-1}\big) = \sum_{h=1}^{H} f_h(m_h^{a-1}) \leq f\big(\mathbf{m}^{a}\big) = \sum_{h=1}^{H} f_h(m_h^{a})$$

Without loss of generality, $a = q(H-1) + h$ and $\mathbf{m}^a$ updates the h-th and $(h+1)$-th elements of $\mathbf{m}^{a-1}$. For notational simplicity, let $m_h^{a-1} = n_h$ and $m_{h+1}^{a-1} = n_{h+1}$ and $m_h^a = m_{h+1}^a = \big(n_h + n_{h+1}\big)/2$. Then

$$f\big(\mathbf{m}^{a-1}\big) - f\big(\mathbf{m}^{a}\big) = \frac{\sigma_h^2}{n_h} + \frac{\sigma_{h+1}^2}{n_{h+1}} - \frac{2}{n_h + n_{h+1}}\Big(\sigma_h^2 + \sigma_{h+1}^2\Big) \leq 0, \quad (4)$$

by applying the case $H = 2$. Finally, we have $f\big(\mathbf{n}\big) \leq 0$ and so V.rss($\mathbf{n}$) $\leq$ V.rss($\mathbf{m}$).

(b) We start with the case of $H = 3$ and $\sigma_1^2 \leq \sigma_2^2 \leq \sigma_3^2$. Under this case, it suffices to find $\mathbf{n} \notin \mathscr{N}$ but V.rss($\mathbf{n}$) $<$ V.rss($\mathbf{m}$).

Suppose we consider $\mathbf{n} \notin \mathscr{N}$ which satisfies (C1) $1 \leq \sigma_3^2/\sigma_2^2 \leq n_3/n_2$, (C2) $n_1 \leq n_2 < m = n/H < n_3$,

$$c_1 \times \frac{\frac{1}{n_1} - \frac{1}{m}}{\frac{1}{m} - \frac{1}{n_3}} \leq \frac{\sigma_3^2}{\sigma_1^2} \quad \text{(C3)}$$

and

$$c_2 \times \frac{\frac{1}{n_2} - \frac{1}{m}}{\frac{1}{m} - \frac{1}{n_3}} \leq \frac{\sigma_3^2}{\sigma_2^2} \quad \text{(C4)}$$

for some positive values c_1 and c_2 such as $1/c_1 + 1/c_2 = 1$.

We again consider the function $f(\mathbf{n})$ = V.rss($\mathbf{n}$) − V.rss($\mathbf{m}$) = $\sum_{h=1}^{3} f_h(n_h)/H^2$ where $f_h(n_h) = \sigma_h^2/n_h - \sigma_h^2/m$ for $h = 1, 2, 3$. Then, by the conditions (C1)–(C4),

$$c_1 f_1(n_1) = c_1\sigma_1^2\left(\frac{1}{n_1} - \frac{1}{m}\right) \leq \sigma_3^2\left(\frac{1}{m} - \frac{1}{n_3}\right) = -f_3(n_3)$$

and

$$c_2 f_2(n_2) = c_2\sigma_2^2\left(\frac{1}{n_2} - \frac{1}{m}\right) \leq \sigma_3^2\left(\frac{1}{m} - \frac{1}{n_3}\right) = -f_3(n_3).$$

Since $m < n_3$ gives $-f_3(n_3) > 0$,

$$\frac{f_1(n_1) + f_2(n_2)}{-f_3(n_3)} \le \frac{1}{c_1} + \frac{1}{c_2} = 1$$

and thus $f_1(n_1) + f_2(n_2) \le -f_3(n_3)$. Finally, we have $\mathbf{n} \notin \mathscr{N}$ but satisfies $f(\mathbf{n}) = \text{V.rss}(\mathbf{n}) - \text{V.rss}(\mathbf{m}) \le 0$. □

3 Graphical Illustration of the Sample Allocation Set $\mathscr{N}$

Here, we fix the set size H at 3 and use a simplex diagram to illustrate the sample allocation set $\mathscr{N}$ proposed in Sect. 2.

First, we consider a hypothetical case with stratum variances $\sigma_1^2 = 1 < \sigma_2^2 = 2 < \sigma_3^2 = 3$, in which the sample allocation set $\mathscr{N}$ is plotted as a gray-shaded region in Fig. 1. In the simplex diagram, a point (x, y, z) with $x + y + z = 100$ implies the percentage of each stratum size relative to the total sample size. Three vertexes represent the allocation schemes having rates N1(100,0,0), N2(0,100,0), and N3(0,0,100), and the line $\overline{\text{N2N3}}$ ($\overline{\text{N1N3}}$ or $\overline{\text{N1N2}}$) becomes the baseline $x = 0$ ($y = 0$ or $z = 0$). A series of lines have been drawn in parallel to each baseline to mark off the percentages, and the percent scale for x (y or z) is laid out along the line $\overline{\text{N1N2}}$ ($\overline{\text{N2N3}}$ or $\overline{\text{N1N3}}$). Then, in Fig. 1, for the case with $\sigma_1^2 = 1 < \sigma_2^2 = 2 < \sigma_3^2 = 3$, the balanced allocation (33.3, 33.3, 33.3), the Neyman

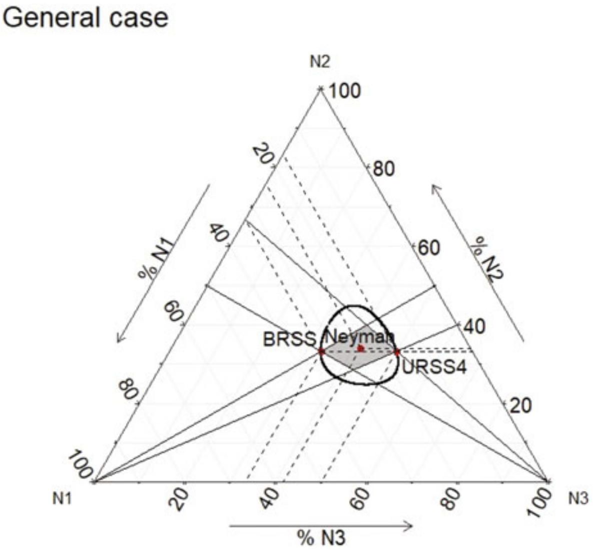


Fig. 1 An illustration of sample allocation schemes (BRSS, balanced allocation; Neyman, Neyman allocation; URSS4, an unequal allocation with $n_{h+1}/n_h = \sigma_{h+1}^2/\sigma_h^2$) for a hypothetical case with set size $H = 3$ and $\sigma_1^2 = 1 < \sigma_2^2 = 2 < \sigma_3^2 = 3$. The oval-like shape represents the set $\mathscr{N}_0$ and the gray-shaded area represents the set $\mathscr{N}$

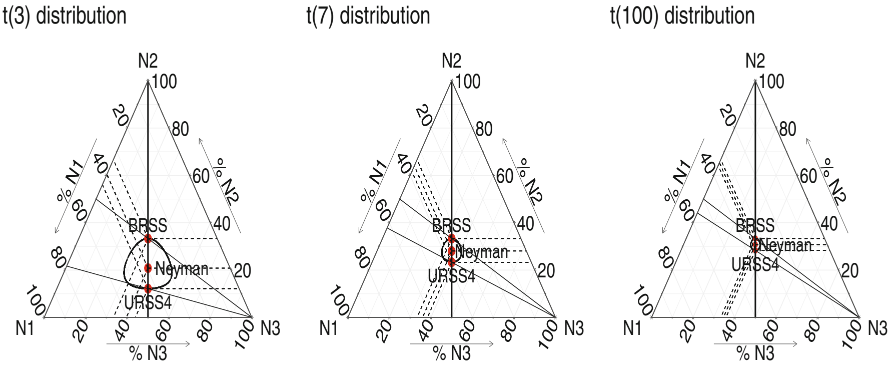


Fig. 2 An illustration of sample allocation schemes in the sets $\mathscr{N}$ and $\mathscr{N}_0$ (BRSS, balanced allocation; Neyman, Neyman allocation; URSS4, an unequal allocation with $n_{h+1}/n_h = \sigma^2_{h+1}/\sigma^2_h$) under t distributions with different degrees of freedom 3, 7, ∞. The oval-like shape represents the set $\mathscr{N}_0$. Due to the symmetry of t distributions, $n_1 = n_3$ so that the set $\mathscr{N}$ is reduced to the segment on the line that is perpendicular to $\overline{\text{N1N3}}$

Table 1 The stratum variances for the set size $H = 3$ under t and gamma distributions

df	t				Gamma		
	$\sigma^2_{[1]}$	$\sigma^2_{[2]}$	$\sigma^2_{[3]}$	α	$\sigma^2_{[1]}$	$\sigma^2_{[2]}$	$\sigma^2_{[3]}$
3	2.601	0.720	2.601	1	0.111	0.361	1.361
7	0.885	0.538	0.885	2	0.095	0.201	0.559
100	0.575	0.454	0.575	3	0.080	0.139	0.336

allocation (24.1, 34.1, 41.8), and the allocation (16.7, 33.3, 50) according to the rule $n_{h+1}/n_h = \sigma^2_{h+1}/\sigma^2_h$, are plotted as points labeled "BRSS," "Neyman," and "URSS4," respectively. Further, the oval-like area surrounded by the thick curved line is the efficient sample allocation set $\mathscr{N}_0$ that contains $\mathscr{N}$.

Secondly, we consider symmetric distributions and in Fig. 2, we show the sample allocation sets under t-distributions with different degrees of freedom $df = 3,7,100$, whose stratum variances are given in Table 1. Note that normal distributions are a special case of t-distributions with $df = \infty$. Due to the symmetric property that yields $\sigma^2_{[1]} = \sigma^2_{[3]}$ and so $n_1 = n_3$, the set $\mathscr{N}$ is reduced to a segment on the line that is perpendicular to $\overline{\text{N1N3}}$ within the set $\mathscr{N}_0$. As df increases, the ratio of variances, $\sigma^2_{[2]}/\sigma^2_{[1]} = \sigma^2_{[2]}/\sigma^2_{[3]}$, decreases and the set $\mathscr{N}$ becomes smaller along the line, approaching the point of BRSS.

Thirdly, we consider asymmetric distributions and show the sample allocation sets under gamma distributions in Fig. 3. For these gamma distributions, shape and rate parameters are both set to α, and thus they have mean 1, variance $1/\alpha$, and skewness $2/\sqrt{\alpha}$. We set $\alpha = 1, 2, 3$, and report the corresponding stratum variances in Table 1. Compared to symmetric distributions, there is no equality in stratum variances, and the sample allocation set $\mathscr{N}$ is a polygon with four edges. Also, as α increases, the skewness decreases so that the variances become more homogeneous. This makes $\mathscr{N}$ become smaller.

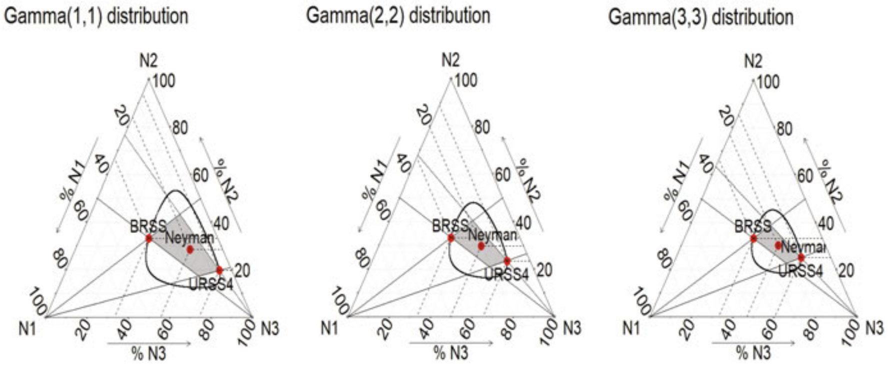


Fig. 3 An illustration of sample allocation schemes in the sets $\mathscr{N}$ and and $\mathscr{N}_0$ (BRSS, balanced allocation; Neyman, Neyman allocation; URSS4, an unequal allocation with $n_{h+1}/n_h = \sigma^2_{h+1}/\sigma^2_h$) for gamma distributions with different $\alpha = 1, 2, 3$ with mean 1. The oval-like shape represents the set $\mathscr{N}_0$ and the gray-shaded area represents the set $\mathscr{N}$

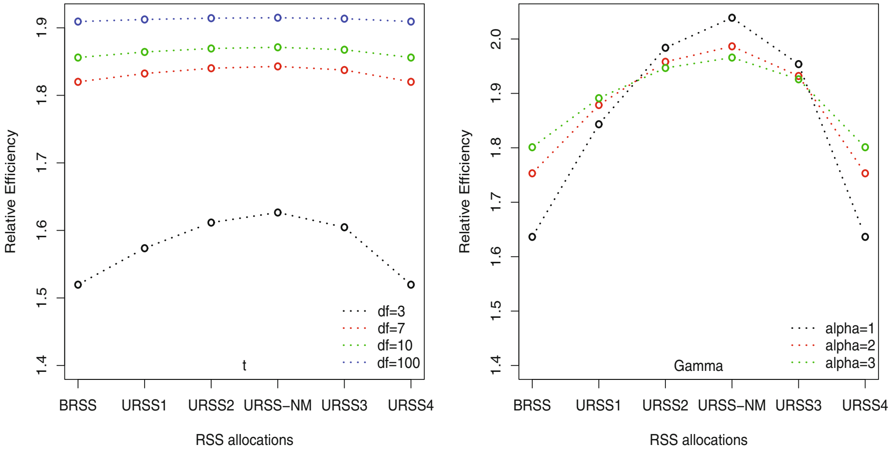


Fig. 4 Comparison of theoretical RE of BRSS to those of various URSS schemes (URSS-NM, the Neyman allocation; URSS1, 1/3 NM and 2/3 BRSS; URSS2, 2/3 NM and 1/3 BRSS; URSS3, 1/2 NM and 1/2 URSS4; URSS4, the scheme with $n_{h+1}/n_h = \sigma^2_{h+1}/\sigma^2_h$) under perfect ranking

Figure 4 shows theoretical RE values of RSS mean estimators based on six allocation schemes in the set $\mathscr{N}$, including BRSS (equal allocation), URSS-NM (the Neyman allocation), and four other unequal allocation schemes: URSS1 (1/3 NM and 2/3 BRSS), URSS2 (2/3 NM and 1/3 BRSS), URSS3 (1/2 NM and 1/2 BRSS), and URSS4 with $n_{h+1}/n_h = \sigma^2_{h+1}/\sigma^2_h$. The RE of the RSS mean estimator $\widehat{\mu}_{\text{RSS}}$ with the sample allocation $\mathbf{n}$ over the SRS mean estimator $\widehat{\mu}_{\text{SRS}}$ with the same sample size $n = \sum_{h=1}^{H} n_h$ is defined as the ratio of variances:

$$RE = \frac{\operatorname{Var}(\hat{\mu}_{\text{SRS}})}{\text{V.rss}(\mathbf{n})} = \frac{\sigma^2/n}{\frac{1}{H^2}\sum_{h=1}^{H}\sigma_h^2/n_h}.$$

All allocation schemes use theoretical allocation proportions to compute their REs (i.e., non-integer sample sizes are allowed). Clearly, the theoretical REs of all five unequal allocation schemes in the set $\mathscr{N}$ are larger than that of the BRSS counterpart. We mention that for the t distributions, as df increases, the heavy-tailedness becomes less severe and the URSS allocation schemes get closer to the balanced BRSS allocation, and so the lines are quite flat.

As shown in Theorem 1-(b), the set $\mathscr{N}$ is nested in the set $\mathscr{N}_0$. Ideally, one may want to figure out a sufficient and necessary condition for allocation schemes in $\mathscr{N}_0$. As $\mathscr{N}$ outlines a sufficient condition only, we are interested in comparing the coverage of $\mathscr{N}$ relative to $\mathscr{N}_0$. In Table 2, we report the ratio of the area of $\mathscr{N}$ to that of $\mathscr{N}_0$, $|\mathscr{N}|/|\mathscr{N}_0|$, as the relative probability of points lying inside the inscribed set $\mathscr{N}$ over the set $\mathscr{N}_0$ for the t and gamma distributions. We denote the area ratios as AR and AR* by considering all the real valued points and only the integer valued points s.t. $n = \sum_{h=1}^{H} n_h$, respectively. Note that for the t distributions, as $\mathscr{N}$ is only a segment, AR is zero in theory. To compute AR for the gamma distributions, we randomly generate 10,000 allocation schemes using Monte Carlo simulation and count how many in $\mathscr{N}$ and $\mathscr{N}_0$. We find that the relative size of $\mathscr{N}$ to $\mathscr{N}_0$ (the ratio AR*) is large when the sample size n is small, in which RSS has been proved to be most useful. In addition, the ratio AR* decreases to the ratio AR, as n increases.

4 Sample Allocation Adjustment

4.1 Local Ratio Consistent and Approximate Neyman Allocations

Based on the sufficient set $\mathscr{N}$ characterized in Sect. 2, for $\mathbf{n} = (n_1, n_2, \ldots, n_H) \notin \mathscr{N}$, we propose the so-called local ratio consistent (LRC) allocation $\mathbf{n}^{LRC}$ to move $\mathbf{n}$ into $\mathscr{N}$ by local adjustment (i.e., adding a few samples to some of the strata). For the purpose of comparison, we also consider a naive adjustment method that leads to approximate Neyman (AN) allocation $\mathbf{n}^{AN}$.

The first adjustment procedure that yields $\mathbf{n}^{LRC}$ attains the local ratio consistency via the following steps. Again, we relabel the strata to satisfy $\sigma_1^2 \le \sigma_2^2 \le \cdots \le \sigma_H^2$.

1. For the current allocation $\mathbf{n} = (n_1, n_2, \ldots, n_H)$, we define, for $h = 1, 2, \ldots, H-1$,

$$u_h = \frac{n_{h+1}}{n_h} \cdot \frac{\sigma_h^2}{\sigma_{h+1}^2} \quad \text{and} \quad \ell_h = \frac{n_{h+1}}{n_h}.$$

Table 2 The ratio of areas $|\mathscr{N}|/|\mathscr{N}_0|$ for the set size $H = 3$. The numbers in parenthesis are "$|\mathscr{N}|$/all" and "$|\mathscr{N}_0|$/all," where "all" is the number of all possible integer allocation schemes

	AR		AR*			
Distribution	α		$n = 6$	$n = 12$	$n = 24$	$n = 48$
t	3	0	0.333(1/10,3/10)	0.333(2/55,6/55)	0.130(3/253,23/253)	0.068(6/1081,88/1081)
	7	0	1(1/10,1/10)	1(1/55,1/55)	0.5(2/253,4/253)	0.176(3/1081,17/1081)
	∞	0	1(1/10,1/10)	1(1/55,1/55)	1(1/253,1/253)	0.500(2/1081,4/1081)
Gamma	1	0.355	0.400(2/10,5/10)	0.412(7/55,17/55)	0.431(25/253,58/253)	0.403(94/1081,233/1081)
	2	0.343	1(2/10,2/10)	0.556(5/55,9/55)	0.395(15/253,38/253)	0.368(56/1081,152/1081)
	3	0.323	0.500(1/10,2/10)	0.500(4/55,8/55)	0.423(11/253,26/253)	0.361(39/1081,108/1081)

and compute these quantities.

2. Let $h^\star = \operatorname{argmax}_{h=1}^{H-1} u_h$ and if $u_{h^\star} > 1$, add one to $n_{h^\star}$, $n_{h^\star} \leftarrow n_{h^\star} + 1$.
3. Let $h_\star = \operatorname{argmin}_{h=1}^{H-1} \ell_h$ and if $\ell_{h_\star} < 1$, add one to $n_{h_\star+1}$, $n_{h_\star+1} \leftarrow n_{h_\star+1} + 1$.
4. If $u_h \leq 1 \leq \ell_h$ for $h = 1, 2, \ldots, H-1$, stop the procedure and report the current allocation $\mathbf{n}$. Otherwise, we iterate steps 1–3.

We remark that $h^\star$ and $h_\star$ can not be equal to each other, because for every $h = 1, 2, \ldots, H-1$, if $u_h > 1$, then $\ell_h > 1$; and, similarly, if $\ell_h < 1$, then $u_h < 1$.

The second adjusted allocation $\mathbf{n}^{AN}$ is based on the Neyman allocation. Let $\mathbf{n}^N = \left(n_1^N, n_2^N, \ldots, n_H^N\right)$ denote the Neyman allocation for a fixed total sample size $n = \sum_{h=1}^H n_h$ and a set size H as in the original allocation $\mathbf{n}$. We then simply define the AN allocation by

$$\begin{aligned}\mathbf{n}^{AN} &= (n_1^{AN}, n_2^{AN}, \cdots, n_H^{AN}) = \left(\max(n_1^N, n_1), \max(n_2^N, n_2), \cdots, \max(n_H^N, n_H)\right)\\ &:= \max\left(\mathbf{n}^N, \mathbf{n}\right),\end{aligned}$$

where $n_h^{AN} = \max(n_h^N, n_h)$ with the total sample size $n^{AN} = \sum_{h=1}^H n_h^{AN}$. Then we have additional samples at the h-th stratum $n_{h+}^N = \max(0, n_h^N - n_h)$ and additional total samples $n_+^N = \sum_{h=1}^H n_{h+}^N$. Unlike $\mathbf{n}^{LRC}$, there is no guarantee that $\mathbf{n}^{AN}$ is in $\mathscr{N}$ or $\mathscr{N}_0$. However, due to more samples used and proximity to the Neyman allocation, $\mathbf{n}^{AN}$ is very likely to be more efficient than the initial $\mathbf{n}$.

4.2 Comparison via Simulation

We generate 10,000 sample allocations from the multinomial distribution with parameter $p = (1/6, 1/3, 1/2)$ with a size of $n = 12, 24, 48$. To compare the two adjustment methods, we compute the proportion of the allocation schemes updated from the original schemes, the average number of additional samples, the average RE, and the average efficiency gain (EG) per an additional sample over 10,000 replicates for each setting considered in Table 3. Note that the average number of additional samples, RE, and EG are computed for the cases where at least one stratum of LRC allocation is updated (i.e., $\mathbf{n}^{LRC}$ has at least one additional sample). The EG of $\mathbf{n}^A$ versus $\mathbf{n}$ per an additional sample is defined as

$$\mathrm{EG}(\mathbf{n}^A) = \frac{\mathrm{RE}(\mathbf{n}^A) - \mathrm{RE}(\mathbf{n})}{n_+^A}$$

where the superscript $A \in \{`LRC', `AN'\}$ denotes the adjustment method of sample allocation and n_+^A is the number of the required additional total samples by the corresponding method A.

We assume that the data are generated from gamma distributions with equal shape and rate parameters $\alpha \in \{1, 2, 3\}$ yielding the variance $1/\alpha$. Accordingly,

Table 3 Comparison between the original and two adjusted allocation schemes: $H = 3$, $p = (1/6, 1/3, 1/2)$, and 10,000 replicates. The numbers in the parentheses represent standard deviation

α	n	adj.	%($\mathbf{n} \notin \mathscr{N}$)	AVG(n_+^A)	AVG(RE)	AVG(EG)
1	12	$\mathbf{n}^{LRC}$	57.61%	1.999 (1.522)	1.887 (0.105)	0.128 (0.064)
		$\mathbf{n}^{AN}$		2.434 (1.159)	1.938 (0.074)	0.126 (0.060)
	24	$\mathbf{n}^{LRC}$	44.09%	2.444 (1.851)	1.902 (0.079)	0.084 (0.066)
		$\mathbf{n}^{AN}$		3.855 (1.404)	1.959 (0.057)	0.065 (0.040)
	48	$\mathbf{n}^{LRC}$	25.66%	2.782 (2.344)	1.913 (0.058)	0.039 (0.028)
		$\mathbf{n}^{AN}$		6.518 (2.161)	1.974 (0.040)	0.026 (0.016)
2	12	$\mathbf{n}^{LRC}$	70.81%	2.305 (1.677)	1.893 (0.049)	0.140 (0.088)
		$\mathbf{n}^{AN}$		2.069 (0.907)	1.893 (0.061)	0.147 (0.081)
	24	$\mathbf{n}^{LRC}$	66.37%	2.811 (2.203)	1.913 (0.047)	0.074 (0.052)
		$\mathbf{n}^{AN}$		3.415 (1.263)	1.940 (0.038)	0.066 (0.044)
	48	$\mathbf{n}^{LRC}$	55.84%	3.305 (2.492)	1.927 (0.039)	0.035 (0.021)
		$\mathbf{n}^{AN}$		5.433 (1.852)	1.957 (0.025)	0.027 (0.017)
3	12	$\mathbf{n}^{LRC}$	78.86%	2.756 (1.840)	1.917 (0.038)	0.115 (0.064)
		$\mathbf{n}^{AN}$		2.134 (0.950)	1.898 (0.067)	0.127 (0.064)
	24	$\mathbf{n}^{LRC}$	78.81%	3.418 (2.305)	1.919 (0.030)	0.065 (0.044)
		$\mathbf{n}^{AN}$		3.585 (1.318)	1.926 (0.036)	0.061 (0.041)
	48	$\mathbf{n}^{LRC}$	73.65%	4.242 (3.012)	1.921 (0.031)	0.032 (0.020)
		$\mathbf{n}^{AN}$		5.448 (1.939)	1.939 (0.022)	0.028 (0.018)

the stratum variances are $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = \{(0.11, 0.36, 1.36), (0.10, 0.20, 0.56), (0.08, 0.14, 0.34)\}$. For RSS schemes, we consider $H = 3$, $n = \{12, 24, 48\}$ for the original sample allocation, all with perfect ranking.

Table 3 shows that as α increases (so that the skewness decreases), the initial sample allocations $\mathbf{n}$ is more frequently not in $\mathscr{N}$, and this is because the (relative) area of the set $\mathscr{N}$ decreases as shown in Fig. 3. We also find that, if the initial allocation is not in $\mathscr{N}$ and is adjusted to $\mathbf{n}^{LRC}$ and $\mathbf{n}^{AN}$, then $\mathbf{n}_+^{LRC}$ (the number of added samples by LRC) tends to be smaller than $\mathbf{n}_+^{AN}$ (the number of added samples by AN). This is further confirmed by Table 4, which reports distributions of the number of additional samples by LRC and AN and that of their difference for the case with $n = 12$ and $\alpha = 1$. For this reason, in Table 3, the average RE of $\mathbf{n}^{LRC}$ is smaller than that of $\mathbf{n}^{AN}$. Nevertheless, the EG per one additional sample of $\mathbf{n}^{LRC}$ is larger than that of $\mathbf{n}^{AN}$ in the cases. To sum up, these results show that $\mathbf{n}^{LRC}$ tends to require fewer additional samples than $\mathbf{n}^{AN}$ to make the design more efficient and is cost-effective in the sense that EG per one additional sample is larger than $\mathbf{n}^{AN}$.

Table 4 Distributions of no. of additional samples using LRC and AN and the distribution of the difference in total sample size between LRC and AN with $H = 3$, $n = 12$, $p = (1/6, 1/3, 1/2)$, and the gamma distribution with $\alpha = 1$

%	n_+^{LRC}	n_+^{AN}	$n^{AN} - n^{LRC}$
0(<0)	42.39	7.65	28.74 (7.12)
1	30.33	33.84	43.73
2	14.77	30.56	17.96
3	5.41	17.31	2.45
4+	7.10	10.64	–

5 Data Example

In practice, BRSS is often the default RSS design and is frequently used due to its simplicity in implementation and inference. However, the issue of missing data is common in many studies and results in unbalancedness in RSS. For example, in the following education study, there are various reasons that students or schools drop out of assessment tests.

In this section, following Wang et al. (2017), we simulate a realistic situation, where study designs are embedded with BRSS when recruiting experimental units but the collected data have a URSS design at the end of the experiments due to missingness. We use a dataset from the High School Longitudinal Study of 2009 (HSLS09) and preprocess it as in Wang et al. (2017) to examine the performance of the two local adjustment procedures in Sect. 4. The HSLS09 data contain results of the students' assessments throughout secondary and postsecondary years from the National Center for the Education Statistics (NCES) website. The 12,533 students involved are considered as the population, and the 2012 math theta scores (X2TXMTH) are thought of as the response variable. The 2012 math theta scores (X2TXMTH) and the 2009 math theta scores (X1TXMTH) are used to rank students in RSS, for perfect and imperfect ranking, respectively. The correlation between X1TXMTH and X2TXMTH is about 0.78.

Suppose we aim to estimate the mean score to evaluate high school students' math ability. We consider the experiment design with $H = 3$, $n_h = m \in \{4, 8, 12\}$ for $h = 1, 2, 3$ and so $n \in \{12, 24, 48\}$, but we assume that outcomes are missing completely at random with missing rate $\phi \in \{10\%, 20\%\}$. We treat the resulting sample allocation from data with missing values as the "original" URSS allocation. Since for real data, the underlying distribution of the response variable is not known, we estimate the stratum variances σ_h^2's using the corresponding sample variances based on data from the "original" allocation.

For the given URSS allocation $\mathbf{n}$, we compute the integer valued Neyman allocation $\mathbf{n}^N$ using Wright (2012) and apply the two local adjustment methods to obtain $\mathbf{n}^{LRC}$ and $\mathbf{n}^{AN}$. We repeat the procedure 10,000 times and compute the performance measures introduced in Sect. 4.2. Table 5 reports the results, including the proportion of the updated allocation schemes, the average number of additional

Table 5 HSLS09 data example: comparing performance of two adjustment methods that yield $\mathbf{n}^{LRC}$ and $\mathbf{n}^{AN}$ in percentage of the number of additional samples and relative efficiency. The numbers in parentheses are standard deviations. The stratum variances σ_h^2's are estimated using the corresponding sample variances based on the "original" allocation (i.e., URSS allocation caused by missing data). The RE is the ratio of the empirical mean squared errors of 10,000 RSS estimates over that of 10,000 SRS estimates with the same sample size

			Perfect ranking			Imperfect ranking		
ϕ	n	adj.	$\%(\mathbf{n} \notin \mathscr{N})$	AVG(n_+^A)	RE	$\%(\mathbf{n} \notin \mathscr{N})$	AVG(n_+^A)	RE
0.1	12	$\mathbf{n}^{LRC}$	67.59 %	1.263 (0.714)	1.853	68.37 %	1.244 (0.688)	1.437
		$\mathbf{n}^{AN}$		1.794 (0.919)	1.792		1.831 (0.929)	1.392
	24	$\mathbf{n}^{LRC}$	74.15 %	1.647 (0.478)	1.920	73.34 %	1.645 (0.479)	1.455
		$\mathbf{n}^{AN}$		2.348 (1.187)	1.860		2.390 (1.208)	1.438
	48	$\mathbf{n}^{LRC}$	79.41 %	2.404 (1.338)	1.950	78.30 %	3.298 (1.331)	1.499
		$\mathbf{n}^{AN}$		3.169 (1.615)	1.901		3.225 (1.589)	1.414
0.2	12	$\mathbf{n}^{LRC}$	70.79 %	1.573 (0.495)	1.902	70.50 %	1.589 (0.492)	1.506
		$\mathbf{n}^{AN}$		1.853 (0.914)	1.831		1.864 (0.916)	1.422
	24	$\mathbf{n}^{LRC}$	79.30 %	2.292 (1.255)	1.937	79.00 %	2.282 (1.249)	1.456
		$\mathbf{n}^{AN}$		2.401 (1.184)	1.871		2.416 (1.199)	1.418
	48	$\mathbf{n}^{LRC}$	82.58 %	3.170 (2.028)	1.946	82.14 %	3.161 (2.022)	1.468
		$\mathbf{n}^{AN}$		3.256 (1.595)	1.943		3.305 (1.611)	1.414

samples, and empirical RE, for both perfect and imperfect ranking. Here, the RE is the ratio of the empirical mean squared errors of 10,000 RSS estimates over that of 10,000 SRS estimates with the same sample size. Again, the measures are calculated for the cases when $\mathbf{n}^{LRC}$ has additional samples to the original $\mathbf{n}$. Table 5 reassures our finding in Sect. 4.2 that on average, n_+^{LRC} is smaller than n_+^{AN} and the efficiency gain by one additional sample in $\mathbf{n}^{LRC}$ is larger than that of $\mathbf{n}^{AN}$ in all designs considered (since RE for $\mathbf{n}^{LRC}$ is already higher than that for $\mathbf{n}^{AN}$ even with fewer additional samples). These findings are true for both perfect and imperfect ranking, even when the stratum variances are unknown and have to be estimated from the data.

6 Conclusion

We conclude the paper with a brief summary. We consider a set $\mathscr{N}$ of sample allocation schemes for unbalanced ranked set sampling (URSS), which is a subset of $\mathscr{N}_0$, the collection of all allocation schemes giving more efficient mean estimation than their BRSS counterparts. The set $\mathscr{N}$ is characterized by local conditions on the sample sizes of adjacent strata, and this allows us to move a less efficient URSS allocation scheme $\mathbf{n}$ into $\mathscr{N}$ by adding a few samples into a few strata. We illustrate the set $\mathscr{N}$ with $H = 3$ using a simplex diagram for various underlying distributions. We further consider two procedures to adjust $\mathbf{n}$, which yields the local ratio consistent (LRC) allocation $\mathbf{n}^{LRC}$ and approximate Neyman (AN) allocation $\mathbf{n}^{AN}$, respectively. We numerically compare the two methods via simulation and find that $\mathbf{n}^{LRC}$, which locally adjusts $\mathbf{n}$ based on $\mathscr{N}$, tends to require fewer extra samples and have higher efficient gain per sample than $\mathbf{n}^{AN}$. Our data example using the High School Longitudinal Study of 2009 (HSLS09) confirms the finding from the simulation, in which stratum variances have to be estimated from data. It also illustrates the usefulness of the LRC method in situations when BRSS is initially planned, but missing data causes a URSS scheme that needs to be adjusted.

Our discussion in this study focuses on the set size $H = 3$. For large H, the cases with more number of rank strata, we expect that the efficient gain of LRC over AN becomes large. It is because the AN allocation globally depends on the allocations of many other strata and tends to require more additional samples to make the new unbalanced design more efficient than the BRSS compared to the LRC allocation.

Acknowledgments This work was supported by the National Research Foundation of Korea (Grant No.: NRF-2017R1D1A1B03032073).

References

Ahn, S., Wang, X., & Lim, J. (2017). On unbalanced group sizes in cluster randomized designs using balanced ranked set sampling. *Statistics & Probability Letters, 123*, 210–217.

Bhoj, D., & Chandra, G. (2019). Simple unequal allocation procedure for ranked set sampling with skew distributions. *Journal of Modern Applied Statistical Methods, 18*(2), eP2811.

Bocci, C., Petrucci, A., & Rocco, E. (2010). Ranked set sampling allocation models for multiple skewed variables: An application to agricultural data. *Environmental and Ecological Statistics, 17*(3), 333–345.

Chen, Z., & Bai, Z. (2000). The optimal ranked-set sampling scheme for parametric familites. *Sankhya: The Indian Journal of Statistics, Series A, 62*(2), 178–192.

Chen, H., Stasny, E. A., & Wolfe, D. A. (2006). Unbalanced ranked set sampling for estimating a population proportion. *Biometrics, 62*(1), 150–158.

McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research, 3*(4), 385–390.

Ozturk, O., & Wolfe, D. A. (2004). Optimal allocation procedure in ranked set sampling for unimodal and multi-modal distributions. *Environmental and Ecological Statistics, 7*, 343–356.

Stokes, S. L., & Sager, T. W. (1988). Ocharacterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association, 83*(402), 374–381.

Takahasi, K., & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics, 20*, 1–31.

Wang, Y. G., Chen, Z., & Liu, J. (2004). General ranked set sampling with cost considerations. *Biometrics, 60*(2), 556–561.

Wang, X., Ahn, S., & Lim, J. (2017). Unbalanced ranked set sampling in cluster randomized studies. *Journal of Statistical Planning and Inference, 187*, 1–16.

Wright, T. (2012). The equivalence of Neyman optimum allocation for sampling and equal proportions for apportioning the U.S. house of representatives. *The American Statistician, 66*(4), 217–224.

On the Versatility of Capture-Recapture Modeling: Counting What We Don't See

James D. Nichols

Abstract Initial development of capture-recapture modeling occurred almost exclusively within the disciplines of wildlife management and animal ecology. Virtually all methods for surveying animals "miss" individuals; i.e., some unknown fraction of animals present in surveyed areas goes undetected. In order to draw inferences about all animals actually present, we must deal with this nondetection. In addition, we sometimes misclassify animals as to species, sex, reproductive condition, etc., requiring us to deal with probabilities of misclassification. Capture-recapture models differ from many other kinds of statistical models in that they incorporate parameters that deal with both the process being studied (e.g., population size, survival rate, recruitment rate) and the sampling process giving rise to the data (e.g., capture or detection probability, correct classification probability). Many other disciplines face these same kinds of counting errors, nondetection and misclassification. These disciplines include epidemiology, medicine, social sciences, paleobiology, remote sensing, military imaging, philately, space exploration, quality control, and software development. This chapter includes a brief history of capture-recapture modeling, an introduction to the logic underlying basic models, a discussion of nontraditional uses of these models, and recommendations for additional potential uses.

1 Introduction

Wildlife biologists and animal ecologists realized early on that their methods for surveying animal populations did not provide accurate counts. Animals are missed by virtually all survey methods, and biologists were forced to develop methods that produced not only counts of animals detected but also estimates of those present, but not detected. Several clever approaches have been developed to deal with this

J. D. Nichols (✉)
Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, FL, USA

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_4

issue of nondetection (Seber, 1982; Williams et al., 2002; Seber & Schofield, 2019), and they have seen wide use in ecology and wildlife biology for decades. Another problem faced when surveying animal populations and communities is misclassification. Some of the same factors that cause animals to be difficult to detect (e.g., reliance on partial sightings or auditory cues) can also result in misclassification. For example, an auditory bird survey may result in species misidentification. A visual survey of a sexually monomorphic bird species may result in sex being misclassified. A sighting survey of reproduction in manatees may result in a reproductive female being misclassified as nonreproductive because her young is too far away from her or obscured by her body. Methods for dealing with nondetection have thus been modified to incorporate misclassification as well (Pradel, 2005; MacKenzie et al., 2018).

These same problems in counting and classifying characterize other scientific disciplines as well, but with much less corresponding effort to deal with them. Here, I first provide a brief history of both capture-recapture and closely related occupancy modeling. Then, I survey uses of these models in disciplines other than ecology and wildlife biology and identify opportunities for even greater use.

2 Capture-Recapture

In this section, I provide a brief introduction to capture-recapture modeling by describing the basic ideas underlying several classes of capture-recapture models. Data are typically summarized as capture histories depicting whether or not an individual was captured or detected at each sampling occasion of the study. Rather than developing the full likelihoods for these different classes of model, I define parameters and then write out probability structures for sample capture histories as an abbreviated way of explaining the thinking that underlies these models.

2.1 2-Sample, Closed Population, Single State

The most basic capture-recapture estimator is based on the recognition that the proportion of a specific type of individual or entity in a representative sample from a population should be roughly equal to that in the population itself. Let M be the known number of animals of a certain type in a total population of N individuals. Then define m as the number of animals of that type in a sample of n individuals. If the sample is representative of the population, in the sense of a similar proportion of marked animals, then we expect:

$$\frac{m}{n} \approx \frac{M}{N}. \tag{1}$$

We can rearrange (1) to obtain the following estimator for N:

$$\hat{N} = \frac{nM}{m}. \tag{2}$$

Expression (2) is known as the Lincoln-Petersen estimator, and the prototypical study to which it applies entails two sampling occasions separated by a short time interval over which the sampled population is assumed to be "closed," with no animals entering or departing. On occasion 1, M animals are captured and marks are applied to them. On occasion 2, another sample of n animals is captured, m of which are found to be marked. This estimator has been independently derived a number of times, initially by Laplace (1786), who used it to estimate the human population of France, and later by Lincoln (1930) to estimate the number of waterfowl in late summer to autumn in North America.

The estimator in (2) can also be viewed as a precursor to the general Horvitz and Thompson (1952) estimator for a population total. M is the number of animals sampled on occasion 1, and m/n estimates p_1, the probability that a member of the population of size N is caught in occasion 1:

$$\hat{N} = \frac{M}{\hat{p}_1}. \tag{3}$$

The estimator of (2) and (3) and its associated variance have been derived using both hypergeometric (fixed sample size) and binomial (random sampling) likelihoods (reviewed by Seber, 1982, Williams et al., 2002, Seber & Schofield, 2019).

The statistics used in capture-recapture modeling are most frequently written as the number of animals exhibiting each possible capture or detection history. For a 2-sample study, there are only three such statistics denoted as x_{ij}, where $i = 1$ if caught on occasion 1 and 0 if not caught then, and $j = 1$ if caught on occasion 2, and 0 if not.

$x_{11} = m =$ number of animals caught on occasions 1 and 2,
$x_{10} = M - m =$ number of animals caught on occasion 1 but not on occasion 2,
$x_{01} = n - m =$ number of animals caught on occasion 2 but not on occasion 1.

The number of animals not captured at either sampling occasion, x_{00}, is unknown, and the problem of estimating total abundance, N, is equivalent to the problem of estimating x_{00}.

Given that an animal is a member of the sampled population, the probabilities of it exhibiting each observable capture history are then:

$$\begin{aligned} \Pr(11) &= p_1 p_2, \\ \Pr(10) &= p_1(1 - p_2), \\ \Pr(01) &= (1 - p_1)p_2, \end{aligned}$$

where p_t is the capture probability for sampling occasion t. Under this 2-sample model, N, p_1, and p_2 can all be estimated. It is termed a "single state" model because all animals are assumed to have the same probabilities of appearing in a sample (i.e., no stratification by age, sex, size, etc.)

2.2 >2-Sample, Closed Population, Single State

This approach was extended to multiple sampling occasions (Schnabel, 1938; Darroch, 1958) for closed populations. Capture histories were modeled as in the 2-sample case. For a study with five sampling occasions, the probability of an animal in the sampled population showing a capture history of 01010 is:

$$\Pr(01010) = (1 - p_1)p_2(1 - p_3)p_4(1 - p_5).$$

Each possible capture history has an associated probability such as above, and we know how many animals exhibited the history, so we can develop a corresponding likelihood and estimate the capture probabilities and abundance. Subsequent developments included consideration of behavioral response of animals to initial capture, heterogenous capture probabilities, and other generalizations (Otis et al., 1978; Chao and Huggins, 2005a 2005b; Seber & Schofield, 2019).

2.3 >2 Samples, Open Populations, Single State

Capture-recapture methods were extended to "open" populations as well, where sampling occasions could be separated by long time intervals such that gains and losses to the population could occur between occasions (e.g., Jackson, 1933, 1939; Cormack, 1964; Jolly, 1965; Seber, 1965). These models were also based on multinomial likelihoods and required additional parameters for survival of an animal from one sampling occasion to the next. For example, let p_t denote capture probability for sampling occasion t, and let ϕ_t denote the probability that an animal alive at sampling occasion t survives until occasion $t+1$ and remains in the sampled population. The conditional probability associated with capture history x_{01010} in a 5-occasion study is:

$$\Pr(01010|\text{release in } 2) = \phi_2(1 - p_3)\phi_3 p_4(1 - \phi_4 p_5). \qquad (4)$$

The last terms in parentheses include both the possibility that the animal survived until 5 but was not caught and the possibility that the animal did not survive.

Likelihoods conditional on new releases in each sampling period can be used to estimate capture probabilities and survival probabilities and, assuming that animals that were and were not previously captured exhibit the same capture probabilities, occasion-specific abundance (see expression 3). Subsequent parameterizations include different ways of modeling the entry of new animals into the sampled population (e.g., Crosbie & Manly, 1981; Pradel, 1996; Schwarz & Arnason, 1996), as opposed to simply conditioning on entries as in (4).

2.4 >2 Samples, Open Populations, Multiple States

Arnason (1972, 1973) introduced the concept of multiple states in which an animal could be captured, where states initially represented different locations and were later generalized to characteristics of individual animals such as age, reproductive condition, body mass, etc. The first multistate models to be widely used allowed capture and survival parameters to depend on age, for studies in which sampling occasions were separated by time intervals that corresponded to the exact interval required for an animal to make the transition from 1 age class to the next (Manly & Parr, 1968; Pollock, 1981; Stokes, 1984). These age-specific models are much simpler than the general models of Arnason (1972, 1973) because of the deterministic, unidirectional nature of age transitions.

In the general multistate models of Arnason (1972, 1973), state transitions are stochastic, necessitating additional new parameters for transitions between states (also see Hestbeck et al., 1991; Brownie et al., 1993; Schwarz et al., 1993). Define θ_t^{rs} as the probability that an animal in state r at occasion t that survives until occasion $t + 1$ is in state s at $t + 1$. Define S_t^r as the probability that an animal in state r at sampling occasion t is still alive and in the sampled population at occasion $t + 1$, and p_t^r as the probability that an animal in state r at occasion t is captured at t. Capture histories must now indicate the state of the animal at each capture. In a study area with two locations, 1 and 2, a capture history of 0102 would indicate an animal first captured in state/location 1 at sampling occasion 2, not captured at occasion 3, and captured in state 2 at occasion 4. The number of animals showing this history is denoted as x_{0102}, and the probability that an animal released in occasion 2, state 1, will exhibit this history and thus appear in this statistic is:

$$\begin{aligned}&\Pr(0102|\text{release in state 1 at occasion 2})\\&= S_2^1[(1-\theta_2^{12})(1-p_3^1)S_3^1\theta_3^{12} + \theta_2^{12}(1-p_3^2)S_3^2(1-\theta_3^{21})]p_4^2.\end{aligned} \tag{5}$$

The portion of expression (5) in brackets reflects the state uncertainty of the animal at occasion 3 and can be viewed as a mixture model incorporating the possibilities that the animal was in state 1 or state 2. Likelihoods are conditional on new releases in each state in each sampling occasion.

These multistate models assume the ability to classify an animal to its appropriate state without error at each capture, and they have been generalized to deal with state uncertainty and misclassification (Kendall et al., 2003, 2004; Nichols et al., 2004; Pradel, 2005). These generalizations include additional classification parameters and sometimes use ancillary data to reduce uncertainty in modeling the capture and classification processes.

2.5 Occupancy Models, Closed System, Single State

Occupancy models extend the thinking underlying capture-recapture from individual animals to a set of locations or sites. The question for a single site is whether a focal species is present or not, and the objective of the modeling is to estimate occupancy, the probability that a site is occupied by the focal species. The motivation for these models is possible nondetection; surveys of sites sometimes "miss" detecting a species, despite presence of the species at the site. A key distinction between the occupancy problem for sites and the capture-recapture problem for individuals within a population is that the number of sites is known and sites can be sampled at every occasion, although the result of the sampling is still characterized by the uncertainty of possible nondetection. Early versions of occupancy models were developed by Geissler and Fuller (1987), Azuma et al. (1990), Nichols and Karanth (2002), and most current modeling is based on MacKenzie et al. (2002).

Sample units for occupancy studies may be naturally occurring units such as ponds or woodlots, or they may be cells in a grid superimposed on a continuous area. Each unit is surveyed on multiple occasions within a relatively short time period (e.g., 2 weeks) over which there are no changes in occupancy. Detection histories are analogous to capture histories and denote the sequence of detections and nondetections at each site. The statistics resulting from such a study are the numbers of sites exhibiting each possible detection history, e.g., x_{101} is the number of sites at which the species was detected on sampling occasions 1 and 3 of a 3-occasion study, but not occasion 2.

The modeling of the detection history data is similar to that for individual animal capture-recapture as well. Define p_t as the probability of detecting the focal species at a sample unit on sample occasion t, and ψ as the probability that a sample unit is occupied by the species. The probability that a surveyed sample unit shows detection history 101 is (MacKenzie et al., 2002):

$$\Pr(101) = \psi p_1(1 - p_2)p_3. \tag{6}$$

The probability for a site at which the species was not detected in any of the three surveys is:

$$\Pr(000) = \psi(1 - p_1)(1 - p_2)(1 - p_3) + (1 - \psi). \tag{7}$$

We know the species was present for detection history 101, as it was detected, and we assume no false positives (Eq. 6). However, history 000 admits more uncertainty, as there are two possibilities: the species was present and not detected or the species was absent (Eq. 7).

2.6 Occupancy Models, Open System, Single State

Open systems are those for which changes in occupancy status of sites may occur between some sample occasions. Define a primary sample occasion as a relatively short period (e.g., a specific month each year) during which occupancy status of a site is not likely to change. Multiple secondary samples (e.g., four survey days) occur within each primary period. However, the sites are permitted to be open to changes in occupancy between primary periods. For a study with three secondary occasions within each of two primary occasions, a detection history of 101,000 denotes a site with detections at secondary occasions 1 and 3 of primary occasion 1 and no detections in any of the three secondary occasions of primary occasion 2. Barbraud et al. (2003) developed an early model for such data, and the approach of MacKenzie et al. (2003) is the basis for most current modeling.

The modeling of detection probability requires an extra subscript for the two kinds of sampling occasions. Let p_{tk} denote the detection probability associated with secondary period k of primary period t. The possibility of changes in site occupancy requires two new parameters: ε_t is the probability that a site is unoccupied by the species at primary occasion $t + 1$, given that it was occupied at occasion t (local extinction); γ_t is the probability that a site is occupied by the species at occasion $t + 1$, given that it was not occupied in period t (local colonization). The probability associated with the above detection history is thus:

$$\Pr(101\,000) = \psi_1(p_{11}(1 - p_{12})p_{13})[\varepsilon_1 + (1 - \varepsilon_1)(1 - p_{12})(1 - p_{22})(1 - p_{23})]. \tag{8}$$

The portion of (8) in brackets reflects the uncertainty about whether the species went locally extinct at the site or instead persisted but went undetected. The likelihood is then the product of these probabilities for the detection histories of all sites.

2.7 Occupancy Models, Multiple States, False Positives

Sometimes we may want to characterize occupied sites by "state," where state carries additional information about an occupied site. A common situation is where a site occupied by a species can be classified into multiple states that can be ordered by the degree of uncertainty characterizing the state classification (Royle, 2004; Royle & Link, 2005, Nichols et al., 2007, MacKenzie et al., 2009). For example, assume interest in a species and an associated pathogen, such that we designate state

0 as a site not occupied by the focal species, and state 1 as a site occupied by the species but where no individuals of the species have been infected by the pathogen. State 2 denotes occupancy, with pathogen infection of at least one member of the species. In addition to these three true states, we define three observation states that can apply to a site at any secondary occasion survey: 0 = no detection of the species; 1 = detection of the species, but no detection of the pathogen; and 2 = detection of both the species and pathogen. Observation state 0 admits the most uncertainty, as true state may be 0, 1, or 2. For observation state 1, true state may be 1 or 2. Under the assumption of no false-positive errors, observation state 2 is unambiguous, only occurring when true state = 2. The notation and modeling of multistate occupancy become increasingly complex (see MacKenzie et al., 2009, 2018).

The initial development of occupancy modeling assumed no false positives, where these refer to the investigator claiming to detect a species, when the species is actually absent from the sample unit. False positives typically occur when the investigator mistakes an individual or sign of one species for that of another. For example, the pugmark (track) or scat of a large leopard may be mistakenly recorded as that of a tiger. Royle and Link (2006) developed a general, single-season occupancy model that incorporates both nondetection and false positives. Miller et al. (2011) developed models that use two (or more) different detection methods to deal with false positives, and these have been extended to multiple designs (Chambert et al., 2015) and multiple seasons (Miller et al., 2013; MacKenzie et al., 2018).

2.8 Software

Computations for capture-recapture estimation of focal parameters and their variance-covariance structures are relatively complex, such that development of software has been critical to the use of these methods. Early software focused on specific parameterizations of capture-recapture models, whereas development of numerical differentiation algorithms has led to more flexible software, permitting inference for user-specified models. A variety of software packages now exists for implementing capture-recapture analyses. For example, one website (https://www.capturerecapture.co.uk/software.html) managed by R. McCrea provides links to a number of available capture-recapture packages.

Program MARK (White & Burnham, 1999; Cooch & White, 2022) implements closed and open capture-recapture models, occupancy models, and a variety of other models useful for inferences about demographic parameters. Program PRESENCE (Hines, 2006) was developed specifically for occupancy models. Historically, PRESENCE has incorporated new classes of occupancy models before other occupancy software. Program M-SURGE (Choquet et al., 2004) was developed to implement multistate capture-recapture models and is based on sufficient statistics, resulting in typically faster computation times than software such as MARK, which is based on individual capture history data. Program E-SURGE (Choquet et al.,

2012) provides a general analytic framework for implementing multistate models in the presence of state uncertainty.

2.9 Summary

Both capture-recapture modeling and closely related occupancy modeling have undergone substantial evolution since their initial development for relatively simple inference problems. Most of this development has been motivated by scientists investigating animal populations and has focused on extensions and generalizations to either estimate additional parameters (i.e., beyond abundance and occupancy) or relax restrictive assumptions.

3 Beyond Traditional Applications

There have been many nontraditional uses of capture-recapture thinking and methodology. Reviews of social science and medical applications include Bohning (2008), Chao (2014), Bird and King (2018), and Bohning et al. (2018). The applications discussed in this chapter are not exhaustive but are illustrative of the diverse estimation problems to which these methods have been applied. Most of these nontraditional uses begin with a focus on abundance of some focal entity, combined with a recognition that the entity is frequently undercounted using the standard survey methods of the discipline.

3.1 Human Health and Epidemiology

3.1.1 Population-Level Inferences

Uses of capture-recapture models for human health applications have a fairly long history, with key early contributions by Wittes and Sidel (1968), Fienberg (1972), Wittes (1974), Wittes et al. (1974), Hook et al. (1980), Hook and Regal (1982, 1992), LaPorte et al. (1992), McCarty et al. (1993), and LaPorte (1994) and useful reviews by IWGDMF (1995a, 1995b), Hook and Regal (1995, 1999), and Chao et al. (2001). Virtually, all of these epidemiological uses are based on data from incomplete lists.

Some capture-recapture applications focus on single lists consisting of frequency distributions of encounters. A single list might include the number of infected individuals for which there was a single recorded encounter (e.g., blood test result, hospital visit), the number with exactly two encounters, three encounters, etc., with the objective to estimate the number of infected individuals that were never encountered. For example, Polonsky et al. (2018) used single list data to estimate the completeness, and thus effectiveness, of contact tracing.

Multiple list data are typically records of individuals infected with a particular disease or mortalities associated with a specific disease or other cause. In closed populations, the appearance of some individuals on one list and not another is clear evidence of nondetection, and early uses of multiple lists entailed first matching names that appear on multiple lists and then counting the total number of unique individuals. This approach does not include in the total count the number of individuals appearing on none of the lists, and inference about this number motivates the use of capture-recapture.

Multiple list data are encoded as individual capture or detection histories (Sects. 2.1 and 2.2), and the entire data set includes a detection history for every individual appearing on at least one list. For example, Hook et al. (1980) analyzed a data set consisting of three lists of individuals with spina bifida in New York state, 1969–1974. The lists were based on (1) birth certificates, (2) death certificates, and (3) medical rehabilitation records. Closed capture-recapture models were then used to estimate the total number of cases and disease “prevalence,” defined as the proportion of individuals in a population that is infected, or as the probability that a randomly selected individual in a population is infected. Numerous applications of capture-recapture to inferences about numbers of cases and prevalence now exist in the scientific literature.

Multiple list data differ from animal capture data in several ways that must be considered when selecting or developing capture-recapture models for epidemiological uses. The multiple lists are analogous to the multiple sampling periods of the animal ecologist, but unlike these animal sampling periods, there is frequently no natural temporal ordering of list data. Time-specificity of capture probabilities corresponds to list-specificity of detection probabilities. Certain kinds of behavioral response models in capture-recapture are based on temporal order of sampling occasions, and models (e.g., log-linear) for list data have been developed for more general kinds of dependence of detection probabilities for individuals among the different lists (see IWGDMF, 1995a, 1995b; Hook and Regal, 1995; Chao et al., 2001; Rivest and Lvesque, 2001).

Heterogeneous capture probabilities are sometimes associated with identifiable covariates, permitting inference based on a general Horvitz-Thompson approach (e.g., Huggins, 1989, 1991; also see Wang et al., 2006). Several approaches have been developed for the more difficult problem of heterogeneous capture probabilities that cannot be readily associated with covariates (e.g., Burnham & Overton, 1978; Chao, 1987; Norris & Pollock, 1995; Haas & Stokes, 1998; Dorazio & Royle, 2003; Haas et al., 2006). Problems deciding whether two similar records really match (represent the same individual) can occur when constructing detection histories from lists, and approaches for dealing with this problem (e.g., Seber et al., 2000; Lee et al., 2001) are similar, in some ways, to approaches for dealing with tag loss (e.g., Arnason & Mills, 1981; Kremers, 1988; Nichols & Hines, 1993) and misreading (e.g., McClintock et al., 2014).

Open capture-recapture models are used to estimate survival rates, numbers of new recruits, and abundance for populations open to gains and losses across multiple sampling occasions. In the context of disease dynamics, multistate models for open populations can be especially useful, where states are defined, for example, as susceptible, infected, and recovered (SIR), a classification system used in classical compartmental disease models (Kermack & McKendrick, 1927; Bailey, 1975; Cooch et al., 2012). List data based on hospital visits or longitudinal data from studies with imperfect follow-up can be used to develop detection histories for such analyses, and estimated parameters include probabilities of state transition (e.g., the transition from susceptible to infected) and state-specific mortality rates. A feature of multistate capture-recapture models that is especially important for epidemiological uses is state-specific detection probabilities (e.g., infected individuals will typically have higher probabilities of detection for hospital lists than susceptible individuals). Uncertainty in state assignment (e.g., a false-negative or false-positive pathogen test result) led to the development of models to deal with this issue (reviewed by Lebreton et al., 2009), and the multi-event approach of Pradel (2005) provides a general approach to this problem (Conn & Cooch, 2009; Choquet et al., 2013; Benhaiem et al., 2018). Multistate capture-recapture models have been recommended for use in estimating epidemiological state transition probabilities and mortality rates (Jennelle et al., 2007; Cooch et al., 2012; Nichols et al., 2017), but such uses have been relatively rare (but see Viallefont & Auget, 1999) for human diseases.

Occupancy models have several potential uses for epidemiological studies. One use entails viewing individuals as the sample units and focusing on presence or absence of the disease organism (e.g., Bailey et al., 2014; MacKenzie et al., 2018). Multistate occupancy models can be used to estimate transitions (including infection rate) among SIR model states and state-specific mortality rates, as with multistate capture-recapture. Pathogen tests for a random or representative sample of individuals can be used with standard occupancy models to estimate prevalence in the case where false negatives (nondetection) are possible (e.g., Lachish et al., 2012; Nichols et al., 2021), and even infection intensity (Miller et al., 2012). Such testing programs should typically include a subset of individuals that receive multiple tests in order to deal with nondetection.

Occupancy models can also be used to model spatial dynamics of disease spread. Data are based on tests of individuals, but now the sample unit is a location (e.g., a county or city), and interest is in whether any infected individuals are present (McClintock et al., 2010; Bailey et al., 2014; MacKenzie et al., 2018). List data could come from hospital visits, and the replication required for the most general occupancy modeling could be obtained by treating each day or week as a sampling occasion. The ability to deal with imperfect detection is especially important in such studies, as detection probabilities are likely to vary among different locations (e.g., urban locations vs. rural locations far from medical centers).

3.1.2 Individual-Level Inferences

Decisions about individual treatment and quarantine depend on the same diagnostic test results that populate lists. When such tests admit false negatives and positives, it is useful to estimate the probability that a specific individual is infected, conditional on the test result(s). Define p_{lm} as the probability that a test result $(x_k = l)$ indicates individual k to be in disease state l, given that true disease state is m. Define state $z_k = 1$ to mean that individual k is infected and state $z_k = 0$ to mean uninfected. Then p_{11} is the probability of correctly detecting infection when present, and its complement $(1 - p_{11})$ is the probability of a false negative. Similarly, p_{10} is the probability of a false positive, incorrectly declaring an individual in state $z_k = 0$ to be infected. Because these probabilities are conditional on the unknown true state of the tested individual, statements about the probability of true infection are also conditional on the underlying pathogen prevalence, ψ. All of the above parameters (detection/classification and prevalence) can be estimated directly using single-season occupancy models (Miller et al., 2011; Chambert et al., 2015; MacKenzie et al., 2018).

The conditional probability that an individual testing positive is actually infected ("positive predictive value") can be written as:

$$\Pr(z_k = 1|x_k = 1) = \frac{\psi p_{11}}{\psi p_{11} + (1 - \psi)p_{10}}. \tag{9}$$

The conditional probability that an individual testing negative is truly not infected (termed "negative predictive value") can be written similarly as:

$$\Pr(z_k = 0|x_k = 0) = \frac{(1 - \psi)(1 - p_{10})}{\psi(1 - p_{11}) + (1 - \psi)(1 - p_{10})}. \tag{10}$$

Note that if prevalence parameters are likely to differ for different groups of individuals (e.g., those exhibiting symptoms and those not), then group-specific prevalence parameters should be estimated and used. If the probabilities of an accurate test result [(9) and (10)] are thought to be too small for important decisions about individual treatment, then multiple tests can be used to increase them (e.g., Nichols et al., 2021).

An advantage of the occupancy approach over that frequently used by epidemiologists is that all of the relevant parameters can be estimated together in a joint likelihood. The probabilities of an individual being infected are computed directly as derived parameters, with the associated estimates of sampling variance properly accounting for the variances and covariances of the different parameter estimates.

3.2 Social Sciences

3.2.1 Census

Governments of most countries conduct periodic "censuses" of population size and distribution. However, direct counts are seldom possible, and virtually all census methods miss individuals (false negatives). Laplace (1786) was the first to derive the estimator (2) and used it to compute the human population size of France using two lists of citizens and their degree of overlap. Sekar and Deming (1949) appeared to derive the estimator (2) independently of (Laplace, 1786) and (Lincoln, 1930) and used it to draw inferences about the numbers of human births and deaths in a district near Calcutta, India. Application of capture-recapture methods (sometimes referred to as "multiple systems estimation" in the social science literature; Fienberg & Manrique-Vallier, 2009; Bird & King, 2018) to problems in the social sciences has increased in recent decades prompting methodological reviews (e.g., Bohning, 2008, Bird & King, 2018) and a book (Bohning et al., 2018).

Capture-recapture models for closed populations have been extended by scientists working with the United States Census Bureau and used with post-enumeration surveys to estimate the census undercount (Wolter, 1986, 1990; Cowan & Malec, 1986; U.S. Census, 2021). Evaluation of census coverage using post-enumeration surveys along with capture-recapture estimation has been recommended by the United Nations (Demographic and Social Statistics Branch, United Nations Statistics Division, 2009) and is being used by various countries (e.g., UK Abbott, 2009; Turkey, Ayhan & Ekni, 2003; Australia, Australian Bureau of Statistics, 2012). In addition to use with standard governmental censuses, capture-recapture approaches have been especially useful for providing inferences about "hidden" populations, groups of individuals that are especially difficult to count using conventional surveys, frequently because they do not wish to be counted (e.g., Sudman et al., 1988).

3.2.2 Homeless

Homeless persons are a problematic group for conventional governmental census methods, as they typically lack a mailing address and are not motivated to provide census information. Fisher et al. (1994) obtained list data for homeless persons from multiple sources including hospitals, local social service agencies, a healthcare center designated for homeless, and hostels and used capture-recapture to estimate the homeless population in an area of London. Their estimate was approximately three times larger than the number of list-identified individuals. Such multiple-list approaches have been used with capture-recapture modeling to estimate homeless populations elsewhere as well (e.g., Baltimore, Cowan et al., 1986; Budapest, David & Snijders, 2002). Berry (2007) used an observational approach to identify homeless individuals in Toronto on the street during multiple sampling occasions.

Closed capture-recapture models were used to estimate the homeless population, and detection probabilities of about 0.2 indicated the importance of dealing with nondetection.

An alternative approach to multiple lists is to insert some number of “marked” (M) or “planted” individuals into the focal homeless population and then survey the population directly, estimating detection probabilities as the proportion of planted individuals that is detected (Eq. 3). Laska and Meisner (1993) identified 103 sites frequented by homeless persons in a region of New York City and planted persons in a random sample of 41 of these sites. Census Bureau enumerators were then sent to directly survey homeless persons at these sites. Detection probability was estimated to be 0.48 and used to estimate the total number of homeless in the surveyed areas.

3.2.3 Problem Drug Users

Capture-recapture models have been used with list data on individuals to estimate numbers of problem drug users in various locations. For example, King et al. (2014) used four list sources, probation records, drug intervention program prison assessments, drug treatment facility records, and drug intervention program community assessments, to estimate the number of injecting drug users and heroin-associated deaths in England. They used a Bayesian approach to incorporate prior information into their capture-recapture modeling, obtaining estimates for England, as well as for specific regions within the country. Both prevalence of problem drug use and detection probabilities (probability that a problem drug user appears on at least one list) showed substantial regional variation. Approaches based on similar list data were used to estimate numbers of injecting drug users in Scotland (King et al., 2013). Capture-recapture approaches to inference about problem drug use are numerous and include inferences about the number of HIV-infected injecting drug users in Bangkok (Mastro et al., 1994), prevalence of opiate use in Dublin (Comiskey and Barry, 2001), prevalence of problem drug use in London (Hickman et al., 1999) and six French cities (Vaissade & Legleye, 2008), the risk of arrest of drug dealers and users in Quebec (Bouchard & Trembley, 2005), and the number of heroin users in the Australian Capital Territory (Larson et al., 1994).

3.2.4 Criminal Activities

Greene and Stollmack (1981) applied closed population capture-recapture methods to records from approximately 6000 males arrested at least once in Washington, D.C., 1974–1975. They estimated a total criminal population of about 30,000 individual criminals. Using these same data, Greene (1984) later applied an open population model permitting inferences about growth rate of the offender population, survival probabilities, and average criminal career length. Bouchard et al. (2019) used capture-recapture with arrest and rearrest record data from Quebec to estimate the number of criminals involved with illegal amphetamine-like stimulants.

They estimated that total arrests were only about 12% of those actually engaged in illegal activities and subject to arrest. Bouchard (2007) used capture-recapture methods with arrest data to estimate the number of criminal marijuana growers in Quebec, 1998–2002. Charette and van Koppen (2016) used capture-recapture methods to investigate selectivity in crime punishment, concluding that black male offenders were more likely to be arrested and punished than members of other demographic groups.

Cases of domestic violence in the Netherlands, 2006–2007, were estimated by van der Heijden (2014) using capture-recapture methods with police register records. Their estimates indicated that about 22% of offenders were actually observed and recorded by police. Silverman (2014) used capture-recapture modeling of multiple list data to estimate the number of victims of human trafficking in the UK, 2013. Data on individual victims came from six lists: local authority, police force, national government organization, nongovernment organization, National Crime Agency, and the general public. The estimated victim population was four to five times larger than the number of individuals detected.

Corlatti et al. (2019) studied illegal poaching of red deer in a park in the central Italian Alps, 2007–2017. They estimated age- and sex-specific mortality rates of deer associated with poaching and non-poaching sources using open, multi-event models with data for tagged red deer. Their modeling included parameters for tag loss and the possibility of misclassifying the cause of death (by poaching or not) and provided strong evidence of higher poaching mortality for older males than any other age-sex class.

Barber-Meyer (2010) proposed use of occupancy models with data on species (e.g., tiger parts and products) sold illegally at souvenir shops, traditional medicine stores, etc., within towns. Replication is provided by the multiple stores and shops within each town. Towns were the sample units, such that occupancy estimated the proportion of towns at which the focal species was illegally sold, and multiseason models could be used to estimate occupancy dynamics over time. Sharma et al. (2014) used reports of annual tiger poaching events reported by the Wildlife Protection Society of India, in conjunction with multiseason occupancy modeling, to estimate the prevalence of tiger poaching during periods of 3–7 years in 605 districts throughout India over a 40-year period. Results provided maps of tiger poaching crime and information about covariates associated with such crime.

Yeo et al. (2017) used eBay postings to estimate aspects of illegal elephant ivory trade dynamics in the UK. Each posting was identified by a description, item number, and seller identification, permitting identification of the item in subsequent postings. Postings were surveyed once per week for eight consecutive weeks, March–May 2014. Detection histories were developed for every item and used with open-population capture-recapture models to estimate numbers of items, as well as weekly survival (persistence in the eBay market) and entry probabilities. The authors concluded that a large fraction of illegal ivory sale items had very low probabilities of detection.

3.2.5 World Conflicts

Armed conflicts throughout the world result in numbers of persons being killed or disappearing, and “counts” of these victims are typically biased low. Capture-recapture methods have been used with casualty list data to estimate numbers of victims associated with conflicts in Peru (Ball et al., 2003; Manrique-Vallier et al., 2013) and Colombia (Lum et al., 2013); number of deaths in Kosovo, March–June 1999 (Ball & Asher, 2002); and the number of persons killed by state forces in Guatemala, 1981–1983 (Ball, 2000). For example, in the Guatemala analysis, lists of victims were provided by the following three sources, the Commission for Historical Clarification, the International Center for Human Rights Research, and the Catholic Church’s Interdiocesan Project for the Recuperation of Historical Memory. The estimated number of killings was about three times larger than the sum of victims identified via the three lists.

Social conflict events from some parts of the world are not well reported, such that counts of such events are typically biased low. Hendrix and Saleyhan (2015) used closed population capture-recapture models to estimate the number of social conflict events occurring across Africa in 2012. They obtained detection/nondetection data on 1443 events from the Social Conflict in Africa Database. They used data from two independent news agencies, Associated Press (AP) and Agence France-Presse (AFP), compiling statistics on numbers of events reported only by AP, only by AFP, and by both agencies. They concluded that these two news sources captured approximately 76% of all events in Africa and that the nondetection rate was predictably smaller for deadly events, events of a larger magnitude, and events associated with government repression.

3.3 Quality Control

Capture-recapture models have been used for several specific problems associated with quality control. Jewell (1985) noted that defects or errors can occur in production of various manufactured goods, in computer software, in manuscripts, etc. and recommended capture-recapture approaches for estimating numbers of them. Quality control efforts typically involve inspectors or proofreaders who examine products for defects or errors, but errors may go undetected. One approach to estimating number of errors/defects in the face of nondetection is to employ multiple inspectors or proofreaders. In the case of three inspectors, for example, each error detected by at least one inspector is represented by a row of three entries (one entry for each inspector), with a 1 denoting detection of the error by the particular inspector and a 0 denoting nondetection. Chao and Yang (1993) used this approach with computer code examined by multiple coders looking for errors and estimated the number of errors remaining (undetected). White et al. (1982) used this approach with multiple proofreaders of a large manuscript and estimated the number of undetected errors.

3.4 Remote Sensing

"Remote sensing" refers to use of aircraft or satellites to obtain information about the earth. Photography and video are typically used to provide images, which are then examined by individuals or computers in order to enumerate focal entities (e.g., wetlands, woodlots) or compute area measurements of specific cover types. However, such analyses of remote sensing images are usually characterized by two types of errors, nondetection and misclassification. These errors can sometimes be dealt with via replication provided by multiple observers in aircraft or multiple persons processing the same image. Capture-recapture methods are then used to estimate number of entities, for example, using the number of entities detected by just one observer, two observers, etc. (Magnusson et al., 1978; Cook and Jacobson, 1979).

More commonly, a sample of area covered by a survey is visited by ground observers providing direct counts and classifications, known as ground truthing. The number of ground truth entities that is correctly detected or classified via the remote images is then used with capture-recapture thinking to estimate detection and correct classification probabilities (see Maxim et al., 1981; Maxim & Harrington, 1982, 1983). Veran et al. (2012) focused on the question of land cover dynamics, noting that classification errors can be made at times t, $t + 1$, or both times, leading to large errors in estimates of land cover state transition probabilities. They proposed use of ground truth data with multistate capture-recapture models that included state misclassification as a means of directly estimating land cover transitions in the face of classification errors.

3.5 Paleobiology

Paleobiologists have long recognized that nondetection is an important issue for analyzing fossil data (e.g., Foote & Raup, 1996). Analyses that do not account for nondetection are subject to serious errors, as detection probabilities are thought not only to be substantial but also to vary across time and space (Brett, 1998). Paleobiological data consist of records of fossil taxa found via sampling at different strata (different geologic time horizons) and locations. Capture-recapture analyses typically treat each lower-level taxon (e.g., family) within a higher-level taxon (e.g., phylum) as an "individual." Detection histories for each lower-level taxon can be developed using spatial samples (analogous to multiple lists) within some time stratum and area of interest, providing the data for estimation of total taxa using closed capture-recapture models. Detection histories for focal taxa developed from different geologic strata (time horizons) at the same sampling location, or even worldwide, can be used with open capture-recapture models to estimate number of taxa and rates of both local and global taxonomic origination and extinction for lower-level taxa (Nichols & Pollock, 1983). Capture-recapture models were introduced to paleobiology over 40 years ago (Rosenzweig & Duek, 1979; Nichols & Pollock, 1983; Conroy & Nichols, 1984; Nichols et al., 1986), but they have seen only limited use (Connolly & Miller, 2001a, 2001b, 2002).

Occupancy models have multiple uses for fossil data as well, with focus on a specific taxon, rather than a group of lower-level taxa. Detection-nondetection data from replicate local samples can be used in conjunction with occupancy models to estimate geographic distribution (Liow &Nichols, 2010; Liow, 2013). Detection histories based on different time horizons from multiple locations can be used with occupancy modeling to estimate local probabilities of colonization and extinction as well (Liow &Nichols, 2010; Liow, 2013). Occupancy models were introduced to paleobiologists much more recently than capture-recapture models (Liow &Nichols, 2010; Liow, 2013; MacKenzie et al., 2018), and paleobiological use of occupancy approaches has been limited (but see Lawing et al., 2021).

3.6 *Miscellaneous Applications*

National databases for traffic accidents are maintained by law enforcement agencies in many countries, but accidents are thought to be underreported, leading to many efforts to estimate their true numbers using capture-recapture. Razzak and Luby (1998) compiled lists of police accident records and emergency ambulance service records over a 10-month period during 1994 in Karachi, Pakistan. Their estimates indicated that official records accounted for 56% of traffic accident deaths and only 4% of serious injuries. Capture-recapture inferences about traffic accidents have been used in various other locations including Nicaragua (Tercero & Andersson, 2004), Ethiopia (Abegaz et al., 2014), and Mali (Sango et al., 2016).

Beirne and Lambin (2013) studied volunteer "citizen scientists" who worked on a project to remove invasive mink from a large area of Scotland. Their objective was to draw inferences about volunteer retention (tendency to remain in the program) and the factors that affected it. They described the potential utility of open capture-recapture approaches, but collected data on volunteer activity data via telephone every 6 months and were thus able to use known-fate models (Pollock et al., 1989). They identified volunteer vocation and recent trapping and removal success as key determinants of retention in the program.

Interest in vocabulary size has prompted literary scholars to count the number of individual words that an author uses in her/his writing, but this number is likely smaller than the number actually known to the author. Efron and Thisted (1976) counted the number of words used once, twice, three times, etc., in samples of Shakespeare's writing in order to estimate the total number of words that he knew. Words counted in the samples totaled 31,534, and capture-recapture estimators indicated that he knew about 35,000. Capture-recapture methods have also been used to estimate the song repertoire size of birds (Garamszegi et al., 2002).

An archaeological use of capture-recapture modeling was provided by Holst (1981), with subsequent reanalyses using different capture-recapture estimators by

Esty (1982, 1983) and Chao (1984). The problem was to estimate the number of different "dies" that produced a set of 204 coins in ancient India. The data were the number of dies that produced only a single coin in the sample, two coins in the sample, three, etc.

Herendeen and White (2013) collected data on appearances of specific rare stamps over the years from sources such as auction catalog, retail price lists, copies of expert certificates, and similar records. Each stamp has an individual identifying number. Herendeen and White (2013) viewed each year as a sampling occasion and used closed population models to estimate the total number of stamps still in existence.

Nichols et al. (2013a, 2013b) used capture-recapture thinking to estimate detection and classification probabilities for military imaging systems. Vessels of different classes (defined by size, and military vs. civilian status) were experimentally positioned at different distances from two new cameras. "No vessel" was one of the experimental possibilities as well. Resulting data were used to develop a model for detection and classification probabilities as a function of distance and vessel type. Capture-recapture model selection was used to infer that distance relationships were dependent on camera type but characterized by a common slope across vessel types (Nichols et al., 2013b).

K.H. Pollock (pers. comm.) used capture-recapture models to estimate the number of man-made objects orbiting earth. The field of astronomy is characterized by substantial nondetection, with detection probabilities a function of telescope type as well as distance, brightness, and size of focal object, and a number of potential uses of capture-recapture thinking can be envisaged.

4 Discussion

The problems of nondetection and misclassification characterize numerous types of count data. The various applications described in Sect. 3 have hopefully supported this assertion, and there are certainly many more applications that can be imagined. The adoption of capture-recapture thinking has not been as rapid as might be hoped for any of the disciplines of Sect. 3, and rate of adoption has varied among these disciplines. For example, my impression based on literature review is that epidemiological and human health applications are somewhat more common than those dealing with social sciences, whereas adoption within paleobiology has been very slow.

One possible explanation underlying this variation in rate of adoption involves the perceived severity of nondetection and misclassification (Nichols, 2019). For example, the development of capture-recapture thinking in the fields of wildlife and animal ecology is likely a natural response to the well-known nondetection problems associated with virtually all animal survey methods. Not only do surveys miss animals, but the fraction missed can be very large. In contrast, epidemiological and social science data based on counts of humans have historically been thought to be closer to truth, although this perception is changing. Indeed, some of the detection probability estimates of Sect. 3 are quite small.

A second factor that may affect rate of adoption of capture-recapture is the cost of incorrect counts and estimates, and thus the scrutiny that analytic results receive. Results of epidemiological surveys and medical diagnoses are viewed as extremely important and are often carefully reviewed, as misleading inferences can have detrimental consequences that may be readily apparent. In contrast, inferences in the social sciences are certainly important, but not so highly scrutinized, and misleading inferences are less likely to be recognized. The greater the degree of scrutiny, the greater the expected attention to analytic details and inferential errors.

Another factor that may affect rate of adoption is the funding available to a discipline. Good funding helps ensure collaboration of statisticians, who are able to deal with the added complexity of modeling sampling processes. Epidemiology and human health are among the better funded scientific disciplines in most countries.

One more factor influencing methodological adoption is likely the familiarity of scientists with capture-recapture approaches. My search for uses of capture-recapture models for inferences about criminal activity produced a number of papers that used capture-recapture to study the criminal activity of animal poaching. I am guessing that the appearance of disproportionate numbers of applications for this particular type of crime resulted from prior familiarity of scientists investigating such crimes with capture-recapture approaches.

This relatively slow adoption of robust methods for dealing with nondetection and misclassification begs the question: what are the alternatives to modeling these components of the sampling process? The most common alternative appears to be to view the problems as so small and insignificant that they can be safely ignored. For example, this has been the case with remote sensing uses, as ground truthing data have provided clear evidence of nondetection and misclassification. However, estimates of these errors are frequently presented, claimed to be small, and then ignored in analysis (see discussion in Veran et al., 2012). I suspect that this alternative is also prevalent in disciplines where errors are not so readily estimated, but rather assumed or claimed to be small and thus not worthy of the effort to deal with nondetection.

A second alternative to the use of capture-recapture is to try to identify the key sources of variation in detection probability or misclassification, to develop models for each of these component processes separately, and then to combine these models to provide overall inferences about detection probability and the focal parameters that they influence. I encountered this approach at a 2011 workshop dealing with nondetection. A biostatistician working for the Centers for Disease Control and Prevention outlined this approach to inference about detection probability for a focal disease. Her strategy was to develop a model for each of 10–12 sources of variation in detection and to then combine these models into an overall model for the sampling process. At the time of the workshop, two of these models had been developed. In contrast to this incremental approach, capture-recapture requires some form of replicate sampling (e.g., multiple lists) and then directly uses the information about nondetection available in detection history data. Detection probability, as well as focal parameters such as numbers of cases or disease prevalence, is estimated without the need for identification and modeling of all factors affecting nondetection. If potential influencing factors can be identified, then they can be treated as covariates in capture-recapture modeling, with the result that their influence can be formally tested and, if found to be important, included in the modeling of detection probability.

A problem related to slow adoption of capture-recapture approaches is the limited inferences to which they are applied. The early history of capture-recapture in wildlife and animal ecology was dominated by a focus on numbers. Capture-recapture estimators for closed populations were used to provide estimates of population size for specific locations and times. However, abundance is not necessarily interesting by itself, but rather is more usefully viewed as a state variable in studies of dynamical processes. The primary interests are sources of spatial and temporal variation in abundance, and the ability of human actions to influence population size. This recognition eventually led to increased interest in capture-recapture models for open populations that experience dynamical changes between sampling occasions. Today's capture-recapture studies of wildlife populations tend to focus on the processes of birth, death, and movement, and on transition probabilities governing changes of state within individuals. Similarly, occupancy estimates themselves are not viewed as especially interesting, and focus has shifted to the probabilities of local extinction and colonization that govern occupancy dynamics.

This review of nontraditional applications of capture-recapture models suggests to me that most of these studies are focusing on numbers of focal entities. As noted in Sect. 3.1.1, epidemiological list data from hospital visits and data from longitudinal studies with incomplete follow-up can be used with multistate capture-recapture models for open populations to draw inferences about state transition probabilities and state-specific mortality rates required by SIR (susceptible, infected, recovered) models. These models provide a way to deal with the state-specific detection probabilities likely to exist in longitudinal data. For example, an individual in the infected state at sampling occasion t is more likely to be found on a hospital list or re-encountered in a longitudinal study at that occasion than an individual in the susceptible or recovered state. Accompanying information on public health interventions or even individual treatments can be used with these models to directly test the efficacy of interventions and treatments. Despite the potential utility of multistate (Lebreton et al., 2009) and multi-event (Pradel, 2005) capture-recapture models, I saw little evidence that these approaches are being used in disciplines other than animal and wildlife ecology.

The primary interest of most capture-recapture applications in criminology (Sect. 3.2.4) was in numbers of criminals or victims, or in detection probability when this equated with probability of arrest. These studies were not focused on the influences of laws or enforcement interventions on criminal activity, or on the effects of crimes on victims. Contrast this with a number of the studies of wildlife poaching crimes which included investigations of poaching-related mortality rates (Corlatti et al., 2019), and of relationships between dynamics of poaching activity and species distributions (Marescot et al., 2019; Moore et al., 2021), and between ranger (enforcement) activity and poaching activity (Moore et al., 2017). I suspect that a main reason for this difference in uses of capture-recapture for wildlife vs. other crimes stems from the familiarity of persons focused on wildlife crimes with these more complicated models and their utility.

With the exception of these investigations of wildlife crimes, the majority of the nontraditional uses of capture-recapture have focused on numbers. Such studies can be very useful when these numbers are incorporated into a larger sampling scheme designed to test hypotheses about system dynamics or effects of potential interventions. However, for most of the reviewed papers describing nontraditional uses of capture-recapture, this was not the case. I believe that studies that go beyond estimates of numbers to focus on system dynamics and key relationships (e.g., effects of interventions or treatments) are much more likely to be useful to both science and decision-making. Capture-recapture and occupancy models developed for open populations are especially useful for investigating underlying processes, and I would hope that we see the same increases in use of these models in nontraditional applications as we did in the fields of wildlife and animal population ecology.

In summary, studies in many disciplines are based on count data, yet counts are frequently inaccurate because of nondetection and misclassification. The fields of wildlife and animal ecology recognized these problems nearly a century ago and began to develop capture-recapture approaches to deal with them. Other disciplines began to adopt these methods and, in some cases, to modify them for their specific applications (e.g., log-linear models for closed populations to deal with list dependence). However, the integration of capture-recapture methods into the toolboxes of scientists of non-ecological disciplines has been incomplete and slower than might be hoped. In addition, the bulk of non-traditional uses of capture-recapture models has focused on estimation of totals for counted entities. I hope that use of capture-recapture models for nontraditional applications increases and that such uses better exploit the open-population models that permit inferences about dynamical processes.

Acknowledgments Thanks to Dainel F. Heitjan and H.K.T. Ng for the invitation to contribute to this volume, for useful comments on the initial manuscript, and for information about nontraditional uses of capture-recapture thinking. Constructive comments on the manuscript were also provided by an anonymous reviewer. Thanks to Ken Burnham and Ken Pollock for providing information on nontraditional uses of capture-recapture and to Evan Cooch for discussions of this topic. Thanks especially to Dr. Lynne S. Stokes for the 3 years that she spent working at a US Fish and Wildlife Service research lab early in her career. Lynne was a great teacher and collaborator and played a major role in increasing my statistical capabilities, as well as those of our entire research group.

References

Abbott, O. (2009). 2011 UK Census coverage assessment and adjustment methodology. *Popular Trends, 137*, 2532.

Abegaz, T., Berhane, Y., Worku, A., Assrat, A., & Assefa, A. (2014). Road traffic deaths and injuries are under-reported in Ethiopia: A capture-recapture method. *PLoS ONE 9*(7), e103001. https://doi.org/10.1371/journal.pone.0103001

Arnason, A. N. (1972). Parameter estimates from mark-recapture experiments on two populations subject to migration and death. *Researches on Population Ecology, 13*, 97–113.

Arnason, A. N. (1973). The estimation of population size, migration rates, and survival in a stratified population. *Researches on Population Ecology, 15*, 1–8.

Arnason, A. N., & Mills, K. H. (1981). Bias and loss of precision due to tag loss in Jolly-Seber estimates for mark-recapture experiments. *Canadian Journal of Fisheries and Aquatic Sciences, 38*, 1077–1095.

Australian Bureau of Statistics. (2012). 2011 census of population and housing: details of undercount. Report 2940.0. Canberra: Australian Bureau of Statistics.

Ayhan, H.Ö., & Ekni, S. (2003). Coverage error in population censuses: The case of Turkey. *Survival Methods, 29*, 155165.

Azuma, D. L., Baldwin, J. A., & Noon, B. R. (1990). Estimating the occupancy of spotted owl habitat areas by sampling and adjusting for bias. USDA Gen. Tech. Rep. PSW-124, Berkeley, CA.

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases* (2nd ed.). Macmillan.

Bailey, L. L., MacKenzie, D. I., & Nichols, J. D. (2014). Advances and applications of occupancy models. *Methods in Ecology and Evolution, 5*(12), 1269–1279.

Ball, P. (2000). The Guatemalan commission for historical clarification: Intersample analysis. In P. Ball, H. Spirer & L. Spirer (Eds.), *Making the case: Investigating large scale human rights violations using information systems and data analysis* (pp. 259–285). American Association for the Advancement of Science.

Ball, P., & Asher, J. (2002). Statistics and Slobodan: Using data analysis and statistics in the war crimes trial of former president Milosevic. *Chance, 15*, 17–24.

Ball, P., Asher, J., Sulmont, D., & Manrique, D. (2003). How many Peruvians have died? An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000 Washington, DC: Report to the Peruvian Commission for Truth and Justice (CVR).

Barber-Meyer, S. M. (2010). Dealing with the clandestine nature of wildlife-trade market surveys. *Conservation Biology, 24*, 918–923.

Barbraud, C., Nichols, J. D., Hines, J. E., & Hafner, H (2003). Estimating rates of extinction and colonization in colonial species and an extension to the metapopulation and community levels. *Oikos, 101*, 113–126.

Beirne, C., & Lambin, X. (2013). Understanding the determinants of volunteer retention through capture-recapture analysis: Answering social science questions using a wildlife ecology toolkit. *Conservation Letters, 6*, 391401.

Benhaiem, S., Marescot, L., Hofer, H., East, M. L., Lebreton, J.-D., Kramer-Schadt, S., Gimenez, O. (2018). Robustness of eco-epidemiological capture-recapture parameter estimates to variation in infection state uncertainty. *Frontiers in Veterinary Science, 5*. https://doi.org/10.3389/fvets.2018.00197

Berry, B. (2007). A repeated observation approach for estimating the street homeless population. *Evaluation Review, 31*, 166–199.

Bird, S. M., & King, R. (2018). Multiple systems estimation (or capture-recapture estimation) to inform public policy. *Annual Review of Statistics and Its Application, 5*, 95–118.

Bohning, D. (2008). Editorial—recent developments in capture-recapture methods and their applications. *Biometrical Journal, 50*, 954956.

Bohning, D., van der Heijden, P. G. M., & Bunge, J. (Eds.). (2018). *Capture-recapture methods for the social and medical sciences*. CRC Press

Bouchard, M. (2007). A capture-recapture model to estimate the size of criminal populations and the risks of detection in a marijuana cultivation industry. *Journal of Quantitative Criminology, 23*, 221–241.

Bouchard, M., Morselli, C., MacDonald, M., Gallupe, O., Zhang, S., & Farabee, D. (2019). Estimating risks of arrest and criminal populations: Regression adjustments to capture-recapture models. *Crime & Delinquency, 65*, 1767–1797.

Bouchard, M., & Trembley, P. (2005). Risks of arrest across drug markets: A capture-recapture analysis of "hidden" dealer and user populations. *Journal of Drug Issues, 35*, 733–754.

Brett, C. E. (1998). Sequence stratigraphy, paleoecology, and evolution; biotic clues and responses to sea-level fluctuations. *Palaios, 13*, 241–262.

Brownie, C., Hines, J. E., Nichols, J. D., Pollock, K. H., & Hestbeck, J. B. (1993). Capture-recapture studies for multiple strata including non-Markovian transition probabilities. *Biometrics, 49*, 1173–1187.

Burnham, K. P., & Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika, 65*, 625–633.

Chambert, T., Miller, D. A. W., & Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology, 96*, 332–339

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics, 11*, 265–270.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics, 43*, 783–791.

Chao, A. (2014). *Capture-recapture for human populations*. Wiley stats ref: Statistics reference Online. Wiley.

Chao, A., & Huggins, R. M. (2005a). Classical closed-population capture-recapture models. In S. C. Amstrup, , T.L. McDonald & B. F. J. Manly (Eds.), *Handbook of capture-recapture analysis* (pp. 22–35). Princeton Univ. Press.

Chao, A., & Huggins, R. M. (2005b). Modern closed-population capture-recapture models. In S. C. Amstrup, , T.L. McDonald & B. F. J. Manly (Eds.), *Handbook of capture-recapture analysis* (pp. 58–87). Princeton Univ. Press.

Chao, A., Tsay, P. K., Lin, S. -H., Shau, W. -Y., & Chao, D. -Y. (2001). The applications of capture-recapture models to epidemiological data. *Statistics in Medicine, 20*, 3123–3157.

Chao, A., & Yang, M. C. K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika, 80*, 193201.

Charette, Y., & van Koppen, V. (2016). A capture-recapture model to estimate the effects of extra-legal disparities on crime funnel selectivity and punishment avoidance. *Security Journal, 29*, 561–583.

Choquet, R., Reboulet, A. M., Pradel, R., Gimenez, O., & Lebreton, J. D. (2004). M-SURGE new software specifically designed for multistate capture-recapture models. *Animal Biodiversity and Conservation, 27*(1), 207–215.

Choquet, R., Rouan, L., & Pradel, R. (2012). Program E-SURGE: A software application for fitting multievent models. In D. L. Thomson, E. G. Cooch, M. J. Conroy (Eds.), *Modeling demographic processes in marked populations* (Vol. 3, pp. 845–865). Environmental and Ecological Statistics. https://doi.org/10.1007/978-0-387-78151-8

Choquet, R., Carrie, C., Chambert, T., & Boulinier, T. (2013). Estimating transitions between states using measurements with imperfect detection: Application to serological data. *Ecology, 94*, 2160–2165.

Comiskey, C. M., & Barry J. M. (2001). A capture-recapture study of the prevalence and implications of opiate use in Dublin. *European Journal of Public Health, 11*, 198–200.

Conn, P. B., & Cooch, E. G. (2009). Multi-state capture-recapture analysis under imperfect state observation: An application to disease models. *Journal of Applied Ecology, 46*, 486492.

Connolly, S. R., & Miller, A. I. (2001a). Global Ordovician faunal transitions in the marine benthos: Proximate causes. *Paleobiology, 27*, 779–795.

Connolly, S. R., & Miller, A. I. (2001b). Joint estimation of sampling and turnover rates from fossil databases: Capture-mark-recapture methods revisited. *Paleobiology, 27*, 751–767.

Connolly, S. R., & Miller, A. I. (2002). Global Ordovician faunal transitions in the marine benthos: Ultimate causes. *Paleobiology, 28*, 26–40.

Conroy, M. J., & Nichols, J. D. (1984). Testing for variation in taxonomic extinction probabilities: A suggested methodology and some results. *Paleobiology, 10*, 328–337.

Cooch, E. G., Conn, P. B., Ellner, S. P., Dobson, A. P., & Pollock, K. H. (2012). Disease dynamics in wild populations: Modeling and estimation: A review. *Journal of Ornithology, 152*, 485–509.

Cooch, E. G., & White, G. C. (2022). *Program MARK – A Gentle Introduction* (22nd ed.). http://www.phidot.org/software/mark/docs/book/

Cook, R. D., & Jacobson, J. O. (1979). A design for estimating visibility bias in aerial surveys. *Biometrics, 35*, 735–742.

Corlatti, L., Sanz-Aguilar, A., Tavecchia, G., Gugiatti, A., & Pedrotti, L. (2019). Unravelling the sex- and age-specific impact of poaching mortality with multievent modeling. *Frontiers in Zoology, 16*, 20. https://doi.org/10.1186/s12983-019-0321-1

Cormack, R. M. (1964). Estimates of survival from the sightings of marked animals. *Biometrika, 51*, 429–438.

Cowan, C. D., & Malec, D. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association, 81*, 347–353.

Cowan, C. D., Breakey, W. R., & Fischer, P. J. (1986). The methodology of counting the homeless. In *Proc. Surv. Res. Meth. Sect.* (pp. 170–175). American Statistical Association.

Crosbie, S. F., Manly, B. F. J. (1981). Parsimonious modeling of capture-mark-recapture studies. *Biometrics, 41*, 385–398.

Darroch, J. N. (1958). The multiple-recapture census. I. Estimation of a closed population. *Biometrika 45*, 343–359.

David, B., & Snijders, T. A. B. (2002). Estimating the size of the homeless population in Budapest, Hungary. *Quality & Quantity, 36*, 291–303.

Demographic and Social Statistics Branch, United Nations Statistics Division. (2009). *Manual on census evaluation. Post enumeration surveys*. United Nations.

Dorazio, R. M., & Royle, J. A. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics, 59*, 351–364.

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika, 63*, 435–447.

Esty, W. W. (1982). Confidence intervals for the coverage of low coverage samples. *The Annals of Statistics, 10*, 190–196.

Esty, W. W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *The Annals of Statistics, 11*, 905–912.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika, 59*, 591–603.

Fienberg, S. E., & Manrique-Vallier, D. (2009). Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. *Advances in Statistical Analysis, 93*, 49–60.

Fisher, N., Turner, S. W., Pugh, R., & Taylor, C. (1994). Estimated numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis. *BMJ 308*, 27–30.

Foote, M., & Raup, D. M. (1996). Fossil preservation and the stratigraphic ranges of taxa. *Paleobiology, 22*, 121–140.

Garamszegi, L. Z., Boulinier, T., Moller, A. P., Torok, J., Michl, G., & Nichols, J. D. (2002). The estimation of size and change of avian song repertoires. *Animal Behaviour, 36*, 623–630.

Geissler, P. H., & Fuller, M. R. (1987). Estimation of the proportion of area occupied by an animal species. In *Proc. Sect. Surv. Res. Meth. Amer. Stat. Assoc.* (pp. 533–538).

Greene, M. A. (1984). Estimating the size of a criminal population using an open population approach. In *Proc. Amer. Stat. Assoc. Surv. Res. Meth. Sec.* (pp. 8–13).

Greene, M. A., & Stollmack, S. (1981) Estimating the number of criminals. In J. A. Fox (Ed.), *Models in quantitative criminology*. Academic Press.

Haas, P. J., & Stokes, S. L. (1998). Estimating the number of classes in a finite population. *Journal of the American Statistical Association, 93*, 1475–1487.

Haas, P. J., Liu, Y., & Stokes, L. (2006). An estimator of number of species from quadrat sampling. *Biometrics, 62*, 135–141.

Hendrix, C. S., & Saleyhan, I. (2015). No news is good news: Mark and recapture for event data when reporting probabilities are less than one. *International Interactions, 41*, 392–406.

Herendeen, D. L., & White, G. C. (2013). Statistical estimates of rare stamp populations. *The History of Science and Technology at the Smithsonian, 57*, 91–100.

Hestbeck, J. B., Nichols, J. D., & Malecki, R. A. (1991). Estimates of movement and site fidelity using mark-resight data of wintering Canada geese. *Ecology, 72*, 523–533.

Hickman, M., Cox, S., Harvey, J. Howes, S., Farrell, M., Frischer, M., Stimson, G., Taylor, C., & Tilling, K. (1999). Estimating the prevalence of problem drug use in inner London: A discussion of three capture-recapture studies. *Addiction, 94*,1653–1662.

Hines, J. E. (2006). *PRESENCE 3.1 Software to estimate patch occupancy and related parameters.* http://www.mbr-pwrc.usgs.gov/software/presence.html

Holst, L. (1981). Some asymptotic results for incomplete multinomial or Poisson samples. *Scandinavian Journal of Statistics, 8*, 243–246.

Hook, E. B., & Regal, R. R. (1982). Validity of Bernoulli census, log-linear, and truncated binomial models for correction for underestimates in prevalence studies. *American Journal of Epidemiology, 116*, 168–176.

Hook, E. B., & Regal, R. R. (1992). The value of capture-recapture methods even for apparently exhaustive surveys: The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *American Journal of Epidemiology, 135*, 1060–1067.

Hook, E. B., & Regal, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews, 17*, 243–264.

Hook, E. B., & Regal, R. R. (1999). Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *Journal of Clinical Epidemiology, 52*, 917–933.

Hook, E. B., Albright, S. G., & Cross, P. K. (1980). Use of Bernoulli census and log-linear methods for estimating the prevalence of spina bifida in livebirths and the completeness of vital record reports in New York State. *American Journal of Epidemiology, 112*, 750–758.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663–685.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika, 76*, 133–140.

Huggins, R. M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics, 47*, 725–732.

International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995a). Capture-recapture and multiple-record systems estimation. I: History and theoretical development. *American Journal of Epidemiology, 142*, 1047–1058.

International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995b). Capture-recapture and multiple-ecord systems estimation. II: Applications in human diseases. *American Journal of Epidemiology, 142*, 1059–1068.

Jackson, C. H. N. (1933). On the true density of tsetse flies. *Journal of Animal Ecology, 2*, 204–209.

Jackson, C. H. N. (1939). The analysis of an animal population. *Journal of Animal Ecology, 8*, 238–246.

Jennelle, C. S., Cooch, E. G., Conroy, M. J., & Senar, J. C. (2007). State specific detection probabilities and disease prevalence. *Ecological Applications, 17*, 154–167.

Jewell, W. S. (1985). Bayesian estimation of undetected errors. In J. M. Bernardo, M. H. DeGroot, D.V. Lindley, A.F.M. Smith (Eds.), *Bayesian statistics* (Vol. 2, pp. 663–671). Elsevier.

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika, 52*, 225–247.

Kendall, W. L., Hines, J. E., & Nichols, J. D. (2003). Adjusting multi-state capture-recapture models for misclassification bias: Manatee breeding proportions. *Ecology, 84*, 1058–1066.

Kendall, W. L., Langtimm, C. A., Beck, C. A., & Runge, M. C. (2004). Capture-recapture analysis for estimating manatee reproductive rates. *Marine Mammal Science, 20*, 424–437.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, 115*, 700–721.

King, R., Bird, S. M., Overstall, A., Hay, G., & Hutchinson, S. J. (2013). Injecting drug users in Scotland, 2006: Listing, number, demography, and opiate-related death-rates. *Addiction Research & Theory, 21*, 235–246.

King, R., Bird, S. M., Overstall, A., Hay, G., & Hutchinson, S. J. (2014). Estimating prevalence of injecting drug users and associated heroin-related death-rates in England using regional data and incorporating prior information. *Journal of the Royal Statistical Society: Series A, 177*, 1–28.

Kremers, W. K. (1988). Estimation of survival rates from a mark-recapture study with tag loss. *Biometrics, 44*, 117–130.

Lachish, S., Gopalaswamy, A. M., Knowles, S. C. L., & Sheldon, B. C. (2012). Site-occupancy modelling as a novel framework for assessing test sensitivity and estimating wildlife disease prevalence from imperfect diagnostic tests. *Methods in Ecology and Evolution, 2012*, 339–348.

Laplace, M. (1786). Sur les naissances, les mariages et les mortes. *Historie Academie Royale Science. Ann'ee, 1783*, 693–702.

LaPorte, R. E. (1994). Assessing the human condition: Capture-recapture techniques. *BMJ, 308*, 5–6.

LaPorte, R. E., McCarty, D. J., Tull, E. S., & Tajima, N. (1992). Counting birds, bees, and NCDs. *Lancet, 339*, 494–495.

Larson, A., Stevens, A., & Wardlaw, G. (1994). Indirect estimates of 'hidden' populations: Capture-recapture methods to estimate the numbers of heroin users in the Australian Capital Territory. *Social Science & Medicine 39*, 823–831.

Laska, E. M., & Meisner, M. (1993). A plant-capture method for estimating the size of a population from a single sample. *Biometrics, 49*, 209–220.

Lawing, A. M., Blois, J. L., Maguire, K. C., Goring, S. J., Wang, Y., & McGuire, J. L. (2021). Occupancy models reveal regional differences in detectability and improve relative abundance estimations in fossil pollen assemblages. *Quaternary Science Reviews, 253*, 106747.

Lebreton, J.-D., Nichols, J. D., Barker, R., Pradel, R., & Spendelow, J. (2009). Modeling individual animal histories with multistate capture-recapture models. *Advances in Ecological Research, 41*, 87–173.

Lee, A. J., Seber, G. A. F., Holden, J. K., & Huakau, J. T. (2001). Capture-recapture, epidemiology, and list mismatches: Several lists. *Biometrics, 57*, 707–713.

Lincoln, F. C. (1930). Calculating waterfowl abundance on the basis of banding returns. U.S. Dept. Agric., Circ. No. 118, Washington, D.C.

Liow, L. H. (2013). Simultaneous estimation of occupancy and detection probabilities: An illustration using Cincinnatian brachiopods. *Paleobiology, 39*, 193–213.

Liow, L. H., & Nichols, J. D. (2010). Estimating rates and probabilities of origination and extinction using taxonomic occurrence data: Capture-mark-recapture (CMR) approaches. In J. Alroy & G. Hunt (Eds.), *Quantitative methods in paleobiology* (pp. 81–94). The Paleontological Society.

Lum, K., Price, M. E., & Banks, D. (2013). Applications of multiple systems estimation in human rights research. *American Statistician, 67*, 191–200.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., & Langtimm, C. A. (2002). Estimating site occupancy when detection probabilities are less than one. *Ecology, 83*, 2248–2255.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization and local extinction probabilities when a species is not detected with certainty. *Ecology, 84*, 2200–2207.

MacKenzie, D. I., Nichols, J. D., Seamans, M. E., & Gutierrez, R. J. (2009). Dynamic models for problems of species occurrence with multiple states. *Ecology, 90*, 823–835.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K.H., Bailey, L. A., & Hines, J. E. (2018). *Occupancy modeling and estimation* (2nd ed.). Academic Press.

Magnusson, W. E., Caughley, G. J., & Grigg, G. C. (1978). A double-survey estimate of population size from incomplete counts. *Journal of Wildlife Management, 42*, 174–176.

Manly, B. F. J., & Parr, M. J. (1968). A new method of estimating population size, survivorship, and birth rate from capture-recapture data. *Transactions of the Society for British Entomology, 18*, 81–89.

Manrique-Vallier, D., Price, M. E., & Gohdes, A. (2013). Multiple systems estimation techniques for estimating casualties in armed conflicts. In T.B. Seybolt, J. D. Aronson, B. Fischhoff (Eds.), *Counting civilian casualties: An introduction to recording and estimating nonmilitary deaths in conflict* (pp. 165–182). Oxford Univ. Press.

Marescot, L., Lyet, A., Singh, R., Carter, N., & Gimenez, O. (2019). Inferring wildlife poaching in southeast Asia with multispecies dynamic occupancy models. *Ecography, 43*, 239–250.

Mastro, T. D., Kitayaporn, D., & Weniger, B. G. (1994). Estimating the number of HIV-infected injection drug users in Bangkok: A capture-recapture method. *American Journal of Public Health, 84*, 1094–1099.

Maxim, L. D., & Harrington, L. (1982). Scale-up estimators with size-dependent detection. *Photogrammetric Engineering and Remote Sensing, 48*, 1271–1287.

Maxim, L. D., & Harrington, L. (1983). Aerial survey design: A systems-analytic perspective *Photogrammetric Engineering and Remote Sensing, 49*, 1425–1435

Maxim, L. D., Harrington, L., & Kennedy, M. (1981). A capture-recapture approach for estimation of detection probabilities in aerial surveys. *Photogrammetric Engineering and Remote Sensing, 47*, 779–788.

McCarty, D. J., Tull, E. S., Moy, C. S., Kwoh, C. K., & LaPorte, R. E. (1993). Ascertained corrected rates: Applications of capture-recapture methods. *International Journal of Epidemiology, 22*, 559–565.

McClintock, B. T., Nichols, J. D., Bailey, L. L., MacKenzie, D. I., Kendall, W. L., & Franklin, A. B. (2010). Seeking a second opinion: Uncertainty in wildlife disease ecology. *Ecology Letters, 13*, 659–674.

McClintock, B. T., Bailey, L. L., Dreher, B. P., & Link, W. A. (2014). Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity, and misidentification. *The Annals of Applied Statistics, 8*, 2461–2484.

Miller, D. A. W., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. (2011). Improving occupancy estimation when two types of observational error occur: Nondetection and species misidentification. *Ecology, 92*, 1422–1428.

Miller, D. A. W., Talley, B. L., Lips, K. R., & Grant, E. H. C. (2012). Estimating patterns and drivers of infection prevalence and intensity when detection is imperfect and sampling error occurs. *Methods in Ecology and Evolution, 2012*, 850–859.

Miller, D. A. W., Nichols, J. D., Gude, J. A., Rich, L. N., Podruzny, K. M., Hines, J. E., & Mitchell, M. S. (2013). Determining occurrence dynamics when false positives occur: Estimating the range dynamics of wolves from public survey data. *PLoS One 8*. https://doi.org/10.1371/journal.pone.0065808

Moore, J. F., Mulindahabi, F., Masozera, M. K., Nichols, J. D., Hines, J. E.,Turikunkiko, E., & Oli, M. K. (2017). Are ranger patrols effective in reducing poaching-related threats in protected areas? *Journal of Applied Ecology, 2017*, 1–9.

Moore, J., Uzabaho, E., Musana, A., Uwingeli, P., Hines, J., & Nichols, J. (2021). What is the effect of poaching activity on wildlife species? *Ecological Applications, 31*, e02397.

Nichols, J. D. (2019). Confronting uncertainty: Contributions of the wildlife profession to the broader scientific community. *Journal of Wildlife Management, 83*, 519–533.

Nichols, J. D., & Hines, J. E. (1993). Survival rate estimation in the presence of tag loss using joint analysis of capture-recapture and resighting data. In J.-D. Lebreton, P.M. North (Eds.), *The study of bird population dynamics using marked individuals* (pp. 229–243). Birkhauser Verlag.

Nichols, J. D., Hines, J. E., MacKenzie, D. I., Seamans, M. E., & Gutierrez, R. J. (2007). Occupancy estimation with multiple states and state uncertainty. *Ecology, 88*, 1395–1400.

Nichols, J. D., & Karanth, K. U. (2002). Statistical concepts: Assessing spatial distributions. In K.U. Karanth, J.D. Nichols (Eds.), *Monitoring tigers and their prey. A manual for wildlife managers, researchers, and conservationists* (pp. 29–38). Centre for Wildlife Studies.

Nichols, J. D., & Pollock, K. H. (1983). Estimating taxonomic diversity, extinction rates and speciation rates from fossil data using capture-recapture models. *Paleobiology, 9*, 150–163.

Nichols, J. D., Morris, R. W., Brownie, C., & Pollock, K. H. (1986). Sources of variation in extinction rates, turnover and diversity of marine invertebrate families during the Paleozoic. *Paleobiology, 12*, 421–432.

Nichols, J. D., Kendall, W. L., Hines, J. E., & Spendelow, J. A. (2004). Estimation of sex-specific survival from capture-recapture data when sex is not always known. *Ecology, 85*, 3192–3201.

Nichols, J. M., Judd, K. P., Olsen, C. C., Waterman, J. R., & Nichols, J. D. (2013a). Estimating detection and identification probabilities for maritime target acquisition. *Applied Optics, 52*, 2531–2545.

Nichols, J. M., Hines, J. E., & Nichols, J. D. (2013b). Selecting among competing models of electro-optic, infrared camera range performance. *Optical Engineers, 52*, 113108.

Nichols, J. D., Hollman, T., & Grand, J. B. (2017). Monitoring for the management of disease risk in animal translocation programmes. *EcoHealth, 14*, S156–S166.

Nichols, J. D., Bogich, T. L., Howerton, E., Bjrnstad, O. N., Borchering, R. K., Ferrari, M., Haran, M., Jewell, C., Pepin, K. M., Probert, W. J. M., Pulliam, J. R. C., Runge, M. C., Tildesley, M., Viboud, C., & Shea, K. (2021). Strategic testing approaches for targeted disease monitoring can be used to inform pandemic control and vaccine rollout programs. *PLoS Biology, 19*, e3001307. https://doi.org/10.1371/journal.pbio.3001307

Norris III, J. L., & Pollock, K. H. (1995). A capture-recapture model with heterogeneity and behavioral response. *Environmental and Ecological Statistics, 2*, 305–313.

Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs 62*, 3–135.

Pollock, K. H. (1981). Capture-recapture models for age-dependent survival and capture rates. *Biometrics, 37*, 521–529.

Pollock, K. H., Winterstein, S. R., Bunck, C. M., & Curtis, P. D. (1989). Survival analysis in telemetry studies: The staggered entry design. *Journal of Wildlife Management, 53*, 7–15.

Polonsky, J. A., Bhning, D., Keita, M., Ahuka-Mundeke, S., Nsio-Mbeta, J., Abedi, A. A., Mossoko, M., Estill, J., Keiser, O., Kaiser, L., Yoti, Z., Sangnawakij, P., Lerdsuwansri, R., & Vilas, V. J. D. R. (2018). Novel application of capture-recapture methods to estimate the completeness of contact tracing during a large outbreak of Ebola Virus Disease, Democratic Republic of Congo, 2018–2020. *International Journal of Infectious Diseases, 116*, S98.

Pradel, R. (1996). Utilization of capture-mark-recapture for the study of recruitment and population growth rate. *Biometrics, 52*, 703–709.

Pradel, R. (2005). Multievent: An extension of multistate capture-recapture models to uncertain states. *Biometrics, 61*, 442–447.

Razzak, J. A., & Luby, S. P. (1998). Estimating deaths and injuries due to road traffic accidents in Karachi, Pakistan, through the capture-recapture method. *International Journal of Epidemiology, 27*, 866–870.

Rivest, L.-P., & Lvesque, T. (2001). Improved log-linear model estimators of abundance in capture-recapture experiments. *Canadian Journal of Statistics, 29*, 555–572.

Rosenzweig, M. L., & Duek, J. L. (1979). Species diversity and turnover in an Ordovician marine invertebrate assemblage. In G.P. Patil & M.L. Rosenzweig (Eds.), *Contemporary quantitative ecology and related ecometrics. Statistical ecology series* (Vol. 12, pp. 109–119). International Cooperative Publishing House.

Royle, J. A. (2004). Modeling abundance index data from anuran calling surveys. *Conservation Biology, 18*, 1378–1385.

Royle, J. A., & Link, W. A. (2005). A general class of multinomial mixture models for anuran calling survey data. *Ecology, 86*, 2505–2512.

Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology, 87*, 835–841.

Sango, H. A., Testa, J., Meda, N., Contrand, B., Traoré, M. S., Stacini, P., & Lagarde, E. (2016). Mortality and morbidity of urban road traffic crashes in Africa: Capture-recapture estimates in Bamako, Mali, 2012. *PLoS ONE, 11*(2), e0149070. https://doi.org/10.1371/journal.pone.0149070

Schnabel, Z. E. (1938). The estimation of the total fish population of a lake. *The American Mathematical Monthly, 45*, 348–352.

Schwarz, C. J., & Arnason, A. N. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics, 52*, 860–873.

Schwarz, C. J., Schweigert, J. F., & Arnason, A. N. (1993). Estimating migration rates using tag recovery data. *Biometrics, 49*, 177–193.

Seber, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika, 52*, 249–259.

Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters*. MacMillan.

Seber, G. A. F., & Schofield, M. R. (2019). *Capture-recapture: Parameter estimation for open animal populations*. Springer.

Seber, G. A. F., Huakau, J. T., & Simmon, D. (2000). Capture-recapture, epidemiology, and list mismatches: Two lists. *Biometrics, 56*, 1227–1232.

Sekar, C. C., & Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association, 44*, 101–115.

Sharma, K., Wright, B., Joseph, T., & Desai, N. (2014). Tiger poaching and trafficking in India: Estimating rates of occurrence and detection over four decades. *Conservation Biology, 179*, 33–39.

Silverman, B. (2014). Modern slavery: An application of multiple systems estimation Home Office. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/386841/Modern_Slavery_an_application_of_MSE_revised.pdf

Stokes, S. L. (1984). The Jolly-Seber method applied to age-stratified populations. *Journal of Wildlife Management, 48*, 1053–1059.

Sudman, S., Sirken, M. G., & Cowan, C. D. (1988). Sampling rare and elusive populations. Science 240, 991–996 (1988)

Tercero, F., & Andersson, R. (2004). Measuring transport injuries in a developing country: An application of the capture-recapture method. *Accident Analysis and Prevention, 36*, 13–20.

U.S. Census Bureau. Post Enumeration Surveys. U.S. Census Bureau, Washington, DC (2021). https://www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/pes.html

Vaissade, L., & Legleye, S. (2008). Capture-recapture estimates of the local prevalence of problem drug use in six French cities. *European Journal of Public Health, 19*, 32–37.

van der Heijden, P. G. M., Cruyff, M., & Bohning, D. (2014). Capture recapture to estimate criminal populations. In G. J. N. Bruinsmaand, D. L. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 267–278). Springer.

Veran, S., Kleiner, K. J., Choquet, R., Collazo, J., & Nichols, J. D. (2012). Modeling habitat dynamics accounting for possible misclassification. *Landscape Ecology, 27*, 943–956.

Viallefont, A., & Auget, J. L. (1999). [Using capture-recapture models to estimate transition rates between states in interval-censored data] (published in French). *Revue D'epidemiologie et de Sante Publique, 47*, 627–634.

Wang, X., Lim, J., & Stokes, S. L. (2006). Forming post-strata via Bayesian treed capture-recapture models. *Biometrika, 93*, 861–876.

White, G. C., & Burnham, K. P. (1999). Program MARK: Survival estimation from populations of marked animals. *Bird Study, 46*(sup1), S120–S139.https://doi.org/10.1080/00063659909477239

White, G. C., Anderson, D. R., Burnham, K. P., & Otis, D. L. (1982). *Capture-recapture and removal methods for sampling closed populations*. Los Alamos National Laboratory LA-8787-NERP, Los Alamos, New Mexico.

Williams, B. K., Nichols, J. D., & Conroy, M. J. (2002). *Analysis and management of animal populations*. Academic Press.

Wittes, J. T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association, 69*, 93–97.

Wittes, J. T., & Sidel, V. W. (1968). A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases, 21*, 287–301.

Wittes, J. T., Coulton, T., & Sidel, V. W. (1974). Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases, 27*, 25–36.

Wolter, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association, 81*, 338–346.

Wolter, K. M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics, 46*, 157–162.

Yeo, L. M., McCrea, R. S., & Roberts, D. L. (2017). A novel application of mark-recapture to examine behaviour associated with the online trade in elephant ivory. *PeerJ, 5*, e3048.https://doi.org/7717/peerj.3048

Advances in the Use of Capture-Recapture Methodology in the Estimation of U.S. Census Coverage Error

Mary H. Mulry and Vincent T. Mule Jr.

Abstract A post-enumeration survey (PES) is an important tool for assessing the quality of a census and gaining information about how to improve census-taking methodology. The U.S. Census Bureau has implemented a PES to evaluate the coverage error in each U.S. Decennial Census since 1980. A PES uses a second enumeration implemented on a sample basis after a census and subsequently matched to the census using a combination of computer and clerical matching. Then, dual system estimation may be used to estimate the population size. The difference between the PES estimate of the population size and the census total yields an estimate of the net undercount. This chapter focuses on the methodology and estimation of net coverage error in the 2010 Census produced by the 2010 PES. The evaluations of U.S. censuses continue to use the PES methodology to evaluate the coverage of the decennial census. These implementations of the PES have built on the quality control methodology that Dr. Stokes developed for the 1990 PES.

1 Introduction

A post-enumeration survey (PES) is an important tool for assessing the quality of a census and for gaining information about how to improve census-taking methodology. The U.S. Census Bureau has implemented a PES to evaluate the coverage error in each U.S. Decennial Census since 1980. There are two types of coverage error. One type is *overcount*, which occurs when an enumeration is inappropriate, such as entries that are duplicates of other enumerations, for people

Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau

M. H. Mulry (✉) · V. T. Mule Jr.
U.S. Census Bureau, Suitland, MD, USA
e-mail: mary.h.mulry@census.gov

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_5

born after Census Day or for people who died before Census Day. The other type is *undercount*, which occurs when a person who should be counted in the census is not enumerated. The *net coverage error*, which equals the *overcount* minus the *undercount*, provides a measure of the quality of a census.

A PES uses a second enumeration implemented on a sample basis after a census and subsequently matched to the census using a combination of computer and clerical matching. Then, dual system estimation, which is another name for capture-recapture estimation, may be used to produce an estimate of the population size. The difference between the PES estimate of the population size and the census total yields an estimate of the net coverage error. A PES that uses dual system estimation essentially applies a variation of the "capture-recapture" methodology designed for estimating the size of wildlife populations to human populations.

This chapter focuses on the methodology and estimation of net undercount in the 2010 Census produced by the 2010 PES. The data collection methods included new quality control procedures and an estimation approach that differed from the estimation used in the prior PES programs conducted from 1980 through 2000. The implementation of the 2020 PES used essentially the same methodology for data collection and estimation as that employed for the 2010 PES. However, the COVID-19 pandemic resulted in some unexpected delays in the 2020 PES data collection and processing. As a result, the estimates from the 2020 PES will not be available in time to meet the publication deadline for this volume.

Dr. S. Lynne Stokes contributed to the methodology for data collection and estimation for the Post-Enumeration Survey at different points in her career. The discussion of the PES methodology and the evolution of its implementation to evaluate census coverage at the U.S. Census Bureau will include descriptions of her contributions.

The discussion in this document focuses on the evolution of design of the PES as implemented to evaluate the coverage of the decennial censuses conducted from 1980 to 2010. These topics include the following:

- Section 1 is the introduction to the document.
- Section 2 has a brief overview of the recognition that there was a need to evaluate the coverage of the U.S. Censuses.
- Section 3 contains a description of the dual system estimator (DSE) that is used in estimating the coverage error in censuses, including the first attempt to implement a PES aimed at evaluating the coverage of the 1980 Census.
- Section 4 describes the 1990 PES and the role Dr. Stokes played in the evaluation program that informed a decision on whether to use PES estimates to adjust the 1990 Census for coverage error.
- Section 5 explains how the 2000 PES was designed to evaluate the coverage of the 2000 Census and describes evaluations used in the decision on whether the 2000 Census should be adjusted for coverage error.
- Section 6 discusses methodological challenges in the 2010 PES.

- Section 7 describes current research on developing methods for replacing data collected in PES fieldwork with administrative records and third-party data for the US population in DSEs to produce census coverage estimates.
- Section 8 is a summary.

2 Background

The U.S. Constitution requires that a census of the U.S. population be conducted every ten years for the purpose of the apportionment of seats in the House of Representatives among the states. Article 2, Section 2 of the Constitution, states that the "actual enumeration" be used to allocate the seats among the states. The current apportionment method, which was chosen by the House of Representatives, is the Method of Equal Proportions, but other methods have been used over the years (Spencer, 1985).

The first U.S. Census was conducted in 1790. As Secretary of State, Thomas Jefferson's duties included certifying the 1790 Census data. Even though Secretary Jefferson certified the census count, both he and President Washington thought the 1790 Census had undercounted the U.S. population by several hundred thousand (U.S. Census Bureau, 2021a). For years, the prevailing attitude was that the census provided the best information about the size and distribution of the U.S. population. And, even if the census was not perfect, certainly it had better coverage of the population than any source of administrative records available at the time.

New information about the coverage of the census appeared in the early 1940s when the Census Bureau conducted a study that compared the number of males of military age in the 1940 Census to the number found in draft registration records. The study used the demographic method of comparing aggregated totals constructed by a clerical operation. The study estimated that there were 14.9% more Black males of 21–35 years of age registered for the draft than were counted in the census and 2.8% more non-Black males in the same age category (Price, 1947).

This result led to the development of census coverage evaluation methodologies, the first one being Demographic Analysis. The estimates produced by Demographic Analysis are a sum of totals for subpopulations based on aggregating administrative records from different record sources, such as birth and death records, to form an estimate of the total population that can be compared to the total from a census. The 1950 Census was the first census to have its coverage evaluated using Demographic Analysis (Coale, 1955). Demographic Analysis has been used to evaluate the coverage of every U.S. Census at the national level since 1950 and is still used today although the method and data sources have improved over the years. Demographic Analysis does not produce estimates for subnational geographic areas such as states and has limited race results since it uses historical data sources.

The need for estimates of census coverage for geographic and demographic subgroups led to the development of two other methods. One is the Post-Enumeration Survey (PES) used by the USA and several other countries (Mulry, 2014). The other is the reverse record check, developed by the Statistics Canada which relies on an administrative record system that is updated on an ongoing basis between censuses (Statistics Canada, 2007, 2021). One of the innovations in the estimation of the coverage of the 2010 Census came from the PES estimation procedure incorporating the Demographic Analysis results for some hard-to-count subgroups. Section 6 gives the details.

3 Post-Enumeration Survey in 1950, 1960, and 1980

This section provides an overview of the Census Bureau's initial attempts to implement the PES to evaluate the coverage of the 1950, 1960, and 1980 Censuses. A more detailed discussion appears in Mulry (2012).

A Post-Enumeration Survey (PES) is a survey conducted after a census for the purpose of evaluating the coverage of the census. The first PES in the USA was conducted after the 1950 Census and was motivated by the undercount of draft-age males discovered in the 1940 Census. A PES uses two systems, which may be samples. The Census Bureau's implementation uses samples where one is a sample of the population, called the P sample, and the other is a sample of census enumerations, called the E sample. The basic strategy is that enumerators conduct interviews at the addresses in the P sample that include collecting the current household roster along with characteristics and where each person resided on Census Day plus a roster of the people living at the address on Census Day. Then a clerical operation matches the people on the P sample roster at each address in the P sample to the Census enumerations in two phases. In the first phase, those in the P sample that match to a census enumeration at the reported Census Day address receive a status of Match. When the matching operation cannot decide, the person receives a status of Unresolved, and the form is sent back to the field for interviewers to collect more information. The E sample enumerations also receive one of three statuses, Correct Enumeration, Erroneous Enumeration (if person was not a resident at the address on Census Day), or Unresolved. When P sample and E sample people receive an Unresolved status, their forms are sent for further fieldwork to determine each person's Census Day address. If the interviewer conducting the second interview is unable to determine where the person lived on Census, the person retains the status of Unresolved. Each census enumeration that retains an Unresolved status receives an imputed probability of being a Correct Enumeration, and P sample people with an Unresolved status receive an imputed probability of being a Match.

The methodology for collecting and processing the data that the PES collects has evolved over the years. The changes include almost all aspects, such as how the samples are selected, how the P and E sample interviews are implemented, the use of technology, and the estimation approach. Section 3.1 contains a short discussion of the Census Bureau's first attempts in 1950 and 1960 to conduct a PES, and Sect. 3.2 discusses the implementation of the 1980 PES to evaluate the 1980 Census. The

Census Bureau did not conduct a PES after the 1970 Census. The Census Bureau did implement a PES after the 1990 Census and subsequent censuses, and Sects. 4, 5, and 6 contain discussions of these implementations.

3.1 PES in 1950 and 1960

The first attempt to implement a PES was aimed at evaluating the coverage of the 1950 Census (Marks et al., 1953). The P and E samples each had about 25,000 housing units and were selected in a manner that resulted in the chosen areas overlapping as much as possible to reduce the expense of the data collection. The strategy was for the P sample interview to be of much higher quality than the census interview so that the error could be estimated by comparing the results of the P sample to the census results in the E sample. When the P sample results did not agree with the census for the same housing unit, interviewers were sent to collect information to resolve the discrepancies so that errors in the P sample could be identified. Then the corrections could be incorporated into the results of the clerical matching operation.

The strategy relied on these procedures discovering the truth in the sample areas. Then an estimate of the population size could be formed by multiplying the total census count by the ratio defined by the total number of people in the P sample housing units divided by the total number of people in the census in the same housing units as shown below:

$$True\widehat{Population} = (Census\ Count)\, x \frac{\text{number of people in P sample in P} - \text{sample housing units}}{\text{number of people in census in P} - \text{sample housing units}} \tag{1}$$

Unfortunately, the results failed to meet the Census Bureau's quality standards. The PES estimate of population size was lower than the estimates derived from demographic methods (Coale, 1955). The PES estimate of undercoverage was 2.1 million persons, which was 1.4% of the enumerated population, while the demographic method estimated the undercoverage to be 5.4 million which was 3.6% of the enumerated population. The Census Bureau's analyses found that the "minimum reasonable estimate" of undercoverage was 3.7 million which was 2.5% of the enumerated population. Subsequent analyses performed in preparation for evaluating the coverage of the 1960 Census found weaknesses in both the PES data collection and estimation and also in some of the assumptions used in producing the demographic estimates (Marks & Waksberg, 1966). Another concern about the 1950 PES was that some PES interviewers did not follow instructions completely. The interviewers were given a sealed census roster for each address. The interviewer's instructions were to open the envelope after completing the PES interview and compare the new roster with the census roster. Then, while still on the doorstep, the interviewer could identify differences and ask questions to identify errors in one or both rosters. However, there were reports that some interviewers did not ask for a household roster on Census Day but only opened the census roster and verified it.

A second attempt to implement a PES was aimed at evaluating the coverage of the 1960 Census. However, the design had P and E samples that were selected independently and retained the assumption that the P sample interview would be more accurate than the census responses in the E sample. The outcome of the 1960 PES also was not satisfactory. Reminiscent of the results from the 1950 PES, the 1960 PES estimates of population size were lower than the national-level estimates derived from demographic methods (Marks & Waksberg, 1966).

3.2 PES 1980 and Dual System Estimation

The Census Bureau introduced a new design for a PES to evaluate the 1980 Census. The 1980 PES implemented a new estimation method called dual system estimation, which led to a new design for sample selection.

3.2.1 Dual System Estimation

A major part of the new design was using dual system estimation (DSE) which did not require the assumption that the data collected for the P sample was without error (Chandrasekar & Deming, 1949). The method had been used in programs sponsored by the United Nations (UN) that focused on estimating population size in other countries. Implementing the DSE, which is another name for capture-recapture, required only that the P sample be a second enumeration of the population as opposed to being a near-perfect enumeration that was required for the estimation approach used in the 1950 and 1960 PESs. The estimation approach used post-stratification, not the log-linear form of the estimator used in some applications of capture-recapture methods.

Data collection for the P and E samples must satisfy four basic assumptions (Chandrasekar & Deming, 1949). One is that selection for inclusion in the P sample is independent on selection for inclusion in the E sample. This assumption means that the census and the P sample could not share data or information. For example, a census interviewer who also worked on the data collection for the P sample had to work in areas that were not included in the interviewer's census assignments. Second, the probability of being included in the census is not correlated with being included in the P sample. Third, each individual is unique, and records for the individual can be identified on both lists without error. And fourth, there are no spurious events in the E sample list or the P sample list, which for the Census Bureau's PES means that there are no sample records that are duplicates, nonexistent, or not in the population of interest (Mule, 2008).

When the four assumptions hold, the following two ratios of expected values are equal. The ratio on the left is based on the E sample and the ratio on the right is based on the P sample. In some capture-recapture applications, the ratio on the right in Eq. (2) is called the *detection probability*:

$$\frac{E\left(Number\ of\ \widehat{correct\ census}\ enumerations\right)}{E\left(\widehat{Population}\ size\right)} \sim \frac{E\left(Number\ of\ \widehat{matched}\ people\right)}{E\left(Number\ of\ \widehat{survey}\ enumerations\right)} \tag{2}$$

A *correct census enumeration* is one where the person is enumerated at the address where the person lives and sleeps around Census Day, which is April 1 of the census year. The enumeration also is required to be *data-defined*, which means that the record has enough information to identify the person uniquely. An enumeration is classified as data-defined if it has two or more characteristics, one of which may be a name. However, sometimes a data-defined enumeration cannot be uniquely identified, such as when an enumeration with the minimum information has characteristics that are common in their area. A *matched person* is one that has a record in the P sample that can be matched to the person's census enumeration.

Using Eq. (2) and algebra, an estimator of the population size can be constructed as follows:

$$\widehat{Population}\ size = (Number\ of\ \widehat{correct\ census}\ enumerations)\frac{(Number\ of\ \widehat{survey}\ enumerations)}{(Number\ of\ \widehat{matched}\ people)} \tag{3}$$

One aspect of using samples is the need to include both small and large subpopulations, such as race and Hispanic ethnicity groups, and geographic areas such as states and metropolitan areas. Therefore, the sample selection probabilities will be higher for smaller population groups than for the larger groups. The estimation needs to account for the variation in the selection probabilities by incorporating sampling weights equal to the inverse of the selection probabilities.

The formula for the DSE based on samples uses the same formula as in Eq. (3) with the addition of a ratio adjustment of the estimated number of correct enumerations to the number of data-defined enumerations. However, the inclusion probabilities are not equal throughout the population, which affects whether Eqs. (2) and (3) hold. The remedy is to partition the population into groups where the inclusion probabilities are believed to be equal or nearly so. The groups are called *poststrata* (indexed by *j*), and the post-stratified estimator for an area *C* is shown below. The *data-defined census enumerations* are those that have enough information for the matching operation to identify them if they are in the P sample. An enumeration is classified as data-defined if it has two or more characteristics, one of which may be a name. Enumerations that are not data-defined remain in the census but are excluded from the E sample. Therefore, the matching is a three-step procedure where the first step determines if the census enumeration is data-defined, and for those that are, the second step identifies the ones that are correct enumerations, and the third step determines if the P sample person matches to a census enumeration.

The formula for the post-stratified DSE estimate of the population size for an area C, $\widehat{TOTAL_C}$, when using J poststrata is as follows:

$$\widehat{Total}_C = \sum\nolimits_{j \in J} CEN_{Cj} \left[\frac{DD_j}{CEN_j} \frac{\widehat{CE}_j \Big/ \widehat{ETOT}_j}{\widehat{M}_j \Big/ \widehat{PTOT}_j} \right] \tag{4}$$

where

$\mathrm{CEN_{Cj}}$ = number of census enumerations in poststratum j in area C
$\mathrm{CEN_j}$ = number of census enumerations in poststratum j
DD_{j} = number of data-defined enumerations in the census in poststratum j
$\widehat{ETOT_j}$ = estimated number of data-defined enumerations in the E sample in poststratum j
$\widehat{CE_j}$ = estimated number of correct enumerations in the E sample in poststratum j
$\widehat{M_j}$ = estimated number of P sample people in poststratum j that match a census enumeration in the correct location
$\widehat{PTOT_j}$ = estimated number of people in the P sample in poststratum j

3.2.2 1980 PES

The E and P samples for the PES in the 1980 Post-Enumeration Program (PEP) were both nationwide samples that were selected in completely different ways. For the 1980 PES, the P sample used the April and August waves of the current population survey (CPS) which is an ongoing nationwide survey that measures unemployment and is conducted separately from the census. The combination of the two CPS waves included about 124,000 housing units with about half from each wave. The P sample questions appeared on a supplementary questionnaire that was administered after the CPS questions and asked who resided at the address on Census Day. The E sample was constructed by selecting 10 housing units from each enumeration district in the USA which resulted in a sample size of about 110,000 (Fay, 1988). Interviewers visited each housing unit in the E sample and verified that each person listed on the census questionnaire for the address was a resident on Census Day. If the people listed on the census questionnaire had moved, the interviewer sought information about them from neighbors and at the post office (Mulry, 2012).

The clerical matching of the two samples to the census to determine who should be on each list was cumbersome and time consuming. The census file and both sample files needed to be available for the matching. The matching for those who moved between the census and the PES interviews was exceptionally time consuming.

The results of the 1980 PES showed some undercount, but there was a controversy over the best way to construct the estimate of the net undercount (U. S Census Bureau, 1980). Some statisticians inside and outside the Census Bureau were not confident that the implementation of the 1980 PES satisfied the assumptions underlying the DSE. There was a concern that the estimates based on the DSE were affected by correlation bias, so analyses assessed the impact of some of the assumptions by constructing 12 sets of estimates. In the end, the preferred set of PES estimates of net undercount were 1.0% for the USA, 5.7% for Blacks, 4.5% for non-Blacks, and 0.0% for others ((U. S Census Bureau, 1980), p. 9–10). Another concern was that the estimates of net undercount at the national level based on the PES were lower than the estimate from Demographic Analysis which was 1.2% for the USA. Other estimates of net undercount from Demographic Analyses were 4.5% for Blacks and 0.8% for non-Blacks (Long et al., 2003). The estimated net undercount prompted a call to adjust the 1980 Census for the undercount using the 1980 PES data. The Census Bureau opposed adjusting the 1980 Census using PES data and stated so in an announcement. Detroit, New York City, and the State of New York filed a lawsuit asking that the Census Bureau be ordered to adjust the 1980 Census for undercount. These lawsuits were consolidated to the court hearing the New York case. The judge ruled that the Census Bureau's decision was not arbitrary and capricious. Therefore, in the end, the 1980 Census was not adjusted (U.S. Census Bureau, 2021).

4 1990 PES

In the aftermath of the 1980 Census, the Census Bureau decided to prepare in a manner that would enable an adjustment of the results of the 1990 Census if such an adjustment was deemed necessary. The preparations included a research and testing program during the decade leading up to the 1990 Census. The program incorporated test censuses during the decade and a dress rehearsal in 1988, each including a PES. The testing program facilitated refining PES data collection, processing, and estimation methodology.

One of the components of the research and testing program for the 1990 Census was the development of computer matching software that could be used in matching the P sample records to E sample records. The goal was to improve the quality of the matching and to produce the matching results faster than was possible with clerical matching. The development of the new matching software leveraged methodology developed by Fellegi and Sunter (1969) for matching records (Jaro, 1989; Kelley, 1986). In addition, clerical staff conducted a quality control operation on a sample of the computer matching results to assure accuracy.

Another component of the research and testing program was the development of methods to assess the quality of the PES estimates which could be used to evaluate their suitability for adjusting the 1990 Census for undercount. See Hogan (1993) for an overview of the 1990 PES methodology and Belin et al. (Belin et al., 1993) for a discussion of the new approach to imputing enumeration status using hierarchical modeling.

4.1 Dr. Stokes's Contributions to Interviewer Quality Control

Dr. Stokes has made significant contributions to the study of interviewer variance and bias. Interviewer effects on data collected in censuses and surveys can be substantial. Interviewer variance was a major reason the 1970 Census started the collection of census data by mail instead of personal interview (Stokes & Mulry, 1987).

Her interest in interviewer effects and quality control started when she worked at the Census Bureau early in her career and continued during her career as an academic. Initially, her work at the Census Bureau focused on optimizing the design of quality control samples to detect interviews fabricated by interviewers (Biemer & Stokes, 1989).

Fabrications of interviews during the data collection for the estimation of census coverage error is particularly important. A reason is that one of the assumptions underlying the DSE in Sect. 3.2.1 states that the E sample list and the P sample list used in estimation do not contain spurious records, such as fabricated records. When the assumption of no spurious events in data holds, the relationship in Eq. (2) that underlies the DSE holds. Interviewing quality control is therefore essential.

Another reason that the detection of fabricated interviews is important is because the quantity being measured is very small. A relatively small number of errors have the potential for a substantial impact on the estimate. For the past eight censuses, the Census Bureau has measured an error in the census count. For example, the Census Bureau estimates that the 1990 Census count for the population was 1.6% too low and the count for Blacks was 4.4% too low. This type of difference in accuracy is called the *differential undercount*. The differential undercount is important because key uses of the census data are for fixed-sum distributions such as the apportionment of Congress, the drawing of districts for state legislatures, and the federal fund allocation programs.

When Dr. Stokes started research in nonsampling error measurement in surveys, one of her concerns was that estimation of the correlated component of response variance usually assumed a normal distribution whereas most survey data were categorical. The paper "Estimation of the Correlated Component of Response Variance for Categorical Variables" (Stokes & Mulry, 1987) subsequently showed that the assumption could cause substantial underestimation of the sample size during the design of a study to measure the effect of interviewers.

Through continuing research on interviewer effects, Dr. Stokes made significant contributions during the consideration of adjusting the 1990 Census numbers for undercount. In addition, she provided the technical expertise for the evaluation of the effect of interviewer fabrication on the quality of the estimates of undercount. This role was the culmination of the research she conducted under contract with the Census Bureau.

Her work during the evaluation of the 1990 PES focused on assessing the assumption of no fabrication of interviews in the PES data. Despite an elaborate quality control program for the interviewing of the 1990 PES, some fabrication of interviewers was detected during the research studies leading up to the 1990 Census (Stokes & Jones, 1989). One result of the research was that the single-person households were the ones most likely to have fabricated interviews. The rationale was that since there was only one household member, these were the most difficult addresses to find someone at home. Thus, interviewers would make several attempts to make contact, but if they were unsuccessful, out of frustration, the interviewer would use the name on the mailbox and fill in the rest of the information. Therefore, one recommendation at the end of the quality control evaluation was that one-person households be checked at a higher rate than households with more than one person.

4.2 *Outcome of the 1990 PES*

Estimates of the net undercount in the 1990 Census were not used to adjust the census counts although there was litigation that reached the Supreme Count. In 1999, the U.S. Supreme Court ruled in an opinion written by Chief Justice William Rehnquist that the census numbers used for the apportionment of seats in the House of Representatives could not be based on samples because the Constitution required using the "actual enumeration" from the census (Department of Commerce vs United States House, 1999). However, the opinion did not prohibit adjusting the census numbers for other uses. See Prewitt (2012) for a brief discussion of the implications of the decision.

The 1990 PES estimated the net undercount of the U.S. population to be about 1.6% or about 4.0 million people. The estimate of the net undercount for Whites was about 1.8 million people while the net undercount rate for Blacks was about 1.4 million. But because the Black population was far smaller than the White population, the percent net undercount rate of 4.4% for Blacks was higher than for the 0.9% undercount rate for Whites (U.S. Census Bureau, 2021b).

5 2000 Census Accuracy and Coverage Evaluation

After the controversy over the possibility of an adjustment of the 1990 Census counts, the Census Bureau decided to create a process for deciding in March 2001 whether the coverage error in the 2000 Census numbers warranted an adjustment for use in redistricting. The Supreme Court made its decision that prohibited adjustment for redistricting in 1999, but the planning and research for the 2000 Census had started several years earlier. These preparations continued after the Supreme Court decision in 1999 because the census numbers may have needed an adjustment for use in other Census Bureau programs. The work included developing a process and predefined criteria for deciding whether adjustment was appropriate. An evaluation program, called the 2000 Census Accuracy and Coverage Evaluation (ACE), was designed to collect and analyze data that would inform the decision. The ACE included a post-enumeration survey and other analyses.

Dr. Stokes served on the panel "Measuring a Changing Nation: Modern Methods for the 2000 Census" that was convened by the National Academy of Sciences. The panel reviewed the Census Bureau's plans for the 2000 Census and the results of Census Tests. These plans included incorporating applications of new technology in several operations. The Census Bureau sought review and advice concerning the performance of the new technology used in the collection and processing of census data and coverage measurement data in the Census Tests conducted in preparation for the 2000 Census.

Advances in technology enabled innovations in the Census Bureau's collection, processing, and analysis of the census and ACE data to be completed in time to make an adjustment decision in March 2001. Much of the technology had been available previously but had not developed to the point where census planners could count on it for implementation and processing on the large scale and short time frame required to collect data from the 115.9 million housing units in the USA (Woodward & Damon, 2001). In 2000, all census response forms, both mail and Nonresponse Followup (NRFU) operation, were scanned by optical character and mark recognition technologies and converted to electronic format for processing (Kline, 2004).

The ACE interviewers used laptop computers when collecting the PES data. Addresses for each P sample block cluster were loaded into the laptop for the interviewer assigned to the area. Interviewers then used the laptop computers to collect data from respondents. The laptops contained the entire questionnaire, and interviewers were able to transmit the collected data electronically to the processing center. The laptops enabled faster processing and analyses of the data than was possible for the previous PES implementations that used paper questionnaires followed by a keying operation.

The interviewing quality control operation also used laptops. An advantage was that the original census responses for an address could be loaded into the interviewers' laptops. After conducting a quality control interview, the interviewer was able to push a button, and the laptop would present a comparison between the census household roster and the Census Day roster provided during the quality control interview. If there were differences in the two rosters, the quality control interviewer was able to ask questions to resolve any issues while still with the respondent.

The biggest surprise from ACE was the discovery that the estimated number of duplicate enumerations in the 2000 Census was much higher than expected. In addition, mail returns that were thought to be the best responses were included in some of the duplicates that were detected. Another finding was that the duplication occurred more frequently among household members under 30 than over 30. Examples of duplicate enumerations include the following: (1) college students being counted at both their college address and their parents' address, (2) children whose parents are divorced being counted at both parent's addresses, and (3) people who move around Census Day (April 1) being counted at both their old and new addresses since census data collection for NRFU goes into summer.

The discovery of the problem with duplication occurred during implementation of the process the Census Bureau had set up to arrive at a decision on whether to adjust the 2000 Census numbers issued for a purpose other than redistricting the seats in the House of Representatives. Further investigation found that a substantial number of erroneous enumerations had gone undetected in the processing of the ACE.

The Census Bureau continued to study whether to incorporate an adjustment to the census numbers that would be used in producing the intercensal estimates and other census products. The focus was on creating another revision, called ACE Revision II, that would be based on additional research concerning the level of duplication and the possibility of an adjustment for correlation bias in the DSE. At this point, a research project, called the Statistical Administrative Records System (StARS), created with the Census Bureau's newly developed administrative records database methodology, had progressed to the point of being useful in detecting census duplicates without fieldwork. StARS was able to create a database that covered the U.S. population by merging federal administrative records. Linking E sample and P sample records to StARS aided in identifying duplicates and other enumeration errors.

In the end, ACE Revision II estimates included several adjustments. The research with StARS and a clerical matching project produced an estimate of 5.8 million duplicates that was the basis of one of the adjustments of ACE Revision II (Mulry et al., 2006). Because the Demographic Analysis estimates produced a ratio of males to females that was higher than observed in the ACE, an adjustment for correlation bias was included in the ACE Revision II estimates. The correlation bias adjustments were created separately for Blacks and non-Blacks within three age categories: 18–29, 30–49, and 50 and over. However, an adjustment was not included for non-Black males 18–29 years of age because the data did not support the estimation in this category. More details about the adjustment may be found in (Bell, 1993, 2001). In addition, errors found during several evaluation studies were corrected in the data used in forming the ACE Revision II estimates.

The 2000 ACE Revision II estimates were the first PES estimates that measured a net overcount in a census of the USA. Figure 1 displays the net coverage estimates based on the PES and Demographic Analysis methodologies for the 1980, 1990, and 2000 censuses conducted in the USA.

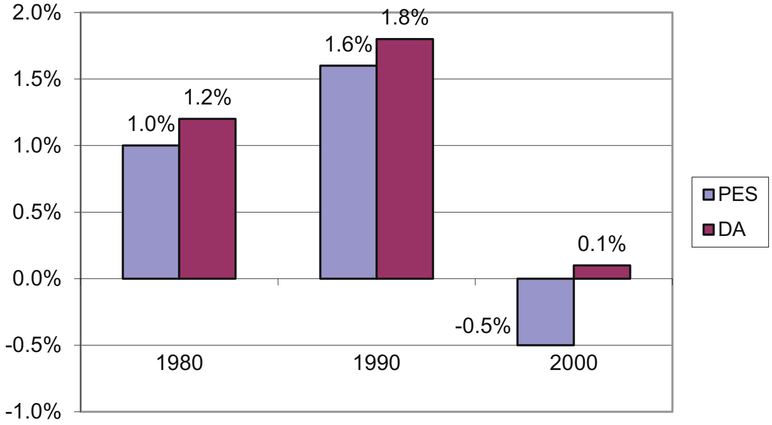


Fig. 1 Percent net undercount estimates from post-enumeration surveys and Demographic Analysis for the 1980–2000 censuses. (Source: Long, Robinson, and Gibson (2003))

6 Innovations in the 2010 PES Methodology

The Census Bureau pursued several innovations in the 2010 implementation of the PES used to evaluate the 2010 Census. The program was called the Census Coverage Measurement (CCM). The enhancements in the CCM involved producing estimates of the components of census coverage error, preparing to include a correction for correlation bias in case one was needed, and using logistic regression instead of post-stratification in the construction of the DSE to produce the estimates of census coverage error. The 2010 CCM focused on measuring the coverage of people in housing units. The CCM also evaluated the coverage of housing units, but those estimates are not discussed in this document.

6.1 Components of Census Coverage Error

One innovation used the PES data to form national-level estimates of the components of census coverage error, namely, the total number of erroneous enumerations and the total number of people missed by the census. Creating these estimates required data processing that differed in some ways from the processing for forming the DSE. For example, a different imputation procedure was needed to compensate for missing data when forming estimates of the erroneous enumerations and the people missed.

The definitions of the four components of census coverage error for persons in housing units are listed below (Mule, 2008). The estimation of the correct, erroneous, and omission components included all the data-defined enumerations in the E sample and did not require that they have a name. This section contains a high-level discussion of the approach to the estimation of the four components of census coverage error. For more details about the estimation method, see Mule (2008). Bell and Cohen (2009) also discuss the 2010 PES. Table 2 contains the estimates for the components of coverage error.

1. *Correct Enumerations.* Estimates of the number of correct enumerations in the final census count were produced at the national level using E sample data, which was a national sample of data-defined enumerations in census housing units. An enumeration was considered correct for component estimation if it was for a person who was counted once and only once in the U.S. housing unit universe. One rule was that if the person was supposed to be enumerated in a housing unit and was included in a housing unit anywhere in the USA, then that person was considered correctly enumerated. If such a person or unit was included multiple times, one of the enumerations was designated correct, and the others were classified as erroneous.

   The estimation approach used a two-stage ratio adjustment to reduce the variability of the estimates and ensure that the sum of estimates for selected subgroups added to the total. The first stage was a ratio adjustment to the E sample weights that was done by identifying cells, which were formed by using characteristics such as race/ethnicity, tenure, age/sex groupings, and then ratio adjusting the sum of the sampling weights in each cell to the total number of data-defined enumerations in the census. The second stage adjustment was applied to each of the first stage cells by a ratio adjustment to the total number of data-defined census enumerations within the cell.
2. *Erroneous Enumerations.* The E sample also was used to produce an estimate of the number of erroneous inclusions in the final census count using the same estimation approach that was used to estimate the number of correct enumerations. Erroneous inclusions consist of duplicate numerations and enumerations of people who should not have been counted in a housing unit. In addition, enumerations for persons born after Census Day and persons who died before Census Day are considered erroneous. The CCM processing identified whether the person should have been counted in the (1) same county but outside of the PES sample block cluster search area, (2) different county in the same state, or (3) different state. The erroneous enumeration estimates used the similar two-stage ratio adjustment.
3. *Whole-Person Census Imputations*. The CCM program tabulated and reported the number of whole-person imputations in housing units directly from the census. The CCM program did not evaluate whether these imputations were correct or erroneous. Whole-person imputations are the result of one or more steps that may include imputing whether a housing unit is occupied, the household size, and the characteristics of the household members.
4. *Omissions*. The CCM program created estimates of the number of omissions of people in housing units from the census. The estimation of the number of

omissions relies on the two following relationships for net error in the census count:

$$Net\ Error = TruePopulation - Census \tag{5}$$

$$Net\ Error = Omissions - Erroneous\ Enumerations \tag{6}$$

Note that Eq. (5) can be rewritten as the following:

$$Omissions = Net\ Error + Erroneous\ Enumerations. \tag{7}$$

Substituting Eq. (5) for *Net Error* in Eq. (7) and some algebra yields the following:

$$Omissions = (TruePopulation - Census) + Erroneous\ Enumerations. \tag{8}$$

Finally, substituting an estimate of the *TruePopulation* size and an estimate of the number of *Erroneous Enumerations* from the PES yields an estimator for *Omissions* as follows:

$$\widehat{Omissions} = True\widehat{Population} - Census + Erroneous\widehat{Enumerations} \tag{9}$$

6.2 Correction for Correlation Bias

Another innovation in the 2010 PES addressed the vulnerability of the DSE to correlation bias, which arises when probabilities of a person or group of people being included in the census and the PES sample are correlated. The remedy was the incorporation of an adjustment for correlation bias. A version of the ratio adjustment for correlation bias first appeared in a revision of 2000 PES estimates. The adjustment was based on the ratio of males to females for Blacks and non-Blacks based on Demographic Analysis estimates of the 2010 population size derived from birth records, death records, and estimates of immigration and emigration (Konicki, 2012; Mulry, 2014).

The correlation bias correction provides a remedy to a violation of the first assumption underlying the DSE (see Sect. 3.2), which requires that inclusion in the census is not correlated with inclusion in the P sample. However, heterogeneity in inclusion probabilities for the census or P sample or both does occur across subgroups. Some people, such as adult males ages 20–35, are hard to count and therefore have lower inclusion rates in both the census and the P sample (Mulry, 2014 p. 50–51).

6.3 Logistic Regression Instead of Post-stratification

Implementations of the PES from 1980 through 2000 used post-stratification in forming the DSEs that were used to evaluate census coverage error. The post-stratified DSE has the disadvantage of requiring an adequate number of observations to produce a reliable estimate of the population defined by a post-stratum. This requirement limits the number of subpopulations for which DSEs can be used to produce census coverage error estimates.

Research during the 1990s demonstrated that PES data collected for forming a post-stratified DSE also could be used in logistic regression models to produce the estimated probabilities needed for constructing a different form of the DSE (Haberman et al., 1998; Alho et al., 1993). This finding enabled creating estimates of population size for subgroups formed using the independent variables in the models and thereby facilitated the construction of estimates of census coverage error for these subgroups. Because the DSEs formed using logistic regression enabled constructing estimates of census coverage error for many more subgroups than were possible when using the post-stratified DSE, the Census Bureau opted to pursue implementing this approach in the 2010 PES.

The form of logistic regression estimator for the DSE uses three separate logistic regression models: one model that predicts the probability of a record being data defined, second that predicts the probability of an E sample record being a correct enumeration, and a third that predicts the probability of a P sample record matching a census record in the search area of its sample block. Then the following formula provides a PES estimate of the population size in poststratum j in area C_j:

$$\widehat{Total}_{Cj} = \sum_{i \in Cj} \left[\pi_{dd,i}{}^{\pi_{CE,i}} \Big/ \pi_{NM,i} \right] \tag{10}$$

where

$\pi_{\text{dd,i}}$ = probability of the i-th record being data defined.
$\pi_{\text{CE,i}}$ = probability of the i-th record in the E sample being a correct enumeration.
$\pi_{\text{M,i}}$ = probability of the i-th record in the P sample matching a census enumeration in the search area of its sample block cluster.

Post-stratification requires partitioning the samples into groups that are large enough to form reliable estimates, which possibly suppresses the variability of the estimated probabilities of inclusion in the E and P samples because every observation in a poststratum receives the same estimated inclusion probability. Using the three separate logistic regression models to estimate the probabilities of being data-defined, a correct enumeration, a nonmatch permits more variability and possibly reduces the risk of violating the assumption that the probability of being included in the census is not correlated with being included in the P sample. This is the second on the list of assumptions underlying the DSE given in Sect. 3.2.

6.4 Consultation with Dr. Stokes and Other Experts About 2010 PES Methodology

The Census Bureau sought review and advice about the 2010 PES estimation plans from outside experts on capture-recapture methodology and dual system estimation. They did this by engaging the Committee of Professional Associations on Federal Statistics (COPAFS) to arrange and conduct a meeting of experts, titled the Census Coverage Measurement (CCM) Workshop. At the meeting, a Census Bureau staff presentation of plans preceded a discussion of the topic that included comments on the proposed plans. The papers that Census Bureau staff prepared for the meeting are available at https://www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/pes.html.

Dr. Stokes was invited to the meeting in recognition of her expertise in capture-recapture estimation methodology and for her contributions concerning the design of the quality control operation for the Census Bureau's 1990 PES fieldwork (Stokes & Jones, 1989; Biemer & Stokes, 1989). Her assignment was to review the plans for the imputation for the estimates of two of the 2010 Census components of coverage error, erroneous enumerations, and correct enumerations. The issue was which of the two proposed methods to use for estimating the probability that a census enumeration was correct. The cell method would assign the correct enumeration rate observed for a cell to each enumeration in the cell. The logistic regression method would instead assign each enumeration a probability estimated from the model. Dr. Stokes recommended the logistic regression approach because the method for selecting independent variables for the model was more straightforward. Although the 2010 PES used the cell method, the plans for the 2020 PES imputation include using logistic regression models.

Dr. Stokes leveraged her expertise to provide useful comments on many other aspects of the plans for the 2010 PES. One suggestion grew out of a discussion of the proposed plan to fill in missing characteristics in 2010 Census enumerations by using the characteristics for the person that could be found in the 2000 Census records. Dr. Stokes suggested going a step further and to consider the 2000 Census to be administrative records and use the 2000 Census records to enumerate some households when the household at an address appears to have the same family structure and the people are ten years younger (U.S. Census Bureau, 2009). The Census Bureau adopted a variation of this proposal in the 2020 Census by using administrative records to enumerate 5.6% of the addresses in the USA (Mulry et al., 2021).

6.5 2010 PES Estimates

The Census Bureau incorporated suggestions from the experts into the plans for the 2010 PES. The results of implementing the new methodology in the 2010 PES

Table 1 National estimates of net undercount by census year from PES

		Net undercount		Percent net undercount	
Year	Census count (thousands)	Estimate (thousands)	Standard error (thousands)	Estimate (%)	Standard error (%)
2010	300,703	−36	429	−0.01	0.14
2000	273,587	−1332*	542	−0.49*	0.20
1990	248,710	3994*	488	1.61*	0.20

Source: The 2010 estimates are from Mule (2012) and the 2000 and 1990 estimates are from Kostanich (2003)

The 2000 and 2010 Census counts exclude persons in group quarters and persons in Remote Alaska

A negative net undercount or percent net undercount estimate indicates an overcount

An asterisk (*) denotes a (percent) net undercount that is significantly different from zero

The standard error estimates are model-based and based on the PES

Table 2 Estimates of the components of 2010 Census Coverage

Components of census coverage	Estimate (thousands)	Standard error (thousands)	Percent (%)	Standard error (%)
Census count	300,703	0	100.0	0
Estimates from PES				
Population size	300,667	429	100	0
Correct enumerations	284,668	199	94.7	0.1
Omissions	15,999	440	5.3	0.1
Net undercount = (PES estimate - census count)	−36	429	−0.01	0.14

Source: Mule (2012). A negative net undercount indicates an overcount

appear in Table 1, which includes the estimates of net undercount in the 2010, 2000, and 1990 censuses in the USA. The estimates of the components of census coverage, correct enumerations, erroneous enumerations, and omissions based on the 2010 PES are shown in Table 2.

7 Current Research

Census Bureau staff currently are looking at ways of improving DSE and census coverage error estimates. Data from administrative records appears to be a fertile ground for research in this area. For the 2020 Census, one innovation was the use of data from federal and third-party sources of administrative records (ARs) to create high-quality household rosters for use in enumerating some households in the 2020 Census Nonresponse Followup (NRFU) operation. The main goal of using ARs in this process was to reduce the cost of the NRFU fieldwork while maintaining its high quality. The use of AR information reduces the number of contact attempts by NRFU enumerators at addresses that were in NRFU because a self-response was not received. AR rosters were used to enumerate addresses only if a self-response was not submitted for the address during the self-response period and if one contact attempt by a NRFU enumerator failed to resolve the status of the address. See Mulry et al. (2021).

In recent years, statistical agencies in other counties have examined the potential for improving DSE estimates for subgroups and their entire populations by incorporating "known" totals from administrative record systems (Bryant & Graham, 2015; van der Heijden et al., 2018, 2020). This is feasible in countries where administrative record systems have high coverage of the population. However, some of these countries have minority groups that are poorly covered by their administrative record systems; thus, these countries are looking for ways to improve estimates for their minorities. Because the USA does not have a single source of administrative records that covers the entire population, the Census Bureau's research is focusing on ways of using these approaches where the "known" totals are from Demographic Analysis. Even though Demographic Analysis estimates are available only at the national level, the intent of the research is to gain knowledge about the strengths and weaknesses of the administrative records for future applications.

As part of the 2020 Census Program for Evaluation and Experiments (CPEX), the Census Bureau is conducting the Administrative Record Dual System Estimation Study that is building on the use of administrative records in the 2020 Census. This project seeks to determine whether administrative records and third-party data for the U.S. population can replace the data collected in PES fieldwork in DSEs. In particular, the project is examining whether the use of administrative records as the second system produces census coverage estimates that are close to the survey-based results. Using administrative records could alleviate the need to conduct the field data collection, develop clerical matching software, and pay the clerical matching personnel costs to produce the DSEs for census coverage estimates. This has the potential to reduce the cost substantially.

The Administrative Record Dual System Estimation Study builds on methods used in other countries for deriving estimates of the population size using files created by linking registers. The linking of registers may not produce a file that covers the entire target population. Van der Heijden et al. (2018) discuss an application of the expectation maximization (EM) algorithm of log-linear models that estimates the part of the population missed by the registers. A novel application creates estimates of the causes of accidents where the cause is recorded in both the police and hospital registers, but the police reporting is more accurate. The paper shows how one can use the EM algorithm to produce estimates. Van der Heijden et al. (2020) describe an application of EM in a census context by using multiple registers to estimate the size of the New Zealand Maori population. The Administrative Record Dual System Estimation project is experimenting with the use of this EM methodology to estimate the size of race and Hispanic-origin populations in the USA. Since race and Hispanic origin are available from responses to the 2020 Census and from historical administrative records, estimates of the size of subpopulations can be compared with the estimates produced by the evaluation.

8 Summary

Dr. Lynne Stokes brought her unique skill set to bear in devising methods for estimating bias due to interview fabrication in the dual system estimator used for estimating census net undercount. She gained an in-depth knowledge of capture-recapture estimation when she worked at the Fish and Wildlife Service, as demonstrated in her paper "The Jolly-Seber Method Applied to Age-Stratified Populations" (Stokes, 1984). When she moved to the Census Bureau, she learned about survey research methodology and the challenge of designing quality control samples to detect interview fabrication as demonstrated in her co-authored paper "The Optimal Design of Quality Control Samples to Detect Interviewer Cheating" (Biemer & Stokes, 1989).

Dr. Stokes applied her background to use interviewing quality control data and the evaluations of the 1990 PES to estimate the number of residual fabrications remaining in the data after the quality control operation identified and corrected some fabrications. In addition to estimating the bias at the national level, she also constructed bias estimates for geographic and demographic subpopulations. Her work on the quality of the PES data was critical to deliberations regarding the adjustment of the 1990 Census.

More importantly, the method that Dr. Stokes used in estimating the residual fabrication errors convinced the Census Bureau of the effectiveness of the quality control operation to the point that it became an accepted practice. The Census Bureau did not construct the estimate of the residual fabrication in the interview data for any of the subsequent PESs. The basic approach to the PES interviewing quality control has remained the same even though technological advancements have enabled enhancements in the operation.

Dr. Stokes has demonstrated a flair for adapting methods developed for one application to other uses. For example, she generalized her work on estimating the amount of residual fabrication in a survey data set to the problems of quality acceptance sampling in manufacturing. The paper she co-authored with her colleague Betsy Greenberg at the University of Texas entitled "Estimating Nonconformance Rate after Zero-Defect Sampling with Rectification" (1992) generated substantial interest among engineers from semiconductor manufacturing settings who adapted the method to their projects. Drs. Stokes and Greenberg next expanded their research topic to include the possibility of misclassification error in quality control operations. This type of error may cause a good batch to fail or a bad batch to pass. In their paper entitled "Repetitive Testing in the Presence of Inspection Errors," Drs. Stokes and Greenberg (2012) formulated a rule about how many times to repeatedly test a batch before considering it to fail. The rule for repetitive testing is used in manufacturing and has numerous potential applications in surveys.

The evaluations of U.S. censuses continue to use the PES methodology to evaluate the coverage of the decennial census. These implementations of the PES have built on the quality control methodology that Dr. Stokes developed for the 1990 PES.

References

Alho, J., Mulry, M. H., Wurdeman, K., & Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual system estimation. *Journal of the American Statistical Association, 88*(423), 1130–1136. https://doi.org/10.1080/01621459.1993.10476386

Belin, T., Diffendal, G. J., Mack, S., Rubin, D. R., Schafer, J. L., & Zaslavsky, A. L. (1993). Hierarchical logistic regression modeling for imputation of unresolved enumeration status in undercount estimation. With discussion and rejoinder. *Journal of the American Statistical Association, 88*(423), 1149–1159. with discussion and rejoinder 1160-1166. https://doi.org/10.1080/01621459.1993.10476388

Bell, R. M., & Cohen, M. L. (2009). *Coverage measurement in the 2010 census*. National Academy of Sciences.

Bell, W. R. (2001). *ESCAP II: Estimation of correlation bias in 2000 A.C.E. using revised demographic analysis results*. Executive Steering Committee for A.C.E. Policy II, Report No. 10. dated October 13, 2001. Washington, DC: U.S. Census Bureau.

Bell, W. R. (1993). Using information from demographic analysis in post-enumeration survey estimation. *Journal of the American Statistical Association, 88*(423), 1106–1118. https://doi.org/10.1080/01621459.1993.10476381

Biemer, P. P., & Stokes, S. L. (1989). The optimal design of quality control samples to detect interviewer cheating. *Journal of Official Statistics, 5*, 23–39. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/the-optimal-design-of-quality-control-samples-to-detect-interviewer-cheating.pdf

Bryant, J. R., & Graham, P. (2015). A Bayesian approach to population estimation with administrative data. *Journal of Official Statistics, 31*(3), 475–487. https://doi.org/10.1515/JOS-2015-0028

Chandrasekar, C., & Deming, W. E. (1949). On a method for estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association, 44*(245), 101–115. https://doi.org/10.1080/01621459.1949.10483294

Coale, A. J. (1955). The population of the United States in 1950 classified by age, sex, and color-a revision of census figures. *Journal of the American Statistical Association, 50*(1), 16–54. https://doi.org/10.1080/01621459.1955.10501249

Department of Commerce vs United States House. 98–404. Supreme Court of the U.S. 1999. https://www.law.cornell.edu/supremecourt/text/525/326.

Fay, R. E. (1988). *The coverage of the population in the 1980 census*. PHC 80-E4. Evaluation and Research Reports.1980 Census of Population and Housing. Washington, DC: U.S. Census Bureau.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*, 1183–1210. Alexandria, VA: American Statistical Association. https://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049?msclkid=bc6f6f01cfe511ec908375c560ae084b

Greenberg, B., & Stokes, S. L. (1992). Estimating nonconformance rate after zero-defect sampling with rectification. *Technometrics, 34*(2), 203–213. https://www.jstor.org/stable/1269236

Haberman, S., Jiang, W., & Spencer, B. (1998). *Development of methodology for evaluating model-based estimates of the population size for states*. NORC Working Paper Series No. WP-2021.03 (with minor updates in 2021). Chicago, IL: NORC. https://www.norc.org/PDFs/Working%20Paper%20Series/WPS_HABERMAN_2021.03.pdf

Hogan, H. (1993). The 1990 post-enumeration survey: Operations and results. *Journal of the American Statistical Association, 88*(423), 1047–1060. https://doi.org/10.1080/01621459.1993.10476374

Jaro, M. (1989) Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association, 84*, 414–420. http://dx.doi.org/10.1080/01621459.1989.10478785

Kelley, R. P. (1986). *Robustness of the Census Bureau's record linkage system*. Proceedings of the American Statistical Association, Section on Survey Research Methods. 620–624. http://www.asasrms.org/Proceedings/papers/1986_116.pdf?msclkid=89715673cfe411ecb1eed28538e496ec.

Kline, D. (2004). *Census 2000 data capture*. Census 2000 Testing, Experimentation, and Evaluation Program. Topic Report No. 3, TR-3. U.S. Census Bureau. Washington, DC. http://www.census.gov/pred/www/rpts/TR3.pdf.

Konicki, S. (2012). *2010 Census coverage measurement estimation report: Adjustment for correlation bias*. DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-11. Washington, DC: U.S. Census Bureau.

Kostanich, D. (2003). *Technical assessment of A.C.E. Revision II*, DSSD A.C.E. Revision II Memorandum Series #PP-61. Washington, DC: U.S. Census Bureau. https://www.nrc.gov/docs/ML1233/ML12335A672.pdf

Long, J. F., Robinson, J., & Gibson, C. (2003). Setting the standard for comparison: Census accuracy from 1940 to 2000. In *2003 proceedings of the American Statistical Association, section on government statistics* (pp. 2515–2524). American Statistical Association.

Marks, E. S., Mauldin, W. P., & Nisselson, H. (1953). The post-enumeration survey of the 1950 census: A case history in survey design. *Journal of the American Statistical Association., 48*(262), 220–243. https://www.jstor.org/stable/2281284

Marks, E. S., & Waksberg, J. (1966). Evaluation of the 1960 census through case-by-case checking. In *1966 proceedings of the American Statistical Association, social statistics section* (pp. 62–70). American Statistical Association.

Mule, T. (2012). *2010 census coverage measurement estimation report: Summary of estimates of coverage for persons in the United States*. DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-01. Washington, DC: U.S. Census Bureau. https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/g-series/g01.pdf

Mule, T. (2008). *2010 census coverage measurement estimation methodology.* DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-18. Washington, DC: U.S. Census Bureau. https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/ccm-workshop/2010-e-18.pdf

Mulry, M. H. (2012). Post-enumeration survey. In M. J. Anderson, C. Citro, & J. Salvo (Eds.), *Encyclopedia of the U.S. Census* (2nd ed., pp. 339–343). Sage/CQ Press.

Mulry, M. H. (2014). Measuring undercounts for hard-to-survey groups (Chapter 3). In R. Tourangeau, N. Bates, B. Edwards, T. Johnson, & K. Wolter (Eds.), *Hard-to-survey populations* (pp. 37–57). Cambridge University Press. https://doi.org/10.1017/CBO9781139381635.005

Mulry, M. H., Mule, T., Keller, A. K., & Konicki, S. (2021). *Overview of Administrative Records Modeling in the 2020 census*. 2020 Census Program Memorandum Series: 2021.10. https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/administrative-record-modeling-in-the-2020-census.pdf

Mulry, M. H., Bean, S. L., Bauder, D. M., Wagner, D., Mule, T., & Petroni, R. J. (2006). Evaluation of census duplication using administrative records. *Journal of Official Statistics, 22*, 655–679. Statistics Sweden, Stockholm, Sweden. http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/evaluation-of-estimates-of-census-duplication-using-administrative-records-information.pdf

Prewitt, K. (2012). Decennial censuses: Census 2000. In M. J. Anderson, C. Citro, & J. Salvo (Eds.), *Encyclopedia of the U.S. Census* (2nd ed., pp. 166–169). Sage/CQ Press.

Price, D. (1947). A check on underenumeration in the 1940 census. *American Sociological Review, 12*(1), 44–49.

Spencer, B. D. (1985). Statistical aspects of equitable apportionment. *Journal of the American Statistical Association, 80*, 815–822. https://doi.org/10.1080/01621459.1985.10478188

Statistics Canada. (2021). *Coverage of the 2016 census: Level and trends*. Ottawa, Ontario: Statistics Canada. https://www150.statcan.gc.ca/n1/en/pub/91f0015m/91f0015m2020003-eng.pdf?st=MOftGoal.

Statistics Canada. (2007). *2006 census technical report: Coverage*. Ottawa, Ontario: Statistics Canada. https://www12.statcan.gc.ca/census-recensement/2006/ref/rp-guides/rp/coverage-couverture/cov-couv_index-eng.cfm.

Stokes, S. L. (1984). The Jolly-Seber methodology applied to age-stratified populations. *Journal of Wildlife Management, 48*(3), 1053. https://doi.org/10.2307/3801468

Stokes, S. L. & Greenberg, B. (2012). Repetitive testing in the presence of inspection errors. *Technometrics, 37*(1), 102–111. https://doi.org/10.1080/00401706.1995.10485893

Stokes, S. L., & Jones, P. M. (1989). Evaluation of quality control procedure for the post enumeration survey. *1999 Proceedings of the Survey Research Section*, Annual Meeting of American Statistical Association. Alexandria, VA: American Statistical Association. 696–698. http://www.asasrms.org/Proceedings/papers/1989_127.pdf

Stokes, S. L., & Mulry, M. H. (1987). On the design of interpenetration experiments for categorical data items. *Journal of Official Statistics, 4*, 389–402. Statistics Sweden, Stockholm, Sweden https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/on-the-design-of-interpenetration-experiments-for-categorical-data-items.pdf

U. S Census Bureau. (1980). Chapter 9. Research, evaluation, and experiments. In *1980 procedural history*. U.S. Census Bureau. https://www2.census.gov/prod2/decennial/documents/1980/proceduralHistory/Chapter_09.pdf

U.S. Census Bureau. (2009). *Transcription of the 2010 census coverage measurement workshop, Jan 12–13*. U.S. Census Bureau.

U.S. Census Bureau. (2021a). Directors 1790 – 1820. In *Census then now*. U.S. Census Bureau. https://www.census.gov/history/www/census_then_now/director_biographies/directors_1790_-_1810.html

U.S. Census Bureau. (2021b). *1980 Overview*. U.S. Census Bureau. https://www.census.gov/history/www/through_the_decades/overview/1980.html

van der Heijden, P. G. M., Cruyff, M. J. L. F., Smith, P., Bycroft, C., Graham, P., & Matheson-Dunning, N. (2020). Multiple system estimation using covariates having missing values and measurement error: estimating the size of the Maori population in New Zealand. *arXiv*. arXiv:2007.00929 [stat.AP].

van der Heijden, P. G. M., Smith, P., Cruyff, M., & Bakker, B. (2018). An overview of population size estimation where linking registers results in incomplete covariates, with an application to mode of transport of serious road casualties. *Journal of Official Statistics, 34*(1), 239–263. https://doi.org/10.1515/JOS-2018-0011

Woodward, J., & Damon, B. (2001). *Housing characteristics: 2000*. Report No. C2KBR/01–13. Washington, DC: U.S. Census Bureau. https://www.census.gov/prod/2001pubs/c2kbr01-13.pdf#:~:text=According%20to%20Census%202000%2C%20there%20were%20115.9%20million,2000%2C%20the%20United%20States%20housing%20inventory%20increased%20by

Part II
Nonsampling Errors in Statistical Sampling

Measurement Issues in Synthesizing Survey-Item Responses

Betsy Jane Becker and Ahmet Serhat Gözütok

Abstract From questions about politics to queries about candy preferences, survey items ask about matters large and small. While statistical approaches to combining survey estimates have been well studied, less attention has been paid to matters of measurement comparability when survey items are being summarized via meta-analysis. We present an overview of this problem. Meta-analyses begin with a defined problem, and relevant studies (here, surveys) are gathered. Studies and their measures should be scrutinized for validity and comparability as part of data collection and evaluation. When summarizing survey items, meta-analysts must represent item responses using indices that are comparable across surveys. However, survey constructs and the items that tap those constructs differ in diverse ways that challenge the meta-analyst. Cook's concept of "heterogeneous irrelevancies" supports the inclusion of diverse survey items in meta-analysis, but the tasks of construct definition and operationalization are key to a successful synthesis of items. Item variation arises from many sources—differences in construct definition, wording of question stems, direction and labeling of response scales, and number and labeling of response options. We describe approaches to dealing with these features using examples from the World Database of Happiness and raise cautions for various stages of the process.

Parts of this work are derived from Ahmet Serhat Gözütok's Ph.D. dissertation entitled "Critical issues in survey meta-analysis" conducted under the supervision of Betsy Jane Becker at Florida State University.

B. J. Becker (✉)
Measurement and Statistics, Educational Psychology and Learning Systems, College of Education, Florida State University, Tallahassee, FL, USA
e-mail: bbecker@fsu.edu

A. S. Gözütok
Measurement and Evaluation in Education, Educational Sciences, Ereğli Faculty of Education, Zonguldak Bülent Ecevit University, Zonguldak, Turkey
e-mail: gozutok@beun.edu.tr

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_6

1 Overview

Surveys are ubiquitous. From polls of political leanings to academic inquiries (Fanelli, 2009) to frivolous studies of candy or soda preferences (e.g., RetailMeNot Editors, 2021), survey items ask us about matters large and small. While statistical approaches to combining quantitative results of surveys have long been of interest (e.g., Kish, 1994, 1999a; Morton, 1999), less attention has been paid to matters of measurement comparability when surveys or survey items are combined in meta-analyses. An early exception to this was Kish's (1999b, p. 131) concern over the measurement challenges faced in cumulating surveys multi-nationally. Of late the scholarship on harmonization of measures has attacked this same problem.

Meta-analyses (Glass, 1976) have the goal of summarizing the "typical" outcome of a set of studies or surveys in terms of strength, direction, and consistency of the findings. In this chapter, we present an overview of conceptual and measurement considerations underlying the synthesis or meta-analysis of survey items, and then briefly characterize the set of techniques called harmonization. We review four survey-item features that impact the quantitative synthesis of survey items. Writings on test validity, item construction, and psychometrics guide this work. To illustrate these ideas, we draw on the World Database of Happiness project by Veenhoven and colleagues (e.g., Veenhoven, 2015; Veenhoven et al., 1993).

2 Introduction to Survey Synthesis

Most of the research to date on cumulating survey results has focused on the nature of the populations to be combined and how their results should be statistically weighted. Kish (1999b) argues that the presence of national surveys (which expanded greatly in the late 1940s) led international entities such as the various agencies of the United Nations to make international comparisons, even when those might not have been statistically justifiable. This growth in cross-national work was followed by many statistical developments including the deliberate design of coordinated national-level studies, derivation of methods for post-stratification weighting, and proposals for new varieties of periodic sampling plans. Kish led the field in this arena, and in 1994 presented five types of multi-population survey designs, based on seven aspects of design. The first three of these aspects relate at least in part to measurement, which is our focus. Kish (2002) later pointed out the connections between his ideas on quantitatively cumulating surveys and meta-analysis—the enterprise of combining studies.

The early focus on statistical analyses for combinations of related surveys may have resulted in part from the fact that early syntheses of surveys estimated parameters based on identical or very similar survey questions. There was little need to consider the nature of the questions asked, avoiding many conceptual and measurement components of the synthesis process. However, such a focus necessarily leads to a narrowed selection of constructs and measures of those constructs compared to what may be seen in the broader literature.

We discuss two classes of approaches to measurement challenges in meta-analysis of surveys. One includes conceptual approaches that deal with the theoretical constructs per se and aim to formalize the meaning behind constructs of interest. van de Water et al. (1996) refer to this as conceptual harmonization. Second are statistical or psychometric approaches that primarily operate on item scale points, distributions of scores, or correlations among items that aim to measure constructs of interest. Properly covering either of these classes of approaches would require a book rather than a book chapter, so we cover only the main aspects of these approaches.

3 The Process of Meta-analysis

Meta-analysis involves the systematic collection of the results of series of related studies, and the eventual quantitative analysis of those results. The process of meta-analysis has components that parallel those of primary research (Cooper, 2017). A simple version of Cooper's steps includes

1. Problem formulation,
2. Literature search,
3. Data evaluation, including representation of study findings,
4. Data analysis,
5. Interpretation of synthesis results, and
6. Public presentation.

We focus on steps 1 and 3, because measurement issues arise primarily at these points.

3.1 Step 1: Problem Formulation for Survey Synthesis

In a typical meta-analysis, a detailed question guides the synthesis process. Meta-analyses often examine the efficacy of interventions, or strengths of relationships. A rationale should be developed for the specific question(s) asked. A successful meta-analysis is based on questions that are not so broad as to be unanswerable, or so narrow that few studies (here, few surveys) address them. In applying this consideration to the synthesis of survey items, we argue that the development of a clear question is critical so that the synthesis team does not end up with a near-infinite set of survey sources, each examining a variation of the true target topic. For example, the World Database of Happiness (WDH; found online at https://worlddatabaseofhappiness.eur.nl) has a bibliography with over 15,500 publications on the topic of happiness, and almost 23,000 distributions of responses to questions on happiness from all over the world. Such a collection of results would swamp an individual meta-analyst; that is why over 100 team members have participated in the accumulation of these results since the 1980s.[1]

The meta-analyst must specify appropriate population(s) of study, develop construct definitions, and delineate an acceptable set of operationalizations of those constructs. The process often begins with an examination of past reviews and existing research; in some fields, scoping reviews (Munn et al., 2018) provide a quick look at the extent of the literature. The creation of lists of keywords and definitions of central concepts are important tasks, as is deciding on the target populations for study, because some constructs will differ by the age, gender, or nationality of respondents.

A key part of problem formulation is to identify the constructs to be studied as the independent and dependent variables. Examination of relevant theories, brainstorming with experts in the field of interest, and use of qualitative research methods such as grounded theory (Wolfswinkel et al., 2013) may help at this stage. Even an idea as simple as happiness may vary in its meaning across cultures (Ye et al. (2015)) or age levels. When several related constructs are of interest (e.g., happiness and life satisfaction in the WDH), the meta-analyst should justify decisions to combine results across those constructs. We posit that the use of frameworks similar to the "blueprints" used in test construction can help outline the components of target constructs (e.g., content focus, behaviors that evidence presence of the construct) and guide collection of desired items. For example, a synthesis on political interest might contain items tapping both engagement/interest and active participation, for different levels of political activity such as local, state, and national campaigns.

Good problem formulation facilitates the creation of inclusion and exclusion rules that help identify appropriate sources of data to address the question of interest. Because aggregate-data meta-analyses synthesize results from completed primary studies, the problem-formulation stage differs from the planning stages of a single survey, or even a multi-site survey program. In a typical survey, measures of the desired construct are developed prior to the survey's administration. In contrast, in meta-analyses one works with existing measures, be they scales or single items. The meta-analyst may aim to gather information about a particular construct only to find it has not been sufficiently studied. For example, in a synthesis of the literature on the management of type 2 diabetes, Brown and colleagues (2016) found that few studies had measured compliance with keeping doctor's appointments.

[1] It should be noted that the WDH is not meant to serve as the source of documents for a single survey synthesis.

3.2 Step 3: Data Evaluation

Efforts to bring together information across independent studies of any kind (including surveys) bring attention to the fact that individual study authors and survey designers have generated a huge diversity of measures on the same or similar topics. This diversity leads to challenges when results are to be accumulated. While any number of authors have commented on the importance of dealing with measurement issues in combining survey results (e.g., Rao et al., 2008, p. 102; Schenker & Raghunathan, 2007, p. 1809), few provide complete solutions for the measurement challenges inherent in the process of combining surveys.

One expects a degree of diversity in outcome measures and study design across studies because the process of science (and the need to publish "new" research) pushes toward uniqueness. Diversity in measures across studies (or surveys) may be great due to differences in construct definitions, or minimal, if construct definitions, item wording, and responses options are similar. Cook (1993) has pointed out that a degree of diversity in measures of a construct can support generalizations. If a varying feature of a set of items, say, strength of item-stem wording, does not relate to how the items function (i.e., to respondent behavior), it tells us that feature is irrelevant to the construct measured. In our example, the meta-analyst would generalize across items of varying strength if wording strength does not relate to response patterns. It is important to identify potential item features at the data-evaluation stage for this reason.

In a typical aggregate-data meta-analysis, reviewers appear to rely on the primary-study authors' claims about what was measured. It is rare to share the exact instruments used in published studies. Also researcher-made measures are often used; these are nearly impossible to obtain. Thus, a great deal of trust, or perhaps mystery, can be involved in construct definition and operationalization in a typical meta-analysis.

In contrast when individual survey items are to be summarized, the exact words used in those items and their response options are obvious. However, this does not remove the necessity for the meta-analyst to assess the nature of the construct(s) tapped, and to ask whether items from different studies measure the same construct. This can be done by way of typical validity-study methods, such as by having experts examine all collected items and rate each on its centrality to the construct of interest and degree of match to the concept definition developed at step 1. These are discussed further below. Only after the constructs are clear should the meta-analyst proceed to the next step of trying to find mathematically sound ways of connecting or comparing the item responses.

4 Harmonization

In part due to this diversity, calls for coordination and (post hoc) harmonization to enable researchers to bring together diverse measures have become more frequent over time, even though this process is rarely reported (Griffith et al., 2015). Early efforts have centered on harmonization as standardization, that is, on creating comparable numerical scoring systems for variables of interest. We refer to this as statistical harmonization, in contrast to conceptual harmonization, discussed above as part of problem formulation. A search of all ProQuest databases for "harmoniz* and measures" in peer-reviewed article titles suggests that attention to harmonization of measures first appeared in the 1990s, setting aside articles on harmonization of physical/scientific indices such as pH and blood counts (Lewis, 1990), or currency-related indices (e.g., Goeltz, 1991). The term harmonization may have grown out of the extensive work on harmonization of laws, regulations, and social policies (e.g., in the European common market), such as in Holloway and Collins (1982) and many other sources.

Initial efforts aimed to make measures of demographics and socioeconomic status more comparable. These were led by ESOMAR—the European Society for Opinion and Marketing Research—which was motivated to aid market researchers (and obviously other commercial entities) to identify "the true diversity of the market place" (ESOMAR, 2003, p. 97) in Europe. Work concerned with the harmonization of measures of human social and cognitive constructs first appeared in the literatures on commerce (e.g., Quatresooz & Vancraeynest, 1992, on demographics) and medicine.

Citations on harmonization grew in the early 2000s as attention was drawn to health-related measures such as activities of daily living that might differ across countries (Nikula et al., 2003) and to measures used in the Health and Retirement Study (e.g., Angrisani & Lee, 2011) and other cross-national health surveys that followed. Interest in the standardization and simplification of measures serves the goals of such cross-national survey efforts. Having similar or related measures and scoring systems facilitates cross-national comparisons (e.g., Bech, 1992, on quality of life), and connections among measures enable researchers to administer fewer measures, thus saving time, money, and the goodwill of participants.

Harmonization may be conducted on measures meant to be of the same construct, or measures of related constructs. The goals of harmonization include avoiding duplication of measurement efforts and ensuring standardization and comparability of measures across different target populations.

4.1 *What Is Harmonization?*

Harmonization is a process of making definitions and measures of a common construct or variable comparable. While many writings on harmonization focus on the translation of numerical scores to compatible scales, we argue that harmonization must involve two components—a conceptual reckoning and clarification of what evidence is suitable to represent the construct, or conceptual harmonization, and a statistical or psychometric component that accounts for how measures of the construct have been scored. The fourth principle of the National Quality Forum (2010) states that the conceptual component should precede the decision on whether to try to statistically harmonize.

4.2 Conceptual Harmonization

The first step in this process must be consideration of the target concepts of the synthesis and any theoretical frameworks that may underlie the measures at hand. The National Quality Forum argues that harmonization must account for the population or populations to whom measures will be administered as well. This is consistent with the idea of consequential validity of any test or measure from the *Standards for Educational and Psychological Testing* (Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014). The measure should be appropriate for all populations to whom it will be administered.

Another relevant concept drawn from the Standards is that of concept underrepresentation—the idea that a measure taps into "less or more than its proposed construct" (p. 12). Assessment of concept representation would be facilitated by the use of a survey blueprint, as we describe above. Many constructs encompass a broad range of measures and diverging conceptualizations, for example, activities of daily living (ADL; Pluijm et al., 2005) and quality of life (QOL; Bech, 1992). Some fields have moved toward the use of common or "core" measures, but we argue that in syntheses, key benefits also arise from diversity in instrument use. Cook (1993) has noted that having a diverse set of measures (or study designs, or populations) assists with generalizability. In particular, if such differences are not associated with study outcomes, our findings can be stated unconditionally. Finding empirical evidence that those features do not matter means that simpler but broader conclusions can be stated. If we narrow our constructs or measures to a select few, we cannot even assess how generalizable our results might be.

Some researchers have provided conceptual models for these variously measured constructs or used frameworks such as classifications of measures (e.g., of attitudes) into affective, cognitive (Crites et al., 1994), and behavioral aspects (Ostrom, 1969). Often theoretical models can assist the meta-analyst in judging whether single items or longer measures "fit" a construct definition. For example, Bech (1992) described a model for health-related quality of life based on six diagnostic components: physical, cognitive, affective, social, economic, and ego-function aspects (PCASEE). Bech's Table 1 relates the six components to specific variables such as sleep (P), concentration (C), depression (A), and so on. However, for others quality of life may be represented in terms of self-assessments: Joyce et al. (1999) listed the individual's assessment of subjective health as the manifestation of Bech's physical aspect, their decision-making capacities as representing cognition, warm feelings toward others as an index of affective QOL, and so on.

Remarkably, though papers can be found with personal definitions of happiness and well-being provided by Veenhoven, few that we examined connect to other scholarship, and many are overly glib and simplistic about defining happiness.[2] Veenhoven (2007, p. 3) states that "'well-being' denotes that something is in a good state." Veenhoven (2009) stipulated that happiness and quality of life and well-being are the same. Further musings dissect quality of life into components, but do not tie the ideas to other literatures that conceptualize or theorize on these constructs, and those literatures are each extensive. Veenhoven (2009) is the most thorough, though its provision of evidence is haphazard, with little attention to the different populations and cultural groups that appear in the database. One finds the conceptual basis for harmonization could be stronger in the happiness realm.

Examination of items is another component of conceptual harmonization. For example, Wang et al. (2014) provide a compendium of items tapping health-related behaviors across a set of surveys that aimed to be comparable to the US Health and Retirement Study (HRS). These reveal the vast array of questions asked with content deemed "similar enough" to be harmonized. Lengthy concordance tables are given for items on smoking, drinking, and physical activity.

Chen et al. (2021) conducted a similar process that they called "pre-statistical harmonization" which involved close inspection of all items, reviews of scoring procedures, and inspection of populations assessed in surveys of behavioral symptoms of dementia. Individual patient data from eight samples allowed them to also conduct statistical harmonization, including psychometric analyses using item response theory, model-fit analyses, and examination of inter-item correlations and cross-tabulations. Experts in the content matter at hand would be critical to this process.

Given sufficient data, one could conduct empirical validity studies of collected items such as investigations of inter-item correlations or factor analyses. However, if each study contributes only a few items to the collection, such studies would require additional new primary data. Seemingly sensible quantitative analyses should be preceded by conceptual harmonization. Even when harmonization is a reasonable goal, surveys intending to include comparable measures may still show notable variation in wording and content. Some of these features are described in the following sections.

[2] Other papers in Veenhoven's extensive writings may present more thorough analyses of relevant theories and evidence.

4.3 Item Features Critical to Harmonization

We next consider four important survey-item features that impact the synthesis of survey items. Findings on item writing and psychometrics guide this work. These features are candidates for coding because they may relate to the responses of participants and may also play a role if statistical harmonization is to take place. We discuss:

- Variation in wording of the question stem or statement
- Number of response-option categories
- Scale direction: Unipolar vs. bipolar scales
- Nature of response options: Labeling and wording of response options

Other features may play roles in cross-survey variation in items (e.g., the use of negatively and positively worded items, per Pilotte and Gable (1990), among many others), but we believe these four features are most important and are moderately easily addressed.

4.3.1 Wording of Item Stems

Differences in the wording of item stems can result in variation in the focus and the strength of the questions asked, and affect the strength of responses from recipients. The impact of wording in surveys is parallel to the way that test-item wording and stem complexity affect difficulty in standard educational exams (e.g., Ascalon et al., 2007). More extreme stem statements are expected to be harder to endorse. Schuman and Presser (1996) discuss various aspects of wording including intensity, centrality, and tone in several chapters in their classic book; there are too many to properly cover here. Additionally, the exact wording differences that are important will surely vary from one meta-analysis to another, so we mention here only the general idea and its importance.

A multifaceted example of stem differences was reported in the RAND project to harmonize measures of health behavior in the elderly. Examining ten different longitudinal studies of aging, Wang et al. (2014) noted that questions about drinking alcoholic beverages varied in the time frames they asked about, and the amounts and specificity of beverages consumed. Items on the frequency of drinking asked about time ranges including the last 7 days, last month, last 3 months, last 6 months, and last year. Some asked about consumption of “any alcohol,” whereas others differentiated wine versus beer versus spirits, and the most focused questions probed for consumption of “normal beer” versus strong beer, or listed specific types of liquor (e.g., wine, beer, and whiskey). Cross-national comparisons are difficult when particular beverages are more popular and readily available in their country of origin (e.g., sake, soju, makgeolli); such wording matters are idiosyncratic but may be key to understanding a particular population.

As an example, suppose we want to summarize information on the portion of the elderly population that is engaged in heavy drinking. One might expect the two features of time frame and amount of alcohol consumed per unit time to work together to represent total alcohol consumption. Thus, it would be important for the meta-analyst to capture such features during data extraction. Some meta-analyses have used coded variables to create additional variables. For example, multiplying the number of treatment occasions for an intervention by the typical session length provides a total-exposure-to-treatment measure. A similar approach could be taken for assessing alcohol consumption. To synthesize data on items that do not allow for similar computations, the meta-analyst could ask expert raters to assess the strength of the item-stem statements and use those assessments as moderators of diversity in the responses.

4.3.2 Number of Response Options

Another feature of item responses that leads to between-surveys variation is the number of response options. Differing numbers of options are the leading reason for using the statistical harmonization methods described below, because the scales of item scores correspond to the number of options available. In some cases, the options are nominally or qualitatively different and for those, scale changes are not needed. In others item responses represent an implied underlying continuum. This distinction has implications for the choice of quantitative approaches to scale harmonization. It is tempting to separate dichotomous items from items with three or more options, but when an item measures an underlying continuum, the difference between two and three or more options is simply one of the granularities of responses. Often surveys require respondents to reply using ordinal scales with varying numbers of categories. Others may allow for continuous responses, for instance, by placing a mark on a line. In any case when synthesizing findings from individual items, the meta-analyst may want to consider the number of options as a potential moderator of between-items differences in results, as there is no way to add response options after the fact.

To enable sensible comparisons, scale conversions have been used to rescale individual item outcomes for ordinal- and interval-scale items, to locate them on a common metric. Many of these are listed in handbooks for statistical harmonization (e.g., Griffith et al., 2013), and we show several examples below. At its simplest, rescaling may entail collapsing responses or applying simple linear transformations of scale points; however, these can lead to potentially idiosyncratic translations across items. More complicated approaches may require assigning labeled points differently across studies, or adopting more sophisticated latent variable models which involve more assumptions and computation (e.g., van den Heuvel et al., 2020).

4.3.3 Nature of Response Options

Responses to items may be partially or fully labeled and may be graphical, numeric, or verbal. When survey-item responses are verbally labeled, the nature of the responses available as answers must be considered. Response-option-label differences have been the focus of efforts to harmonize measures across surveys, especially in the work of Veenhoven and his team (DeJonge et al., 2017), and can be very tricky when different formats appear across studies (e.g., what words are associated with the array of frowning faces on pain-scale items?).

Differences in response labeling are presumed to lead to different choices by participants, and are known to vary across different surveys, contributing to further between-surveys variation. Thus, this feature is one that should be coded or characterized by the meta-analyst. When a finite number of categories is offered, response options may be fully or partially labeled; this has long been noted to affect the reliability of responses (Endig, 1953). Similarly even when respondents are offered a continuum to mark, labels may be specified at different points along the length of the response line, and continuous-looking scales may be assigned integer scores by survey software. Coding these features enables the meta-analyst to empirically assess whether such variations in design affect participant responses.

4.3.4 Item Polarity or Direction

Some surveys use items with responses organized along continua with endpoints that are meant to be opposites; these are bipolar items. For example, ratings may range from “happy” to “unhappy” (or if endpoints have modifiers, “very happy” to “very unhappy,” etc.). In contrast unipolar item responses may run from “not at all happy” to “very happy,” with no coverage of the range representing degrees of unhappiness. This approach to labeling may reflect a potential belief that, say, happiness and unhappiness reflect two separate dimensions, rather than two ends of a continuum. It may be possible to link similar option choices across items of different polarities if verbal labels are assigned to all items.

To combine unipolar and bipolar items presents a challenge to the meta-analysis. Using simple linear transformations (e.g., that move responses to a common scale) will not address the fact that the endpoint of a unipolar scale may correspond more closely to the middle of a bipolar scale than to its end. This is why numerical transformations cannot be blindly applied without consideration of the constructs per se that are tapped by individual items. At the very least, the polarity of items must be coded, and it may be sensible to separately analyze items with different polarity, unless a translation can be found that soundly matches response options across these two item types.

4.4 Statistical Harmonization

Hofer and Piccinin (2009) of the Integrative Analysis of Longitudinal Studies on Aging (IALSA) project have pioneered the idea of harmonization within the study of the psychology of aging. They and others have provided a guide to statistical harmonization aimed largely for use in individual participant meta-analyses (Griffith et al., 2013). Many approaches to this statistical harmonization, along with supportive software (e.g., Adhikari et al., 2021; Fortier et al., 2011; Winters and Netscher, 2016), have been developed. We touch on some of the most common methods here and discuss their weaknesses.

4.4.1 Linear Transformations

A conventional method to locate item responses on a common scale is to place the scale points of the diverse items (i.e., responses to individual items) onto a common secondary scale by applying linear transformations. These date far back (Hull, 1922) and have numerous instantiations. The simplest linear transformations include linear stretching that converts primary-scale response-option scores to a common scale running between prespecified endpoints (e.g., 1–10) and standard linear transformations that shift scores to a scale with a known mean and standard deviation (SD), such as the T score with a mean of 50 and SD of 10, or the well-known z score.

4.4.2 Linear Stretching

If the number of response options of a primary scale is smaller than that of the common scale, the transformation is done by linearly stretching the scale points from the smaller scale onto the larger scale (e.g., moving a 5-point scale to 10-point scale). If the number of primary-scale response options is larger than the number of the common-scale response options, the primary scale is linearly compressed into the boundaries of the common scale (e.g., from a 10-point scale to 5-point scale).

The equation below can be used to stretch or shrink the scale points from one scale X to another, say Y:

$$Y = \left[(X - \min(X)) \times \frac{\max(Y) - \min(Y)}{\max(X) - \min(X)}\right] + \min(Y), \text{ or}$$

$$Y = \left[X \times \frac{\max(Y) - \min(Y)}{\max(X) - \min(X)}\right] - \min(X) \times \frac{\max(Y) - \min(Y)}{\max(X) - \min(X)} + \min(Y), \quad (1)$$

where X is a scale point of the original item; $\min(X)$ and $\max(X)$ are the minimum and maximum possible scale points of that item (not the observed min and max values), respectively; and $\min(Y)$ and $\max(Y)$ are the analogous values on the transformed scale (Card, 2011). This is easily seen to be a linear transform $Y = a + bX$ where

$$a = -\min(X) \times \frac{\max(Y) - \min(Y)}{\max(X) - \min(X)} + \min(Y) \text{ and } b = \frac{\max(Y) - \min(Y)}{\max(X) - \min(X)}.$$

4.4.3 Linear Transformation to a Target Mean and Variance

An obvious second transformation assigns a new mean (say μ_Y) and variance (σ_Y^2) to the scale responses. This is attained by using the linear transform $Y = a + bX$, with

$$a = (\mu_Y S_x - \bar{X}\sigma_Y)/S_x \text{ and } b = \sigma_Y/S_x.$$

Target values are denoted using Greek characters to distinguish from the sample values for the original scales.

4.4.4 Assumptions

Once the scale points across a set of items are on a common scale, a meta-analyst can directly summarize their results across studies. The assumptions of linear translations are that response options are equidistant (i.e., interval scaling), that the most extreme possible responses on all items should be scored as min(Y) and max(Y), and that identical verbal labels do not need to be assigned the same numerical value across items or surveys. Griffith et al. (2015) state without support that these methods also assume normality (we doubt this condition is needed), but rightly note that such translations may run into trouble if the measures have very non-normal distributions (e.g., skewness due to ceiling effects). Indices used may include mean scores if one is willing to assume an underlying continuum, or proportions of participants scoring above (or below) a set cutoff on the new scale.

Zumbo and Woitschach (2021) endorse the concerns we raise and have raised several additional concerns about the family of linear transforms by examining a more stringent mathematical formalization. These authors concur with our assessment that when a scale is fundamentally only ordinal, simplistic translations cannot magically change that fact.

4.4.5 Example

We demonstrate using two survey questions measuring a respondent's degree of happiness, taken from the World Database of Happiness (Veenhoven, n.d.). The first question is "In general, how happy would you say you are these days?" Its response scale runs from 1 to 7 with response-option labels Not happy at all (1), Very unhappy, Somewhat unhappy, Neither happy nor unhappy, Pretty happy, Very happy, and Extremely happy (7). The second question is worded slightly differently: "How happy do you feel as you live now?" This question has four response-option labels scored from 1 to 4: Very unhappy, Not too happy, Pretty happy, and Very happy, respectively. We use the linear-stretching method to transform the ratings of both scales to a common metric running from 0 to 10. After applying this method, the ratings for response labels of the first question become 0, 1.67, 3.33, 5, 6.67, 8.33, and 10, respectively, and those of the second question are 0, 3.33, 6.67, and 10, respectively.

Once the original ratings are linearly transformed to the same metric, the observed means ($\hat{\mu}_y$) and variances (S_y^2) can be calculated for each transformed scale with an underlying continuum by entering the transformed ratings y_j and the proportions of respondents choosing each of the ratings ($P(y_j)$) in the following equations:

$$\hat{\mu}_y = E(Y) = \sum_{j=1}^{J} y_j P(y_j), \tag{2}$$

$$S_y^2 = \sum_{j=1}^{J} (y_j - \hat{\mu})^2 P(y_j), \tag{3}$$

where we sum across all J response categories, $j = 1$ to J. Alternatively, one may calculate the linearly transformed mean and standard deviation directly from the means and standard deviations of the original scales (Kalmijn, 2010). Specifically,

$$\hat{\mu}_y = \left[\hat{\mu}_x - \min(X)) \times \left(\frac{\max(Y) - \min(Y)}{\max(X) - \min(X)}\right)\right] + \min(Y), \tag{4}$$

$$S_y = S_x \times \left(\frac{\max(Y) - \min(Y)}{\max(X) - \min(X)}\right). \tag{5}$$

The linear-stretching method has a few drawbacks especially when used to transform items with verbal response labels. One is that the meta-analyst must assume equidistance between response categories of all the scales being transformed (i.e., assume they have interval-scale properties). This assumption implies that the differences between successive response-option categories are the same within the old and new metrics. For the first question in our example, for instance, the difference in the degree of happiness between "Pretty happy" and "Very happy" should be the same as the difference in the degree of happiness between "Very happy" and "Extremely happy". This is impossible to verify as the measure is inherently ordinal even if the underlying construct is not.

The other issue is that the transformation does not consider either the verbal anchoring of response options or any differences in strength or focus of the question stems. For instance, after linear stretching has been applied, the score of 3.33 on the common 0–10 scale is associated with both the “Somewhat unhappy” verbal label of the first scale and “Not too happy” of the second scale, which is unsatisfying, and implies that they have the same meaning. In contrast at the top end of the scale, 6.67 is assigned to “Pretty happy” on both items, but “Very happy” is assigned 8.33 for the 7-point item versus 10 for the 4-point item. This lack of correspondence is both problematic and confusing.

4.4.6 Nonlinear Transformations

Two nonlinear and nonmathematical transformation approaches can be used to avoid these issues associated with the linear-transformation method. Both approaches are based on using subjective judgments of individuals (e.g., coders or expert judges) to determine corresponding values for possible response-option labels on a secondary numerical scale. The use of raters or coders to rate prompts (e.g., Eagly & Carli, 1981) or other study features (e.g., quality features; Atkins et al., 2004; Guyatt et al., 2008) is common in meta-analysis, and the use of ratings as moderators of study effects is also common. Here raters are asked to evaluate the semantics of the response-option labels.

In one approach, coders are presented with a list of all response-option labels from the primary scales and are asked to assign values to each label on a second common metric that is bounded by predetermined values. In the other approach from Veenhoven et al. (1993), coders evaluate the semantics of each response-option label after reading the question stem and the other response-option labels of the item.

In the first approach, which we call the semantic-judgment-out-of-context method, each response option is rated irrespective of the other response-option labels of the original item and its relative position on the common scale that contains all other response-option labels. This approach does not take into account the differences in the wordings of the question stems or the number or nature of the response options of the items. A weakness of this approach is that differences in question stems may impact how judges interpret the response-option labels being rated. Also, the semantic intensity of a response-option label may be interpreted differently depending on whether it is presented with many other options (e.g., seven categories) or few (e.g., three categories).

One early application of this approach was as the first step in a longer process of scaling items proposed by Jones and Thurstone in 1955. Respondents rated the semantic strength of 51 phrases (response options) used to indicate like or dislike of various foods (e.g., strongly like, tasty, bad). Each phrase was presented independently with no stem (i.e., no prompt of a specific food), and the raters assigned to each an integer value between −4 and 4, inclusive. Scale endpoints were labeled with “Greatest Dislike” and “Greatest Like” and the midpoint (0) was labeled with “Neither Like Nor Dislike.” The authors then applied a modified version of the successive-intervals scaling method (Edwards, 1952) to determine the values of the phrases on the common scale. Consequently, all the phrases were placed on an interval where a neutral label has the value 0.

With the second approach, the semantic-judgment-in-context method, coders assign values to each response-option label considering all aspects of the original item, including the relative position of the label, other response-option labels, the number of response options, and the wording of the question stem. Each individual question and its corresponding response-option labels are presented to the coders separately. For each response option, the coders assign a value that they consider the most appropriate on a secondary scale, say running from 0 to 10. One coder might assign 1 for the option label "very unhappy," 3 for "unhappy," 6 for "happy," and 9 for "very happy." Another coder could rate the same labels differently. Also, coders may assign different values to the same response-option label when it is presented in the context of different questions (e.g., with different question stems and accompanying response-option labels). The final ratings for each item's response options are computed by averaging assigned values across coders. Consequently, response-option labels have unique values specific to each survey item on a set secondary scale.

In the extensive project on happiness, Veenhoven et al. (1993) used this approach to place values of the response options of nine survey items tapping happiness on a common scale. The items had almost identical question stems but differed in numbers and labeling of response options. Ten content experts evaluated the semantics of response-option labels by assigning values on a 0 to 10 scale. The means of those values across experts determined the final ratings of the response options. If a response option appeared in multiple items, its ratings were also averaged across items. Response options used only once retained their original rated values. Consequently, the authors came up with a single common scale having all possible options.

A concern with this approach is that it damps down spread that may be explained by other item features. Consider an item with response-label ratings that all exceed the ratings of the same labels when used with other items. Because the mean response score will replace those higher-than-average ratings, the process has the feature of moving extreme labels closer to the center, and in theory could reorder verbal labels within items, especially if ratings adjacent to the focal option are numerically close.

4.4.7 Comparisons of Approaches

Gözütok (2018) made use of survey items and their descriptive statistics (i.e., original means, standard deviations, proportions of respondents on response options) from Veenhoven's World Database of Happiness collection to illustrate how three scale-transformation methods can be used in conducting a meta-analysis. The original response options of items were transformed to a secondary numerical scale running from 0 to 10 by the transformation methods described above. Then means and standard deviations of the items on the new scale were computed. The means obtained from the transformed scales were treated as study outcomes in three hypothetical meta-analyses based on raw-means synthesis (Bond et al., 2003).

In the three pseudo-meta-analyses, Gözütok (2018) included the wording of the question stems as a moderator variable, along with other survey-item characteristics such as number of response-category options, scale polarity (i.e., unipolar vs. bipolar scales), and scale labeling (i.e., endpoints labeled vs. all points fully labeled). To capture differences in the wording of question stems, he used ratings of the strength of the statement or question about the construct (i.e., happiness). This rating task may be done by meta-analysis coders, content experts, or a sample of target respondents. For example, coders may assign a higher rating of strength to the question stem "Do you feel elated?" and give a lower value to "Do you feel happy?". If so, part of the potential between-studies variance in the study outcomes will be explained by the differences in question-stem strength. The strength ratings did not relate to the mean happiness ratings in these pseudo-meta-analyses, possibly because the actual items on happiness were very similar (i.e., there was little between-items variation in wording). Also concern was raised due to an idiosyncratic rater Gözütok identified in his rater pool.

As part of the World Database of Happiness project, a great deal of work regarding the comparability of survey items of the same construct has been done by Veenhoven and his colleagues. Veenhoven (2008) reported on the International Happiness-Scale Interval Study. Its participants provided interval boundaries on a 0 to 10 scale for each response-option label of a set of country-specific happiness items. Each item stem plus its associated set of J response labels was presented to the participants. A web-based tool called the Scale Interval Recorder (Veenhoven & Hermus, 2006) allowed the participant to slide $(J - 1)$ markers that defined the boundaries between the J verbal labels that were associated with each question. Midpoints of the resultant summarized intervals were used to represent each response option.

This study inspired further innovations such as the continuum approach of Kalmijn (2010), and the reference-distribution method from DeJonge et al. (2014). The continuum approach postulates a latent happiness variable in the population. It is assumed to be continuous and was set arbitrarily to have scores in the interval [0, 10]. Beta-distribution shape parameters that best match the observed data are then found. Kalmijn recommends using a beta distribution which is left-skewed to reflect high levels of happiness in the population.

The reference-distribution method builds heavily on other harmonization methods in that it aims to define intervals to represent ranges for response options. In the reference-distribution method, the boundaries between the response options of the primary scale are derived from a reference distribution instead of from ratings by judges on a scale-interval recorder. Again this approach assumes an underlying latent distribution and uses the beta distribution that fits best to the survey results of the responses of a given sample and item. For interested readers, details and implications of these approaches can be found in DeJonge et al. (2017).

Other methods have been proposed to harmonize data from different survey items. Griffith et al. (2015) provided a summary of statistical approaches used in systematic reviews of cognitive variables. Transformation and other score-conversion techniques were common, but these authors argue for more sophisticated latent variable techniques including factor analysis and item response theory that would be difficult to achieve without individual participant data.

5 Conclusion

Methods and software tools have been developed to support the practical task of harmonization, but few focus on the detailed conceptual components that we argue are crucial. One exception is Fortier et al. (2017) whose step 2a concerns variable definition. Also on the whole, the practice of statistical harmonization remains simplistic. Griffith et al. (2015) and Zumbo and Woitschach (2021) have argued for the use of latent variable modeling in statistical harmonization, which is consistent with mainstream work in measurement and assessment, yet it is rarely used. The problem for the meta-analyst is that unless extensive individual-level data are available, these analyses cannot be conducted. For example, few surveys have simultaneously administered more than one or two of the hundreds of items in the Happiness Database. Also, such approaches assume that the first step of conceptual harmonization has occurred and identified a set of measures worthy of calibration and linking. It is unclear how often this has been done.

One possible route for meta-analysts, though a labor-intensive one, would be to include in the meta-analysis what are called bridge studies (e.g., Perie et al., 2005, which examined changes in the test structure of the National Assessment of Educational Progress or NAEP). The goal would be to map different items onto one calibrated scale. After conceptual harmonization of target items, the meta-analyst would administer those survey items to a new sample from the population of interest. When one considers the massive numbers of highly similar items in the World Database of Happiness, the task seems impossible. However, if multiple subsamples responded to smaller structured subsets of items (e.g., using incomplete blocks designs as have been used in NAEP and other large-scale assessments), a set of calibrations with common anchor items could allow for various forms of equating or linking (Kolen & Brennan, 2004) to be applied. Various designs for effective linking already exist. This would also allow for item analyses to be conducted to check on the structure of the construct as a whole, thus enhancing the logical examinations and construct-validity analyses done as part of conceptual harmonization.

References

Adhikari, K., Patten, S. B., Patel, A. B., Premji, S., Tough, S., Letourneau, N., Giesbrecht, G., & Metcalfe, A. (2021). Data harmonization and data pooling from cohort studies: A practical approach for data management. *International Journal of Population Data Science, 6*(1), 21. https://doi.org/10.23889/ijpds.v6i1.1680

Angrisani, M., & Lee, J. (2011). *Harmonization of cross-national studies of aging to the health and retirement study income measures (WR-861/5)*. RAND Corporation: Santa Monica, Calif. https://doi.org/10.7249/WR861.5

Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education, 20*(2), 153–170. https://doi.org/10.1080/08957340701301272

Atkins, D., Best, D., Briss, P. A., Eccles, M., Falck-Ytter, Y., Flottorp, S., Guyatt, G. H., Harbour, R. T., Haugh, M. C., Henry, D., Hill, S., Jaeschke, R., Leng, G., Liberati, A., Magrini, N., Mason, J., Middleton, P., Mrukowicz, J., O'Connell, D., Oxman, A. D., . . . GRADE Working Group. (2004). Grading quality of evidence and strength of recommendations. *BMJ (Clinical Research Ed.), 328*(7454), 1490–1494. https://doi.org/10.1136/bmj.328.7454.1490

Bech, P. (1992). Issues of concern in the standardization and harmonization of drug trials in Europe: Health-related quality of life, ESCT Meeting, Strasbourg, 23–24 May 1991. *Quality of Life Research: An International Journal of Quality of Life: Aspects of Treatment, Care and Rehabilitation, 1*(2), 143–145. https://doi.org/10.1007/BF00439722

Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8*(4), 406–418. https://doi.org/10.1037/1082-989X.8.4.406

Brown, S. A., García, A. A., Brown, A., Becker, B. J., Conn, V. S., Ramírez, G., Winter, M. A., Sumlin, L. L., Garcia, T. J., & Cuevas, H. E. (2016). Biobehavioral determinants of glycemic control in type 2 diabetes: A systematic review and meta-analysis. *Patient Education and Counseling, 99*(10), 1558–1567. https://doi.org/10.1016/j.pec.2016.03.020

Card, N. A. (2011). *Applied meta-analysis for social science research*. New York: Guilford.

Chen, D., Jutkowitz, E., Iosepovici, S. L., Lin, J. C., & Gross, A. L. (2021). Pre-statistical harmonization of behavrioal [sic] instruments across eight surveys and trials. *BMC Medical Research Methodology, 21*(1), 227. https://doi.org/10.1186/s12874-021-01431-6

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relations. In L. B. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them. New Directions for program evaluation* (Vol. 57). Jossey-Bass.

Cooper, H. M. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Sage.

Crites, S. L., Fabrigar, L. R., & Petty, R. E. (1994). Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues. *Personality and Social Psychology Bulletin, 20*(6), 619–634. https://doi.org/10.1177/0146167294206001

DeJonge, T., Veenhoven, R., & Arends, L. (2014). Homogenizing responses to different survey questions on the same topic: Proposal of a scale homogenization method using a reference distribution. *Social Indicators Research, 117*(1), 275–300. https://doi.org/10.1007/s11205-013-0335-6

DeJonge, T., Veenhoven, R., & Kalmijn, W. (2017). *Diversity in survey questions on the same topic: Techniques for improving comparability*. Springer.

Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. *Psychological Bulletin, 90*(1), 1–20. https://doi.org/10.1037/0033-2909.90.1.1

Edwards, A. L. (1952). The scaling of stimuli by the method of successive intervals. *Journal of Applied Psychology, 36*(2), 118–122. https://doi.org/10.1037/h0058208

Endig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *Journal of Applied Psychology, 37*, 38–41. https://doi.org/10.1037/h0057911

ESOMAR European Society for Opinion and Marketing Research. (2003). The ESOMAR standard demographic classification. In J.H.P. Hoffmeyer-Zlotnik & C. Wolf, (Eds.), *Advances in cross-national comparison*. Boston, MA: Springer. https://doi.org/10.1007/978-1-4419-9186-7_6

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One, 4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Fortier, I., Doiron, D., Little, J., Ferretti, V., L'Heureux, F., Stolk, R. P., Knoppers, B. M., Hudson, T. J., & Burton, P. R. (2011). Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *International Journal of Epidemiology, 40*(5), 1314–1328. https://doi.org/10.1093/ije/dyr106

Fortier, I., Raina, P., van den Heuvel, E. R., Griffith, L. E., Craig, C., Saliba, M., Doiron, D., Stolk, R. P., Knoppers, B. M., Ferretti, V., Granda, P., & Burton, P. (2017). Maelstrom research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology, 46*(1), 103–115. https://doi.org/10.1093/ije/dyw075

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

Goeltz, R. K. (1991). International accounting harmonization: The impossible (and unnecessary?) dream. *Accounting Horizons, 5*(1), 85.

Gözütok, A. S. (2018). Critical issues in survey meta-analysis. Unpublished doctoral dissertation. Florida State University.

Griffith, L., van den Heuvel, E., Fortier, I., Hofer, S. M., Raina, P., Sohel, N., Payette, H., Wolfson, C., & Belleville, S. (2013). Harmonization of cognitive measures in individual participant data and aggregate data meta-analysis. methods research report. (Prepared by the McMaster University Evidence-based Practice Center under Contract No. 290-2007-10060-I.) AHRQ Publication No.13-EHC040-EF. Rockville, MD: Agency for Healthcare Research and Quality. https://www.ncbi.nlm.nih.gov/books/NBK132553/

Griffith, L. E., van den Heuvel, E., Fortier, I., Sohel, N., Hofer, S. M., Payette, H., Wolfson, C., Belleville, S., Kenny, M., Doiron, D., & Raina, P. (2015). Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. *Journal of Clinical Epidemiology, 68*(2), 154–162. https://doi.org/10.1016/j.jclinepi.2014.09.003

Guyatt, G. H., Oxman, A. D., Kunz, R., Vist, G. E., Falck-Ytter, Y., Schünemann, H. J., & GRADE Working Group (2008). What is "quality of evidence" and why is it important to clinicians? *BMJ (Clinical Research ed.), 336*(7651), 995–998. https://doi.org/10.1136/bmj.39490.551019.BE

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods, 14*(2), 150–164. https://doi.org/10.1037/a0015566

Holloway, J., & Collins, D. (1982). Social policy harmonization in the European community. *Journal of Social Policy, 11*, 144–144.

Hull, C. L. (1922). The conversion of test scores into series which shall have any assigned mean and degree of dispersion. *Journal of Applied Psychology, 6*(3), 298–300.

Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Jones, L. V., & Thurstone, L. L. (1955). The psychophysics of semantics: an experimental investigation. *Journal of Applied Psychology, 39*(1), 31–36. https://doi.org/10.1037/h0042184

Joyce, C. R. B., McGee, H. M., & O'Boyle, C. A. (Eds.) (1999). *Individual quality of life*. Routledge.

Kalmijn, W. M. (2010). Quantification of happiness inequality. Unpublished doctoral dissertation. Erasmus University Rotterdam.

Kish, L. (1994). Multipopulation survey designs: Five types with seven shared aspects. International Statistical Review, 62(2), 167–186.

Kish, L. (1999a). Combining surveys: A framework. Bulletin of the International Statistical Institute: Proceedings of the ISI 52nd Session, Finland. https://www.stat.fi/isi99/proceedings/arkisto/varasto/kish0135.pdf

Kish, L. (1999b). Cumulating/combining population surveys. *Survey Methodology, 25*(2), 129–138.

Kish, L. (2002). Combining multipopulation surveys. *Journal of Statistical Planning and Inference, 102*, 109–118.

Kolen, M.J., & Brennan, R.L. (2004). Test equating, scaling, and linking. Springer

Lewis, S. M. (1990). Standardization and harmonization of the blood count: The role of International Committee for Standardization in Haematology (ICSH). *European Journal of Haematology. Supplementum, 3*, 9–13. https://doi.org/10.1111/j.1600-0609.1990.tb01520.x

Morton, S. (1999). Combining surveys from a meta-analysis perspective. Bulletin of the International Statistical Institute: Proceedings of the ISI 52nd Session, Finland. https://www.stat.fi/isi99/proceedings/arkisto/varasto/mort0275.pdf

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*, 143. https://doi.org/10.1186/s12874-018-0611-x

National Quality Forum. (2010). Guidance for measure harmonization: A consensus report. Washington, DC: NQF. https://www.qualityforum.org/Publications/2011/05/MeasureHarmonization_full.aspx

Nikula, S., Jylhä, M., Bardage, C., Deeg, D. J., Gindin, J., Minicuci, N., Pluijm, S. M., Rodríguez-Laso, A., & CLESA Working Group (2003). Are IADLs comparable across countries? Sociodemographic associates of harmonized IADL measures. *Aging Clinical and Experimental Research, 15*(6), 451–459. https://doi.org/10.1007/BF03327367

Ostrom, T. M. (1969). The relationship between the affective, behavioral, and cognitive components of attitude. *Journal of Experimental Social Psychology, 5*(1), 12–30. https://doi.org/10.1016/0022-1031(69)90003-1

Perie, M., Moran, R., & Lutkus, A. D. (2005). *NAEP 2004 Trends in academic progress: Three decades of student performance in reading and mathematics*. (NCES 2005–464). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington, DC: Government Printing Office.

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement, 50*(3), 603–610. https://doi.org/10.1177/0013164490503016

Pluijm, S. M., Bardage, C., Nikula, S., Blumstein, T., Jylhä, M., Minicuci, N., Zunzunegui, M. V., Pedersen, N. L., & Deeg, D. J. (2005). A harmonized measure of activities of daily living was a reliable and valid instrument for comparing disability in older people across countries. *Journal of Clinical Epidemiology, 58*(10), 1015–1023. https://doi.org/10.1016/j.jclinepi.2005.01.017

Quatresooz, J., & Vancraeynest, D. (1992). Harmonisation of demographics in Europe 1991: The state of the art; Part 2: Using the ESOMAR Harmonised Demographics: External and internal validation of the results of the EUROBAROMETER Test. *Marketing and Research Today, 20*(1), 41.

Rao, S. R., Graubard, B. I., Schmid, C. H., Morton, S. C., Louis, T. A., Zaslavsky, A. M., & Finkelstein, D. M. (2008). Meta-analysis of survey data: Application to health services research. *Health Services and Outcomes Research Methodology, 8*(2), 98–114. https://doi.org/10.1007/s10742-008-0032-0.

RetailMeNot Editors. (2021). RetailMeNot Study Finds Reese's and M& M's Are STILL the Most Popular Halloween Candies This Year. RetailMeNot. https://www.retailmenot.com/blog/favorite-halloween-candy-revealed.html

Schenker, N., & Raghunathan, T.E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine, 26*(8), 1802–1811. https://doi.org/10.1002/sim.2801

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Academic Press.

van de Water, H. P., Perenboom, R., J., & Boshuizen, H. C. (1996). Policy relevance of the health expectancy indicator; an inventory in European Union countries. *Health Policy, 36*(2), 117–129. https://doi.org/10.1016/0168-8510(95)00803-9

van den Heuvel, E. R., Griffith, L. E., Sohel, N., Fortier, I., Muniz-Terrera, G., & Raina, P. (2020). Latent variable models for harmonization of test scores: A case study on memory. *Biometrical Journal, 62*(1), 34–52. https://doi.org/10.1002/bimj.201800146

Veenhoven, R. (2007). Subjective measures of well-being. In M. McGillivray (Ed.) Human well-being: Concept and measurement. Palgrave/McMillan.

Veenhoven, R. (2008). The international scale interval study. In V. Møller & D. Huschka (Eds.), *Quality of life in the new millennium: 'Advances in quality-of-life studies, theory and research',* *Part 2: Refining concepts and measurement to assess cross-cultural quality of-life* (pp. 45–58). Social Indicator Research Series, vol. 35. Springer Press.

Veenhoven, R. (2009). How do we assess how happy we are? Tenets, implications and tenability of three theories. In A. K. Dutt & B. Radcliff (Eds.), *Happiness, economics and politics: Towards a multi-disciplinary approach* (pp. 45–69). Edward Elger.

Veenhoven, R. (2015). Concept of happiness. Downloaded from https://worlddatabaseofhappiness-archive.eur.nl/hap_quer/introtext_measures2.pdf

Veenhoven, R. (n.d.) World Database of Happiness, Erasmus University Rotterdam, The Netherlands. http://worlddatabaseofhappiness.eur.nl

Veenhoven, R., & Hermus, P. (2006). *Scale interval recorder. Tool for assessing relative weights of verbal response options on survey questions. Web survey program*. Erasmus University Rotterdam.

Veenhoven, R., Ehrhardt, J., Ho, M. S. D., & de Vries, A. (1993). Happiness in nations: Subjective appreciation of life in 56 nations 1946–1992. Erasmus University Rotterdam.

Wang, S., Min, J., & Lee, J. (2014). Harmonization of cross-national studies of aging to the Health and Retirement study: USER GUIDE, Health behavior, Version A. (WR-861/8) Santa Monica, Calif.: RAND Corporation. https://doi.org/10.7249/WR861.8

Winters, K., & Netscher, S. (2016). Proposed standards for variable harmonization documentation and referencing: A case study using QuickCharmStats 1.1. *PLoS One, 11*(2), e0147795. https://doi.org/10.1371/journal.pone.0147795

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems, 22*(1), 45–55. https://doi.org/10.1057/ejis.2011.51

Ye, D., Ng, Y. K., & Lian, Y. (2015). Culture and happiness. *Social Indicators Research, 123*(2), 519–547. https://doi.org/10.1007/s11205-014-0747-y

Zumbo, B., & Woitschach, P. (2021). A critique of the conventional methods of survey item transformations, with an eye to quantification. In Michalos, A. C. (Ed.), *The Pope of Happiness—A Festschrift for Ruut Veenhoven* (pp. 303–313). Springer. https://doi.org/10.1007/978-3-030-53779-1_30

Two Sources of Nonsampling Error in Fishing Surveys

J. Michael Brick, William R. Andrews, and John Foster

Abstract Two important nonsampling errors arise from nonresponse and noncoverage. Both nonresponse and noncoverage are forms of missing data and are potentially important contributors to the accuracy of survey estimates. This research examines the potential effects of both nonresponse and noncoverage in the Fishing Effort Survey (FES), a survey conducted by the National Oceanic and Atmospheric Administration (NOAA). The difficulty in evaluating nonresponse and noncoverage is that the values for the missing data are not available and proxy measures must be used. Using proxies we find that noncoverage results in very large biases, while the magnitude of nonresponse bias is negligible in comparison. Another important finding is that the rates of missing data due to nonresponse or noncoverage are not predictive of the magnitude of the biases.

1 Introduction

All sample surveys are subject to sampling and nonsampling errors which cause survey estimates to deviate from true population values. Survey sampling texts (e.g., Cochran, 1977 and Lohr, 2019) describe survey design techniques to lower sampling errors in a cost-effective manner, but these texts provide less guidance on reducing nonsampling errors. This research investigates two important sources of nonsampling errors that are forms of missing data: nonresponse and noncoverage. The effects of nonresponse and noncoverage errors in a survey used to estimate fishing effort are examined both separately and jointly. The joint analysis provides

J. M. Brick (✉)
Westat, Rockville, MD, USA
e-mail: mikebrick@westat.com

W. R. Andrews · J. Foster
NOAA Fisheries, Silver Spring, MD, USA
e-mail: rob.andrews@noaa.gov; john.foster@noaa.gov

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_7

insight into the relative magnitudes of the potential error and can help prioritize efforts to improve the overall quality of the survey.

Nonresponse affects virtually all surveys and has been the subject of many articles and texts (e.g., Groves & Couper, 1998; Särndal & Lundström, 2005, and Stoop, 2005). Falling response rates (Atrostic et al., 2001; Williams & Brick, 2017, and Luiten et al., 2020) both in the United States and internationally have greatly heightened interest in nonresponse. For example, Stedman et al. (2019) question whether low-response rates imply surveys are no longer valid research vehicles, and Groves (2006) discusses the value of probability samples with low-response rates.

The research on noncoverage error, another form of missing data, is more diffuse because the source of noncoverage differs across surveys. A more comprehensive review is done by Lessler and Kalsbeek (1992), who discuss nonresponse, noncoverage, and other nonsampling errors such as measurement error. Narrower research on specific instances of noncoverage is illustrated by the work on telephone surveys in the 1980s and 1990s (Thornberry & Massey, 1988), and then again when mobile devices were first introduced (Tucker et al., 2007). Similarly, web surveying has spawned research on noncoverage related to Internet access (Scherpenzeel & Bethlehem, 2010).

Nonresponse and noncoverage both result in missing data. As a result, the two sources of nonsampling error can be investigated, at least theoretically, using the same structure. For example, consider the bias in a sample estimate of the mean, $\bar{y}_{nm} = \sum_{k \in s_{nm}} d_k y_k / \sum_{k \in s_{nm}} d_k$, where d_k is the inverse of the probability of selection for unit k and s_{nm} is the set of non-missing data where the missingness is due to either nonresponse or noncoverage. The bias can be written as

$$Bias(\bar{y}_{nm}) \approx M(\bar{Y}_{nm} - \bar{Y}_m), \tag{1}$$

where M is the percent of missing data, $\bar{Y}_{nm}$ is the mean of the (possibly hypothetical) stratum of those who would provide data, and $\bar{Y}_m$ is the mean of the stratum of those who would not provide data.

Another popular way of characterizing nonresponse is as a function of the correlation between the probability of a sampled unit responding (its response propensity) and the outcome variable (Bethlehem, 1988). This stochastic representation is

$$Bias(\bar{y}_{nm}) \approx \bar{\phi}^{-1} \sigma_\phi \sigma_y \rho_{\phi y}, \tag{2}$$

where $\bar{\phi}$ is the population propensity of providing data, σ_ϕ and σ_y are the standard deviations of the propensity and y variable , and $\rho_{\phi,y}$ is the correlation between the propensity and y. This model is not typically used for noncoverage because coverage propensities—the probability a unit is on the sampling frame—is more difficult to postulate as being random.

For most surveys, either nonresponse or noncoverage errors are examined, but they are rarely examined jointly. One exception is Couper et al. (2007) who do deal with both sources. More theoretical work, like that of Little and Rubin (2019), treat nonresponse and noncoverage as forms of missing data, but do not delve into the implications of the magnitudes of the biases from each source.

Our objective is to examine the potential effects of both nonresponse and noncoverage in a particular survey to better understand the potential implications of each source. The findings help to better understand the optimal approach to managing resources in this survey to improve the accuracy of the estimates. More generally, the development of bias estimates due to each source of missing data for the same survey will help illuminate the nature of both nonresponse and noncoverage and suggest improved ways of thinking about these nonsampling errors for other surveys.

2 The Fishing Effort Survey

The survey that is the focus of our research is the Fishing Effort Survey (FES) conducted by the National Oceanic and Atmospheric Administration (NOAA). The FES is part of NOAA's Marine Recreational Information Program (MRIP), which produces estimates of recreational saltwater fishing catch—estimates that are used in managing fisheries. It is a cross-sectional, household survey that is conducted every 2 months. The key estimates are the total number of private boat and shore-based recreational, saltwater fishing trips taken by residents of coastal states.

The FES is an address-based sample (ABS) where the addresses are stratified into coastal and non-coastal sub-state regions defined by geographic proximity to the coast. Within each geographic strata, addresses are matched to the National Saltwater Angler Registry (NSAR), which is comprised of state lists of licensed saltwater anglers. This matching creates two additional strata: license-matched (households with one or more licensed anglers) and license-unmatched (households that cannot be matched to NSAR). The coastal and license strata were instituted to improve the efficiency of the sample. Within each stratum, addresses are selected in a single stage using simple random sampling. Weights include raking to household control totals derived from the American Community Survey (ACS) and a final poststratification adjustment to the number of households by coastal/non-coastal strata.

The state effort estimates of the number of shore fishing trips and boat fishing trips from the FES are then combined with independent estimates of average catch per trip from the Access Point Angler Intercept Survey (APAIS) to produce estimates of total recreational saltwater catch. Since the FES only samples people who reside in the state, an adjustment is made to the FES estimates to account for the noncoverage of nonresident anglers. Details on the survey protocol and results for 2020 are in the FES annual report (https://media.fisheries.noaa.gov/2021-08/MRIP-Fishing-Effort-Survey-2020-Annual-Report-V2.pdf). More details on estimation methods are given in Papacostas and Foster (2018).

For this research, we focus on the FES conducted in Waves 4 and 5 of 2020 (July–August and September–October time-periods) for four states where a nonresponse

Table 1 Sample sizes, number of completes, and response rates for 2020 standard Fishing Effort Survey and nonresponse follow-up surveys, by state

	Standard survey			Nonresponse follow-up			Overall
State	Sampled	Completed[a]	RR2[b]	Sampled	Completed[a]	RR2[b]	RR2[a]
Overall	28,650	7968	27.9%	15,993	3456	21.9%	42.4%
Florida	4222	1238	28.0%	2235	513	22.4%	42.6%
Massachusetts	6143	2010	31.5%	3160	774	24.3%	47.0%
New York	11,956	2714	26.1%	7253	1343	19.0%	39.8%
North Carolina	6329	2006	28.4%	3345	826	23.1%	43.7%

[a] Includes partial completes with some data that requires editing
[b] RR2 is the American Association of Public Opinion Research response rate 2

follow-up (NRFU) was conducted. The follow-up data are used in the analysis of potential nonresponse bias. The NRFU followed a subsample of the nonrespondents to the standard FES. Details of the NRFU data collection protocol are given in Andrews (2021).

The first columns of Table 1 show the sample sizes, number of completes, and response rates for the standard FES, where the data are aggregated over both waves in each state. Completed surveys include those where all the information requested is 100% reported plus partial completes which include missing or inconsistent information that can be resolved by editing or imputation. About 85–90% of completes require no editing. The overall response rate for the standard survey in these states during the two waves was 27.9%. The subsequent columns show the same information for the NRFU study. The NRFU response rate is based on the nonrespondents sampled for the NRFU. The overall response rates in the last columns combine the standard and NRFU data collection and are weighted to account for the subsampling in the NRFU. The overall response is 42.4%, computed using the AAPOR RR2 formula.

Table 2 shows the percent of households that took a fishing trip (either of any type of fishing or by boat or from the shore) and the mean number of trips of those that did fish by state, geographic area, and license status. We refer to the licensed-matched stratum as “Licensed” and the remainder are “Not licensed.” These estimates are the standard estimates that do not include the NRFU effort. The table demonstrates the considerable variation in fishing by state, area, and license status.

3 Methods of Assessing Bias

The primary problem facing evaluations of nonresponse and noncoverage in surveys is that the values for the missing data are not available except in unusual circumstances. As a result, proxy measures of bias must be substituted for the missing values so that estimates of the effects of the missing data can be computed.

Table 2 Estimated percent of households that fished and mean number of trips, by state and stratum

	Percent			Mean trips	
States	Any fish	Boat fish	Shore fish	Boat	Shore
All four states	11.3	6.9	8.4	6.9	8.6
Coastal county	13.1	8.8	10.4	7.0	8.8
Non-coastal	2.3	1.5	2.4	5.0	5.6
Licensed	43.7	28.7	32.9	9.9	10.0
Not licensed	8.6	5.1	6.3	5.5	8.0
Florida	17.5	10.8	13.1	7.0	9.3
Coastal county	17.5	10.8	13.1	7.0	9.3
Non-coastal	–	–	–	–	–
Licensed	53.6	38.2	39.4	10.4	10.7
Not licensed	13.3	7.6	10.0	5.0	9.3
Massachusetts	8.6	5.3	6.0	6.5	6.2
Coastal county	10.1	6.4	6.8	6.7	6.5
Non-coastal	4.1	1.9	3.7	4.5	4.3
Licensed	58.6	32.8	45.1	10.4	8.6
Not licensed	6.3	4.0	4.2	5.0	4.9
New York	7.1	4.7	4.9	6.7	9.1
Coastal county	10.4	6.9	7.1	6.8	9.7
Non-coastal	1.7	1.0	1.2	5.1	3.0
Licensed	28.9	18.1	21.7	8.8	11.5
Not licensed	6.4	4.2	4.3	6.3	8.6
North Carolina	8.5	4.4	7.1	7.3	6.9
Coastal county	13.8	7.4	11.8	8.1	6.8
Non-coastal	4.4	2.0	3.5	5.0	7.0
Licensed	31.7	18.0	25.3	8.8	8.3
Not licensed	5.2	2.4	4.5	5.8	5.7

The proxy measures used in this analysis are discussed below for both nonresponse and noncoverage.

3.1 Nonresponse Bias

We estimate nonresponse bias in two ways. First, we compare the estimates from the standard survey to the estimates from the data including the standard and NRFU respondents. The difference is a proxy for potential nonresponse bias. This bias proxy assumes the combined data set is unbiased. Although this assumption is unlikely to hold with an overall response rate of 42.4%, Groves and Peytcheva (2008) found this method tends to produce larger estimates of nonresponse bias than other methods.

Table 3 Estimated nonresponse biases of percent who fished, by state and stratum

States	Percent any fish		Percent boat fish		Percent shore fish	
	NRFU	Early/late	NRFU	Early/late	NRFU	Early/late
All four states	0.1	0.1	0.0	0.2	0.1	0.0
Coastal county	0.2	0.2	0.1	0.2	0.2	0.0
Non-coastal	−0.6	0.2	−0.1	0.2	−0.6	0.0
Licensed	0.0	1.2	0.0	1.3	0.2	0.2
Not licensed	0.1	0.0	0.1	0.1	0.0	0.0
Florida	0.3	0.1	0.0	0.2	0.5	−0.3
Coastal county	0.3	0.1	0.0	0.2	0.5	−0.3
Non-coastal	–	–	–	–	–	–
Licensed	0.1	1.6	1.1	2.2	0.1	−0.1
Not licensed	0.3	0.0	−0.1	0.1	0.5	−0.2
Massachusetts	−0.4	0.7	0.1	0.4	−0.5	0.6
Coastal county	−0.4	0.8	0.1	0.5	−0.5	0.6
Non-coastal	−0.1	0.1	0.1	0.0	−0.2	0.3
Licensed	2.5	0.7	0.2	−0.1	2.4	0.5
Not licensed	−0.5	0.6	0.1	0.4	−0.6	0.5
New York	0.3	0.0	0.3	0.1	0.1	0.0
Coastal county	0.5	−0.1	0.3	0.2	0.1	0.0
Non-coastal	0.1	0.3	0.2	0.1	0.1	0.2
Licensed	−0.4	3.4	−1.2	2.1	−0.2	1.9
Not licensed	0.4	−0.2	0.3	0.1	0.1	−0.1
North Carolina	−0.7	−0.1	−0.4	−0.1	−0.6	−0.1
Coastal county	0.4	−0.3	−0.3	−0.4	0.9	−0.1
Non-coastal	−1.5	0.0	−0.6	0.2	−1.7	−0.2
Licensed	−0.5	−1.0	−1.1	−0.2	0.0	−0.9
Not licensed	−0.7	0.1	−0.4	0.1	−0.7	0.1

Another proxy is to compare estimates from early respondents (those who responded to the first mailing) to those of the combined early and late respondents in a level of effort analysis. The assumption that this difference gives an unbiased estimate of the bias is even less likely to hold than the NRFU assumption. Our primary goal for these estimates is to support the development of bounds on the potential bias in the next section.

Table 3 shows the two proxy estimates of nonresponse bias for three key estimates: the percent of households that did any fishing during the time period, the percent that fished from a boat, and the percent that fished from the shore. The NRFU nonresponse bias is computed using Eq. 1, which is equivalent to the difference between the standard estimate and the estimate that includes the NRFU respondents as well as the standard respondents. Both the standard and NRFU estimates went through the full set of adjustments except raking to the ACS.

The early/late bias estimate is the difference between the estimate based only on the early respondents to those based on all respondents to the standard protocol

(NRFU data are not included in the early/late estimates). The early estimate is computed as a domain of the fully weighted standard estimate rather than repeating the weighting steps for this domain.

The nonresponse bias estimates in the table are all relatively small. For example, the 30 bias estimates for the variable any fishing (the eight estimates for each of the four states except Florida which does not have any non-coastal areas) have a mean of 0.2% points. It is also interesting that 10 of the 30 bias estimates are negative because the primary bias concern for the FES is based on the hypothesis that the survey might be subject to topic interest bias (Groves et al., 2004). This hypothesis states that anglers would be more likely to respond to the survey than those who do not fish, so that the extra effort (e.g., the follow-up and more mailings) would increase the proportion of those who did not fish. Since 10 of the 30 estimates are negative (fewer anglers in responding sample) and the sizes of the estimates are relatively small, the NRFU data provide no evidence of substantial nonresponse bias due to topic interest.

3.2 Noncoverage Bias

For noncoverage in the FES, we simulate the effects of not including portions of the population because the ABS frame contains nearly 100% of all residential addresses (Battaglia et al., 2016). As a result, the estimates from the FES are subject to minimal noncoverage, except from nonresident anglers. Noncoverage scenarios are simulated by excluding (1) non-coastal addresses and (2) addresses that are not matched to the license register. These two types of restrictions of the sample have been researched as methods to improve the efficiency of the sample.

The noncoverage bias estimates are computed as the difference between the estimate restricted to either the coastal county or licensed-matched stratum (licensed) and the full sample estimate. The bias estimates were computed by using the full set of weighting procedures except raking, but using only the respondents from the "covered" stratum.

Table 4 shows the percent of fishing households that are excluded under the two scenarios. The magnitude of missing data is relatively small when coverage is restricted to the coastal stratum. In contrast, when the data are limited to the licensed households, the missing data rates are larger and roughly similar to those resulting from nonresponse.

Figure 1 shows both the estimated nonresponse and noncoverage biases for the three estimates of the percent of households that fished in each state and across the four states. The noncoverage bias estimates are labeled "Licensed" and "Coastal" to denote the covered part of the frame, while the nonresponse bias estimates are labeled "Early/late" and "NRFU" as described above. The noncoverage biases are all positive and generally large. Positive biases are expected because addresses in licensed addresses and coastal counties fish more often than those in the non-coastal and not licensed addresses. The magnitude of the noncoverage biases also vary substantially among states.

Table 4 Percent missing data, by source and state

		Noncoverage	
State	Nonresponse	Licensed	Coastal
Overall	57.6	70.0	7.0
Florida	57.4	63.0	0.0
Massachusetts	53.0	70.0	11.0
New York	60.2	86.0	9.0
North Carolina	56.3	53.0	29.0

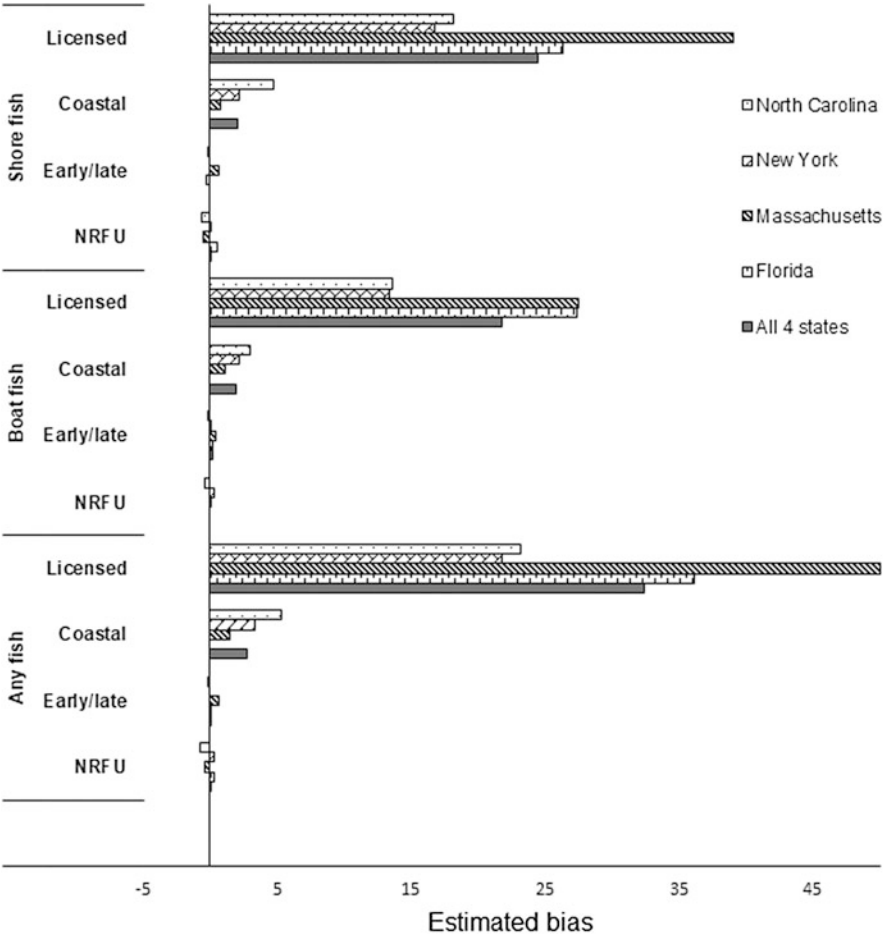


Fig. 1 Estimated nonresponse and noncoverage bias estimates for percent of households reporting any, boat, and shore fishing

The figure clearly shows the noncoverage biases are much larger than the nonresponse biases. The biases from excluding addresses without a license are especially large. For example, the overall noncoverage bias resulting from exclusion of non-coastal addresses is 2.8% points. When addresses without a license match are excluded, the estimated bias is 32.4% points. The corresponding nonresponse biases are less than 0.2% points.

Another key result is that the rate of missing data is a poor indicator of the potential bias from the different sources. For example, for the estimate of fishing prevalence (any fish), excluding non-coastal counties in North Carolina from the sample results in a bias that is more than 13 times higher than the nonresponse bias despite having a much lower rate of missing data. When we examine the joint effect of nonresponse and noncoverage, the dominant contribution of noncoverage is apparent. Table 3 shows that nonresponse bias is very small for all the strata used for simulating noncoverage bias (coastal and licensed). Because nonresponse bias was so small compared to noncoverage bias, we did not directly try to simulate the correlation between the two types of bias. If we assume the effects of the two sources of missing data are independent, an assumption of additive biases that probably overestimates bias, the nonresponse bias adds only slightly to the large positive biases due to noncoverage. At least in the FES, the effect of reducing coverage even to the coastal areas would swamp any nonresponse bias.

4 Bounds on Bias Estimates

The relatively low nonresponse bias estimates are not surprising. An earlier NRFU study done in 2012–2013 in the same four states also found no significant nonresponse bias. Other research also found that excluding non-coastal counties led to higher than desired noncoverage bias and that only including licensed addresses had substantial biases. As a result of these earlier findings, the decision was made to cover all addresses in the FES.

Despite these findings, the response rates in the FES are still low enough that nonresponse bias remains a concern (Stokes et al., 2021). When the fishing effort survey transitioned from a telephone survey to a mail survey, the estimates of the percent of adults fishing increased two- to threefold. The topic interest bias hypothesis was viewed as a realistic cause because the sampled households could see the whole survey immediately and understand the questions being asked.

It is important to understand the FES is subject to other nonsampling errors such as recall error (Andrews et al., 2018). However, recall and many of the other nonsampling errors would tend to reduce the proportion of estimated people who fished.

Since we are interested in looking at bounds, we construct bounds on the nonresponse bias with the topic interest hypothesis in mind. Early on Cochran (1977) discusses bounding the estimated nonresponse bias of a proportion and concluded the bounds could be “distressingly wide" if nonresponse was not negligible.

Following this approach but only allowing the bias to be positive, consistent with the topic bias hypothesis, we obtain the maximum possible bias in this direction. This value is derived by assuming that all fishing households in the sample responded (i.e., 100% response rate for fishing households) and that all nonrespondents were non-fishing households. For example, the FES NRFU had a 45.3% response rate across the four states, so we assume the remaining 54.7% did not fish. With this assumption, the maximum bias is 6.1% points for any fishing, 3.8% points for boat fishing, and 4.5% points for shore fishing. These are large biases relative to the size of the estimates given in Table 2, but assuming a 100% response rate for those who fish is extreme. Furthermore, even these extreme nonresponse assumptions result in nonresponse biases that are far smaller than the noncoverage biases in Fig. 1.

The bounding approach of Montaquila et al. (2008) can be modified to give more insight into the possible bias. Define the response propensities for those who fish (any fish, shore fish, or boat fish) to be ϕ_1 and for those who do not fish to be ϕ_2 . Let P be the proportion who fish. The expected response rate is $\bar{\phi} = P\phi_1 + (1 - P)\phi_2$. Taking expectations of the estimated proportion ($\hat{p}$) over both the sample design and response mechanism, the bias of an estimated proportion is

$$Bias(\hat{p}) = P(\phi_1\bar{\phi}^{-1} - 1). \tag{3}$$

See Hedlin (2020) who derives the same expression.

These equalities are used to show how the response propensity or response rate in the any fish group (ϕ_1) and the bias is related. We take the NRFU as the basis for our values, with the NRFU overall response rate of $\bar{\phi} = 45.3\%$, and its estimate of the proportion who did any fishing of $\hat{p} = 0.112$. Plugging in these values, we obtain the ϕ_1 required to produce a bias of a specified amount shown in Fig. 2. The figure also shows shore fishing ($\hat{p} = 0.083$) and boat fishing ($\hat{p} = .0.069$) curves. The estimate is unbiased if the responses rates are equal ($\phi_1 = \phi_2$).

The maximum positive nonresponse bias for each type of fishing is achieved when the response rate for those who fish (any fish, shore fish, or boat fish) is 100%. For example, for any fish, this bias is 6.1% points and is achieved when any fish $\phi_1 = 100\%$, as mentioned above. More reasonable values for ϕ_1 are some multiple of the response rate for those who do not fish. For example, bounds might be set by allowing ϕ_1 to be $1.2\phi_2$ (a response rate of 54.4%), $1.4\phi_2$ (63.4%), and $1.6\phi_2$ (72.5%). In practice, even a multiple of 1.2 is unusual and would imply a large topic interest effect.

Inspecting Fig. 2 at the points where the curves intersect these response rates shows the biases are relatively small. For example, for shore fishing, the biases are: 1.4, 2.4, and 3.1% points, respectively. For boat fishing, the biases are smaller.

The approach used by Hedlin (2020) bounds the potential nonresponse bias using Eq. 2 by considering different values of $\rho_{\phi,y}$. Hedlin shows that unless the correlation is high, the relative bias is small when the mean response propensity (response rate) is greater than 30%.

Applying this method, we assume the people who fish have a response propensity of ϕ_1, and the overall response propensity is $\bar{\phi}$. The correlation is

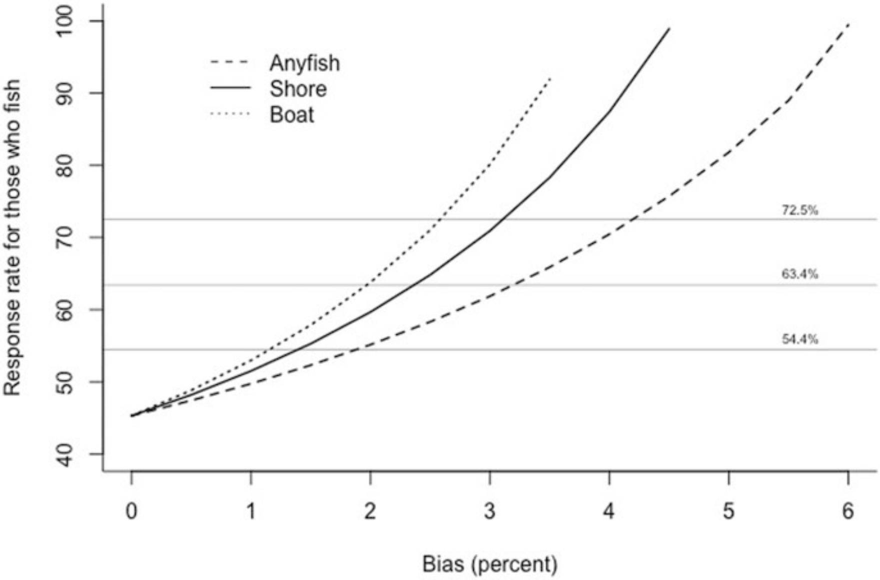


Fig. 2 Relationship between bias of estimate of any fishing, shore fishing, and boat fishing and response rate of those who fish, when overall response rate is 45.3% and estimated fishing proportions are 0.112, 0.083 and 0.069, respectively

$$\rho_{\phi,y} = \frac{P(\phi_1 - \bar{\phi})}{\sqrt{\bar{\phi}(1-\bar{\phi})P(1-P)}} \tag{4}$$

If we assume the extreme bound of $\phi_1 = 100\%$ and substitute for the other quantities in Eq. 4 using the NRFU values from above, the correlation for any fishing is 0.39, for shore fishing is 0.33, and for boat fishing is 0.30. With $\phi_1 = 1.6\phi_2$ (72.5%), the correlations are less than 0.20 for all three statistics. In other words, the correlations required to produce very large nonresponse biases are consistent with very extreme response rate assumptions for the fishing households.

Throughout this evaluation, the sample and weighting methods used to reduce potential nonresponse have not been taken into consideration. For example, the stratification by coastal geography and by license-match status have proven to be very effective in terms of identifying addresses with higher proportions of fishing as shown in Table 2. If the above analysis were repeated within stratum rather than overall, the within-stratum homogeneity of the fishing proportions and response rates would result in even less potential nonresponse bias.

For noncoverage, the bias estimates given earlier are those that we would obtain if no special weighting adjustments were used to reduce the bias. As a result, we discuss the potential to reduce noncoverage biases by weighting. This serves two purposes. First, it provides an alternative way to judge the size of the noncoverage bias if the frame only included coastal addresses or license addresses. Second, it provides another angle on our research goal of exploring the relative magnitude of the nonresponse and noncoverage biases under conditions more favorable to noncoverage.

For the exclusion of addresses in non-coastal areas, an adjustment like that used for nonresident anglers could be employed using data from the APAIS. However, adjusting to a relatively small sample, such as the APAIS, is less efficient than using totals from a census or a very large, high-response rate survey like the ACS. Furthermore, although calibration generally reduces biases for estimates of totals, it is much less effective for estimates of proportions such as the proportion who fish. To be effective for adjusting estimates of proportions such as those studied here, the calibration data would need to be broken into classes or cells with differential coverage rates. Small surveys do not have adequate sample size to provide accurate estimates by classes. The exclusion of those in addresses that do not match to a fishing license is even more problematic. Asking license status in an in-person intercept survey such as the APAIS is fraught with problems since fishing without a license is illegal in many cases. As shown by Tourangeau and Yan (2007), this is precisely the situation in which large biases are common.

5 Discussion

Our analysis explores both nonresponse and noncoverage biases for the FES. The nonresponse bias analysis was feasible largely due to a nonresponse follow-up study. Noncoverage biases were estimated by artificially excluding data that were collected, where including only coastal areas or including only addresses with fishing licenses are designs that have been examined in practice because they are efficient in terms of finding anglers to complete the survey.

The nonresponse bias for the FES is relatively small except under very unexpected assumptions. The 2020 NRFU study and the analysis of the early and late respondents find only small biases. This finding is consistent with an earlier NRFU study. Both of those studies found no evidence to support the topic interest bias hypothesis that would result in overestimates of fishing prevalence.

When bounds based on different assumptions about the response rates for the people who fished and those who did not fish were constructed, nonresponse bias remains small under reasonable assumptions. Substantial nonresponse bias occurs only under the most extreme and unrealistic assumptions (such as assuming everyone who fished responded to the survey). We also translate these response rate assumptions into correlations between fishing and responding, showing again the nonresponse biases are small under realistic assumptions. For the four states in this study, the nonresponse bias for both shore fishing and boat fishing is likely to be no greater than 1–2% points even under relatively unusual assumptions (ratios of response rates of 1.2). This finding contrasts with the very large biases associated with noncoverage in the FES. Large biases occur when the sample is restricted to either just coastal addresses or to addresses with licenses. Weighting methods to reduce these biases are feasible, but available external data sources are unlikely to reduce the noncoverage biases to be close to the magnitude of the nonresponse biases.

An important conclusion is that all missing data are not equivalent. In the FES at least, data that are missing because of noncoverage result in much larger biases than data that are missing due to nonresponse. The noncoverage biases are large even though the missing data rate for the coastal estimates average 7% and go up to 30% in North Carolina. The license noncoverage biases are much larger than even those in the coastal stratum. Despite missing data due to nonresponse being over 50%, the nonresponse bias is small. Clearly, missing data rates are not predictive of biases.

The important determinant of bias is the difference in the characteristics of the missing and non-missing data. This concept is simple to understand for noncoverage. If the percent of people who fish is very different for the covered and non-covered, then the bias will be large and weighting adjustments are unlikely to reduce the biases significantly. The bias estimates for the noncoverage due to sampling only coastal areas and licensed-matched addresses were very large for the percent who fished.

While the post-survey distinction of a respondent stratum and nonrespondent stratum has value in formulating a bias expression like Eq. 1 for nonresponse, it is a model that has limited conceptual appeal. If all sampled units have some chance of responding, then a nonrespondent stratum does not exist. Instead, the response propensity model given by Eq. 2 is more consistent with data collection experiences. For example, multiple attempts are made to interview the same units because the decision to participate is not fixed and may depend on a host of factors.

Another difference is that weighting adjustments for noncoverage rely exclusively on external data, but nonresponse adjustments can use data collected in the survey itself as well as external data. Typical nonresponse weighting class adjustments transfer weights from the nonrespondents to the respondents in the same class before calibration to external data. The survey data allow examination of the homogeneity of the response propensities within the classes. Noncoverage weighting adjustments also typically use classes to reduce bias, but the model for the adjustment cannot be evaluated from the survey data itself. Thus, weighting adjustments may be more effective for nonresponse.

These findings suggest that noncoverage may result in larger biases than nonresponse, even when the missing data rates due to noncoverage are much lower than those due to nonresponse. Surveys need to consider more than just missing data rates when deciding on survey designs and levels of effort. With the FES, saving resources by reducing coverage and using those resources to increase the response rate to the survey would likely increase the biases of the estimates. Each survey needs to evaluate its potential for biases, but the FES results reveal that relying on missing data rates to do this may be very misleading.

References

Andrews, W. R. (2021). *Evaluating nonresponse bias in the MRIP fishing effort survey*. https://apps-st.fisheries.noaa.gov/pims/main/public?method=DOWNLOAD_FR_DATA&record_id=2018

Andrews, W. R., Papacostas, K. J., & Foster, J. (2018). A comparison of recall error in recreational fisheries surveys with one- and two-month reference periods. *North American Journal of Fisheries Management, 38*, 1284–1298.

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*, 209–226.

Battaglia, M. P., Dillman, D. A., Frankel, M. R., Harter, R., Buskirk, T. D., McPhee, C. B., DeMatteis, J. M., & Yancey, T. (2016) Sampling, data collection, and weighting procedures for address-based sample surveys. *Journal of Survey Statistics and Methodology, 4*, 476–500.

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics, 4*, 251–260.

Cochran, W. (1977). *Sampling techniques*. Wiley.

Couper, M., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an internet survey. *Social Science Research, 36*, 131–148.

Groves, R. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*, 646–675.

Groves, R., & Couper, M. (1998). *Nonresponse in household interview surveys*. Wiley.

Groves, R., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly, 72*, 167–189.

Groves, R., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly, 68*, 2–31.

Hedlin, D. (2020). Is there a 'safe area' where the nonresponse rate has only a modest effect on bias despite non-ignorable nonresponse. *International Statistical Review, 88*, 642–657.

Lessler, J., & Kalsbeek, W. (1992). *Nonsampling error in surveys*. Wiley.

Little, R., & Rubin, D. (2019) *Statistical analysis with missing data*. Wiley.

Lohr, S. (2019). *Sampling: Design and analysis*. Chapman and Hall/CRC.

Luiten, A., Hox, J., & de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics, 36*, 469–487.

Montaquila, J., Brick, J. M., Hagedorn, M., Kennedy, C., & Keeter, S. (2008). Aspects of nonresponse bias in RDD telephone surveys. In: L. Tucker, B. de Leeuw, J. Lavrakas, L. Sangser (Eds.), *Advances in telephone survey methodology* (pp. 561–586). Wiley.

Papacostas, K., & Foster, J. (2018). National Marine Fisheries Service's Marine Recreational Information Program survey design and statistical methods for estimation of recreational fisheries catch and effort. https://www.fisheries.noaa.gov/resource/document/survey-design-and-statistical-methods-estimation-recreational-fisheries-catch-and

Särndal, C. E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. Wiley.

Scherpenzeel, A., & Bethlehem, J. (2010). How representative are online panels? Problems of coverage and selection and possible solutions. In *Social and behavioral research and the internet* (pp. 105–132).

Stedman, R. C., Connelly, N. A., Heberlein, T. A., Decker, D. J., & Allred, S. B. (2019). The end of the (research) world as we know it? Understanding and coping with declining response rates to mail surveys. *Society & Natural Resources, 32*, 1139–1154.

Stokes, S. L., Williams, B. M., McShane, R. P., & Zalsha, S. (2021). The impact of nonsampling errors on estimators of catch from electronic reporting systems. *Journal of Survey Statistics and Methodology, 9*, 159–184.

Stoop, I. (2005). *The hunt for the last respondent*. Sociaal en Cultureel Planbureau.

Thornberry, O., & Massey, J. (1988). Trends in United States telephone coverage across time and subgroups. In: R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, J. Waksberg (Eds.), *Telephone survey methodology* (pp. 25–49). Wiley.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859–883.

Tucker, C, Brick, J. M., & Meekins, B. (2007). Household telephone service and usage patterns in the United States in 2004: Implications for telephone samples. *Public Opinion Quarterly, 71*, 3–22.

Williams, D., & Brick, J. M. (2017). Trends in US face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology, 6*, 186–211

Triple System Estimation with Erroneous Enumerations

Paul P. Biemer, G. Gordon Brown, and Christopher Wiesen

Abstract A central assumption in population coverage error estimation is that nonresidential units are not counted (i.e., no erroneous enumerations) and, thus, the only remaining errors are omissions. This assumption is violated in many situations, notably in the US Census 2000, where undetected erroneous enumerations were a primary reason that the post-enumeration survey (PES) results could not be used in census undercount adjustments. This paper develops a latent class modeling approach that allows for varying levels of undetected erroneous enumerations in one of the population lists. Our approach requires three population lists which may be the Census, the PES, and a list derived from merging records from administrative systems. The resulting data take the form of an incomplete contingency table which can be represented by a latent class model where the latent variable is an individual's true status (i.e., resident or nonresident of the population). Latent class analysis is used to estimate the expected values of the observed cells of this table and then to project these estimates onto the unobserved cells in order to estimate the total number of population members. Using artificial populations, the improvement in mean squared error using this approach is evaluated and compared to other modeling approaches from the capture-recapture literature.

P. P. Biemer (✉)
RTI International, Raleigh, NC, USA
e-mail: ppb@rti.org

G. G. Brown
SAS Institute, Cary, NC, USA

C. Wiesen
Odum Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: chris_wiesen@unc.edu

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_8

1 Background

An important issue in estimating the number of persons residing in an area from census data is the evaluation of census coverage error. Various techniques using multiple input sources have been developed for estimating the error in the census count. One widely used method is dual-system estimation (Sekar and Deming, 1949). With this approach, a post-enumeration survey (PES) of the population is conducted, and the persons in the PES are matched to persons in the census enumeration. For the 2000 US Census, the PES involved enumeration of the occupants of 300,000 households in a random national sample of 12,500 housing blocks (Hogan et al., 2002).

For the dual system estimation (DSE) approach, data from the enumeration process and the PES are combined in a 2 × 2 table of counts cross-classifying the presence or absence of persons in the census enumeration with their presence or absence in the PES. The DSE approach provides an estimator of the number of persons in the fourth cell of this table which corresponds to persons missed by both the census and the PES. The sum of the three observed and one estimated cells of the Census-PES cross-classification table provides an estimate of the total population count.

Three key assumptions are made for the DSE approach:

1. *Independence.* The probability of inclusion of an individual on the second list (the PES) does not depend upon inclusion or exclusion from the first list (the census). Failure of this assumption will induce correlations between the errors in the two lists, sometimes referred to as behavioral correlation (Wolter, 1986). If a third list is available, the independence assumption can be tested (see, e.g., Bishop et al., 1975, Chapter 6). Zaslavsky and Wolfgang (1993) provide models for dealing with the behavioral correlation in three systems.
2. *Homogeneity.* The probability of inclusion on a list does not vary from individual to individual. Although this assumption is known not to hold for the population as a whole, various strategies have been used to address the problem of heterogeneous enumeration probabilities, including post-stratification (Sekar and Deming, 1949) and logistic regression (Alho et al., 1993). Methods involving three systems have been explored by Darroch et al. (1993), Fienberg et al. (1999), and Chao and Tsay (1998). This is the correlation bias problem (see, e.g., Wolter, 1986).
3. *Perfect enumeration and matching.* Individuals in both lists are all population members that can be accurately matched between the two lists, and any nonresidents who have been erroneously enumerated can be identified and eliminated. Matching errors can be fairly substantial (Biemer and Davis, 1991a), and methods for dealing with these can be found in Biemer (1988) and Ding and Fienberg (1992). Biemer and Davis (1991b) show how undetected erroneous enumerations can seriously bias the estimates of census coverage error. Usually, the bias is positive, resulting in overcorrecting the census counts for the undercount.

The models considered in this paper seek to address failures of all three assumptions to some extent, but particularly assumptions 1 and 3. The assumption of independence may be relaxed if a third counting system is introduced, for example, an administrative records list (ARL) of persons in the population. Although erroneous enumerations can occur in all three systems, the problem is much greater for administrative lists as will be discussed subsequently. Therefore, the focus in present paper is on erroneous enumerations in the ARL. A subsequent paper (in progress) will extend the ideas of the present paper to erroneous enumerations in all three lists.

Erroneous enumerations (EEs) occur when individuals who are not residents of the target population are erroneously counted as residents. EEs may be persons who were deceased prior to Census Day, born after Census Day, or nonresidents of the target area on Census Day. EEs also include geocoding (or location) errors, duplicated persons, and fictitious or nonexistent persons. In the following, we will refer to all of these entities as nonresidents regardless of their source. Further, any nonresident who is classified as a resident will be called an EE.

In this paper, a statistical framework for dealing with undetected EEs using a latent class modeling (LCM) approach is presented. Latent class models are essentially log-linear models where one or more of the variables are latent or unobservable. Since traditional capture-recapture models can also be written as log-linear models, LCMs are straightforward extensions of the traditional capture-recapture models. LCMs provide a convenient statistical framework for specifying capture-recapture models with undetected EEs as well as missed residents in all three systems. Unfortunately, the identifiability of LCM for population size estimation has never been explored in the literature. Further, little is known about the statistical properties of the LCM estimators in census coverage error evaluation applications.

To simplify the exposition of the general ideas and the theory, the paper is confined to the situation where undetected EEs are present only in the ARL. That is, we assume the census process is successful in identifying and removing EEs in the census and the PES. This somewhat restricted class of models represents an important generalization over the traditional assumption of no EEs and provides a useful alternative to other dual and triple system models for applications where the numbers of EEs in the census and PES are small compared to the number in the ARL. Further, study of this restricted case will provide important insights regarding the much more complex case of EEs in all three systems, the ultimate goal of this research.

The problem of EEs in the census process has long been recognized, and adjustment of the DSE of N for EEs is an essential component of the estimation process. A special survey, referred to as the E-sample (see, e.g., Hogan, 1993), is conducted simultaneously with the PES in order to estimate the number of EEs in the census and adjust the DSE for them. Despite these efforts, some EEs are not identified and are included in the dual system thereby inducing bias in the estimates of N.

In 2000, the US Census Bureau Evaluation Followup (EFU) estimated that about 1.8 million enumerations in the PES were actually EEs (ESCAP, 2001). Further, 365,000 persons classified as EEs were in fact correct enumerations. Based on these results, the Census Bureau concluded that the net undercount was overstated by three to four million persons and, thus, adjustments to the census count on the basis of the PES would substantially overcorrect the population counts in many areas. For the 1990 Census, Biemer and Davis (1991b) reported that the level of misclassified EEs in the 1990 PES exceeded 5% of the PES count for many areas of the country. In the worst areas, the Northeast urban and the Midwest non-central city areas, the EE rate exceeded 20%.

Although the availability of an ARL as the third list provides the means for modeling the correlation between the census and the PES, the risk of including EEs in the estimation process is substantially increased since an ARL may contain many non-population members and duplicate persons that are difficult to accurately identify and remove from the process. An example of an ARL that is being considered for census undercount evaluation purposes is the Census Bureau's Statistical Administrative Records System (StARS; see Judson, 2000). The StARS consists of seven merged databases including IRS returns, selective service files, Medicare enrollment database, Indian Health Service patient file, and the HUD tenant resident certification system. Since individuals may be on two or more of these lists, the potential for duplicate persons on StARS is quite high. The address information on the files may be incomplete or erroneous, thus increasing the opportunity for geocoding errors. The files may not be completely current, which can cause the application of the Census Residency Rules to the StARS to be problematic. Although many EEs can be identified through intensive field follow-up, such evaluations are costly and quite time-consuming, considering the schedule for producing the census counts. In addition, for a large-scale implementation, the error rate of such a field verification process is likely to be unacceptable for census adjustment purposes.

In the next section, we introduce the notation and describe the models that will be used in our study. This includes both models with and without EEs. Section 3 describes the estimators of the total population size that can be obtained from the models and provides an illustration of the ideas using real data. Section 4 reports on an extensive simulation study using artificial populations which compared the estimators under various adverse population conditions. Finally, in Sect. 5, we summarize the results and discuss their implications for using the estimators in census coverage error evaluation studies.

2 Models

This section briefly describes a few of the basic models in the capture-recapture literature, elaborating on two models that will be used extensively in our work. In addition, a new class of capture-recapture models based upon latent class analysis is proposed, and the identifiability and utility of these models are explored.

Let U denote the persons in a target area (i.e., the area to be enumerated by the census). This includes the union of persons included on at least one of the three lists as well as all residents in the area who are not included on any of the lists. Thus, U denotes all actual residents of the target area as well as nonresidents and fictitious persons that are erroneously included on the lists. Let P denote all persons who are true residents of the area and should be counted. Let E denote the complement of P, (i.e., $E = U \sim P$), i.e., persons who are not residents of the target area and should not be counted. The number of persons in U will be denoted by M and the number of persons in P by N. The objective of this research work is to obtain a robust estimate of N based upon data from the census, PES, and ARL when the members of E are classified as in P (i.e, EEs) and the members of P are either missed or classified as in E (i.e., omissions). As previously mentioned, in this paper we assume that EEs enter the estimation process solely as the result of a misclassification by the ARL.

Let X_i denote a dichotomous variable defined for the ith person in U, where $X_i = 1$ if person $i \in P$ and $X_i = 0$ if person $i \in E$. We assume that X_i is an unknown and unobservable (latent) variable for all $i \in U$. For triple system estimation, there are three indicators of X_i corresponding to the census denoted by A_i, the PES denoted by B_i, and the ARL denoted by C_i. Like X_i, each indicator variable takes on the value 1 if person i is classified as in P and 0 if classified in E. Note that the definitions of X_i and its indicators depend upon the definition of the target area. For notational convenience, in the following, we will drop the subscript i when it is clear we are referring to an individual in the universe.

2.1 Model Assumptions and Notation

Let π_x denote $P(X = 1)$, $\pi_{A=a|X=x} = \pi_{a|x} = P(A = a|X = x)$ with analogous definitions for $\pi_{b|x}$ and $\pi_{c|x}$, where x, a, b, and c can be either 1 or 0. The probability the census correctly enumerates a resident in U is $\pi_{A=1|X=1}$, referred to as the correct enumeration probability. An EE occurs when a person in E is classified as in P. Thus, the probability of an EE in the census is $\pi_{A=1|X=0}$.

Let $XABC$ denote the (unobservable) cross-classification table for the variables X, A, B, and C for all $i \in U$, and let (x, a, b, c) denote the cell associated with $X = x$, $A = a$, $B = b$, and $C = c$ in this table. Define $\pi_{xabc} = P(x, a, b, c)$ as the expected proportion in cell (x, a, b, c) and note that π_{abc} can be expressed as

$$\pi_{xabc} = \pi_x \pi_{a|x} \pi_{b|ax} \pi_{c|abx}. \tag{1}$$

Although the $XABC$ table is not observable, (1) is still useful to specify the cell probability for the observable ABC table, i.e.,

$$\pi_{abc} = \sum_x \pi_x \pi_{a|x} \pi_{b|ax} \pi_{c|abx}. \tag{2}$$

When all parameters in this likelihood are identifiable, they can be estimated using maximum likelihood estimation techniques. However, the unrestricted model (2) contains 95 parameters, but only 47 degrees of freedom are available in the ABC table; thus, the model is substantially over-parameterized and not identifiable. Restrictions on the probabilities will be introduced to reduce the number of parameters associated with the model and obtain an identifiable model. The plausibility of these restrictions and other model assumptions for census coverage error evaluation applications is a key issue for the modeling process.

In the next section, we consider models that assume no EEs in the census estimation process, i.e., $\pi_{A=1|X=0} = \pi_{B=1|X=0} = \pi_{C=1|X=0} = 0$. These are traditional capture-recapture models that are appropriate when the probability of undetected EEs in the three systems is negligibly small. In that case, we can ignore the latent variable X in the analysis and consider models for π_{abc} rather than π_{xabc}.

2.2 Models with No Erroneous Enumerations

In this section, we present a few classic closed population models as defined in Pollock et al. (1990) and discuss their utility for coverage error estimation. In order to remain consistent with census terminology, we will use the term "enumeration probability" instead of the traditional term "capture probability" used in the capture-recapture literature. In addition, the models are written using the notation introduced in Sect. 2.1.

2.2.1 Model M$_0$: Equal Catchability Model

Model M$_0$ assumes that every individual in the population has the same probability of being enumerated on each sampling occasion, i.e., $\pi_A = \pi_B = \pi_C = \pi_1$, and enumerations at future time points are independent of previous enumerations. For this model, $\pi_{abc} = \pi_1^{a+b+c}(1 - \pi_1)^{3-a-b-c}$. Although model M$_0$ is very unlikely to hold in practice, it is still important as the basis for all closed population models. Instances where M$_0$ has been used in practice are quite rare; however, it should be noted that, when the enumeration probabilities are in fact equal, inference obtained from model M$_0$ is nearly identical to inference obtained from the next model we will consider—namely, M$_t$ or Schnabel's model.

2.2.2 Model M$_t$: Schnabel's Model

Schnabel (1938) originally developed the M$_t$ model for situations where it may be assumed that every individual in the population has the same enumeration probability within a list, but enumeration probabilities may vary across the lists. As with the M$_0$ model, future enumerations are assumed to be independent of previous enumerations. This model does not allow heterogeneity of enumeration rates within a list or a behavioral response to capture. For this model, $\pi_{abc} = \pi_a \pi_b \pi_c$. The next model we consider allows enumeration probabilities for subsequent enumerations to depend upon previous enumerations.

2.2.3 Model M$_b$: The Trap Response Model

For the M$_b$ model, previously enumerated individuals can have future enumeration probabilities that differ from previously unenumerated individuals. Consequently, enumeration outcomes across the three lists may be correlated. The model specifies that $P(A = 1) = P(B = 1|A = 0) = P(C = 1|A = 0, B = 0) = \pi_u$. Note that π_u is the probability of enumeration for any individual not previously enumerated. The corresponding probability for individuals previously enumerated by the census, the PES, or both is $P(B = 1|A = 1) = P(C = 1|A = 1 \text{ or } B = 1) = \pi_e$. Thus, the cell probabilities for this model can be written as products of π_u, $(1 - \pi_u)$, π_e, and $(1 - \pi_e)$. As an example, $P(A = 1, B = 1, C = 1) = \pi_i \pi_e^2$, $P(A = 0, B = 1, C = 0) = (1 - \pi_u)\pi_e(1 - \pi_e)$, and so on.

In the population census context, correlations may be introduced between the census and the PES due to the reactions of individuals to the census enumeration process. For example, individuals who were enumerated in the census may have enjoyed the experience or may determine that any fears they may have had about the process were unfounded. This reaction might cause their probabilities of enumeration in the PES to be higher than for individuals missed by the census—referred to as "trap happy" behavior. Conversely, individuals whose experience with the census enumeration process was less than favorable might engage in avoidance or "trap shy"5t4rt-= behavior in the PES.

In general, trap shy behavior causes enumeration rates for previously enumerated individuals to decrease, leading to overestimation of the population size. Trap happy behavior causes enumeration rates of previously enumerated individuals to increase, leading to underestimation of the population size. Because of its inherent limitations for population census applications, the M$_b$ model is extended in the next section.

2.2.4 Model M$_{tAB}$: Non-stationary Behavioral Response Models

A natural extension of the M$_b$ is the M$_{tb}$ model which has both time variation and behavioral response to the enumeration process. Although the standard form of the M$_{tb}$ model is not identifiable, a very useful and identifiable model, the M$_{tAB}$ model, can be obtained by imposing a plausible restriction on the M$_{tb}$ model. The index AB on the M$_{tAB}$ model is used to indicate that the model contains one interaction term representing behavioral correlation between the A- and B-lists and that the C-list is assumed to be independent of the other two lists. Under this model $\pi_{abc} = \pi_a \pi_{a|b} \pi_c$ where $\pi_{a|b}$ may differ from π_a.

The motivation for this model stems from the realization that behavioral correlation is likely to be much greater between the census and the PES than for either of these enumerations and the ARL. This is because enumeration by the census or PES depend largely on an individual's attitude toward being interviewed and census participation in general, whereas the listing of an individual on an administrative record usually depends upon factors that provide more direct benefits to the individual; examples are Social Security and Medicare benefits, unemployment compensation, automobile ownership, payment of taxes, and so on. Therefore, whether an individual appears on the ARL should not be greatly influenced by the individual's choice or ability to participate in the census or PES.

Thus, it seems reasonable to assume that being listed on the ARL is uncorrelated with enumeration by either the census or the PES and, consequently, the interaction terms AC and BC are relatively small and negligible. This assumption also greatly reduces the complexity of the models. When both the AC and BC interactions are included in the model, identifiability problems result that can only be remedied by adding parameter constraints which tend to be implausible for census applications. It is possible, however, to fit models that allow for correlations between enumerations in all three lists (see, e.g., Zaslavsky & Wolfgang, 1993), and one such model will be considered later in Sect. 4. In this paper, the M_t and M_{tAB} models are examined in some detail since they appear to be the most likely of the traditional closed population models to mirror triple systems data.

2.3 Models with Erroneous Enumerations in the ARL

For the models presented in this section, several assumptions made for the traditional closed population models are relaxed. We still assume that $\pi_{A=1|X=0} = \pi_{B=1|X=0} = 0$, but now we allow $\pi_{C=1|X=0} > 0$, i.e., EEs are allowed to enter into the census estimation process through the ARL or the C-list. Thus, we define a new class of population size estimation models which we refer to as L-models. The L-model assumptions essentially parallel those made for the M-models discussed previously except now we introduce a latent "true enumeration status" variable to account for the possibility that nonresidents may be misclassified as residents by their inclusion on the C-list.

The assumption of no undetected EEs in the census and PES is consistent with traditional assumptions made for these two systems, but, as previously discussed, these assumptions are unlikely to hold in some enumeration areas. In this regard, the L-models we consider in this paper suffer from the same limitations as the M-models with respect to EEs in the census and PES. The L-models should be preferred when the majority of EEs in the estimation process are introduced through the C-list, as is likely when the C-list is the ARL. It is possible to extend the general ideas described here for modeling EEs in the C-list to the case where non-negligible EEs occur in the A- and B-lists. This research is currently underway and will be reported in a subsequent paper.

Another issue for the L-models is the value of $P(C = 1|X = 0)$, which is the probability that a nonresident in the population is an EE in the ARL. Since all models considered in this paper assume that EEs only enter into the estimation process through the ARL, the only EEs that are of any consequence in the analysis are those that are brought in through the ARL. This implies that, given that we observe an EE in the data, the probability that it was introduced through the ARL is 1. Thus, we know that $P(C = 1|X = 0) = 1$ and $P(A = 1|X = 0) = P(B = 1|[X = 0) = 0$ for all L-models considered.

2.3.1 Model L$_0$: Equal Catchability LCM

The L$_0$ model extends the M$_0$ model to include EEs in the C-list. Like the M$_0$ model, the L$_0$ assumes that every individual in the target population has the same probability of enumeration on all three lists. An additional parameter is included to account for the EEs in the C-list. Let π_x denote $P(X = 1)$, $\pi_{A=1|X=1} = \pi_{B=1|X=1} = \pi_{C=1|X=1} = \pi_1$, and $\pi_{C=1|X=0} = \pi_2$. Then

$$\pi_{xabc} = \pi_x \pi_1^{a+b+c}(1-\pi_1)^{3-a-b-c} + (1+\pi_x)\pi_2^c(1-\pi_2)^{(1-c)}(1-a)(1-b). \tag{3}$$

This model parallels the M$_0$ model and is the least complex of the L-models. Like the M$_0$ model, which is also unlikely to hold in practice, it will not be considered further in this paper.

2.3.2 Model L$_t$: Non-stationary, Independent LCM

The L$_t$ model extends model M$_t$ to reflect EEs in the C-list. Thus, we have

$$\pi_{xabc} = \pi_x \pi_{a|x} \pi_{b|x} \pi_{c|x} \tag{4}$$

which corresponds closely to the classical latent class model for the ABC table except for the structural zero in the 000 cell. If there is no correlation between the A- and B-lists, then the L$_t$ model should provide a good estimate of N. If there is correlation between the A- and B-lists, then the Lt model is not appropriate, and the quality of inference will decline as the magnitude of the AB interaction increases. The L$_t$ model is the least complex of the L-models that may hold, at least approximately, in practice and so is investigated in this paper.

2.3.3 Model L$_{tAB}$: Non-stationary, Behavioral Response Latent Class Model

Among the models considered in this paper, the L$_{tAB}$ is the most complex LCM and the most likely to accurately represent triple systems data. The L$_{tAB}$ model accounts for correlation between the A- and B-lists and for list-dependent enumeration probabilities. Under this model, $\pi_{xabc} = \pi_x \pi_{a|x} \pi_{b|xa} \pi_{c|x}$. Unfortunately, the L_{tAB} model has several parameters that are not estimable when using the full or unconditional version of the likelihood. Additionally, the conditional version of the likelihood, which conditions on the seven observable cells and is used in latent class analysis, is not identifiable. Lack of identifiability implies that additional information must be provided in order to obtain meaningful inference from the L_{tAB} model.

Our solution to the identifiability problem is to assume that the parameter $\pi_{X=1|C=0}$ is equal to a known constant, say γ. Knowledge of $\pi_{X=1|C=0}$ means the conditional likelihood is fully identifiable and allows all parameters of the unconditional likelihood to be estimated. For the majority of our study, we will assume that $\pi_{X=1|C=0} = \gamma$ with negligible error. We also present an example of a potential method for determining γ. A generalization of the L_{tAB} model that allows non-negligible error in the estimate of $\pi_{X=1|C=0}$ is beyond the scope of this paper.

3 Estimation

3.1 Estimating N

Both the method of moments and maximum likelihood estimation methods have been used for parameter estimation in the literature for capture-recapture models. Method of moments estimates are often easy to calculate but can have undesirable properties such as large variance or large bias. Consequently, maximum likelihood estimates are often preferred. The standard MLE method consists of using the conditional likelihood of the model being considered (see White and Burnham, 1999). In this method, the enumeration probabilities are estimated, and an estimate of population size is derived using a Horvitz-Thompson estimator.

For the M-models, the estimator of N has the general form

$$\hat{N}_M = \frac{n}{(1 - \hat{\pi}_{000})} \tag{5}$$

where n is the number of persons enumerated (all assumed to be in P) and $\hat{\pi}_{000}$ is the estimate of the proportion of persons in P in the 000 cell of the ABC table.

For the L-models, we use two methods for estimating N. One method involves estimating π_{000} and π_x using the conditional likelihood for the ABC table (see, for example, Section 6.3 in Bishop et al., 1975). Denoting these estimates by $\hat{\pi}_{000}$ and $\hat{\pi}_x$, respectively, leads to the estimator

$$\hat{N}_L = \frac{m}{(1 - \hat{\pi}_{000})} \hat{\pi}_x \tag{6}$$

where m is the number of persons enumerated (including EEs).

The other method uses the full likelihood by performing a search over likely values of M. For this search method, the initial value of M is set to its minimum value, $M = m$. Then, the likelihood is maximized over the other parameters conditional on $M = m$. The process is then repeated for $M = m + k$, for $k = 1, 2, 3, \ldots$, and so on until the global maximum is found for all of the parameters. The multi-modal nature of the likelihood necessitated the use of this simple search algorithm. Let M_{opt} and π_{opt} denote the estimate of M and π_x from this process. Then the estimator of N from the search method is

$$\hat{N}_L = M_{opt}\pi_{opt}. \tag{7}$$

Bishop et al. (1975) and Cormack (1989) show how the M-models can be fit using traditional log-linear analysis. Haberman (1979) provides a similar structure for estimating LCMs using log-linear analysis with latent variables. For example, the L$_t$ model is equivalent to the following hierarchical log-linear model:

$$\log m_{xabc} = u + u_x^X + u_a^A + u_b^B + u_c^C + u_{xa}^{XA} + u_{xb}^{XB} + u_{xc}^{XC}, \tag{8}$$

where $m_{xabc} = m\pi_{xabc}$ and m is the number of enumerated individuals. This model is represented in shorthand notation by including the highest order terms in braces, viz., $\{AX, BX, CX\}$. Likewise, the L$_{tAB}$ model is represented as $\{AX, BX, CX, AB\}$ with constraints as noted above.

In Sect. 4, we illustrate an application of two of the more complex models described in Sect. 2: the M$_{tAB}$ and L$_{tAB}$ models. These models are applied to data from a study conducted by Zaslavsky and Wolfgang (1993). Estimates from the M$_{tAB}$ and L$_{tAB}$ models are compared to the corresponding estimates from a similar model considered in their paper.

3.2 Illustration Using Real Data

In this section, we illustrate the properties of the M$_{tAB}$ and L$_{tAB}$ estimators using the triple system data reported in Zaslavsky and Wolfgang (1993), hereafter referred to as ZW. For comparison purposes, we also compare these two estimators with a similar estimator proposed by ZW which is based upon method of moments estimation principles.

ZW propose several models for estimating population size using triple system data. Three sources of data were considered in their study from the 1988 Dress Rehearsal: the census, the PES, and the ARL. These sources were labeled E, P, and A, respectively, in their study but are re-labeled as the A-, B-, and C-list, respectively, to be consistent with our current notation.

Among the models used by ZW, the one that most closely resembles our M$_{tAB}$ model is their $\alpha_{EP|A}$ model. The primary difference is that $\alpha_{EP|A}$ allows for behavioral correlations among all three lists, while the M$_{tAB}$ model allows correlation only between the A- and B-lists. Specifically, the $\alpha_{EP|A}$ model forces equality between the AB odds ratio conditioned on $C = 1$ and the marginal AB odds ratio as follows:

$$\alpha_{EP|a=1} = \frac{n_{001}n_{111}}{n_{011}n_{101}} = \frac{n_{00+}n_{11+}}{n_{01+}n_{10+}} \tag{9}$$

while M$_{tAB}$ forces three-way equality between the conditional given $C = 0$, conditional given $C = 1$, and marginal odds ratios as follows:

$$\frac{n_{000}n_{110}}{n_{010}n_{100}} = \frac{n_{001}n_{111}}{n_{011}n_{101}} = \frac{n_{00+}n_{11+}}{n_{01+}n_{10+}}, \tag{10}$$

where '+' indicates summation over the index.

The ZW model uses the observed value of the AB odds ratio, conditioned on $C = 1$ (denoted by $\alpha_{EP|a=1}$) as an estimate of the unconditioned odds ratio. ZW first estimate

$$\hat{n}_{00+} = \alpha_{EP|a=1}\frac{n_{01+}n_{10+}}{n_{11+}} \tag{11}$$

and, thus, an estimate of n_{000} is $\hat{n}_{000} = \hat{n}_{00+} - n_{001}$. In their formulation, the AB odds ratio conditioned on $C = 0$ is not restricted, and C-list is assumed to be dependent on the A- and B-list.

The estimator of n_{000} from the M$_{tAB}$ model is

$$\hat{n}_{000} = n\left(\frac{\hat{\pi}_{000}}{1 - \hat{\pi}_{000}}\right) \tag{12}$$

where now $\hat{\pi}_{000}$ is the MLE of π_{000} under the M$_{tAB}$ model. An approximate expression for $\hat{n}_{000}$ which can be compared the estimator based on (11) is derived in the Appendix 1.

Variances of the estimators were estimated using traditional capture-recapture variance estimation techniques for population size such as those described in Seber (1982). The methodology typically used depends on the Taylor series expansion of the Horvitz-Thompson estimate of population size. Program MARK is a software package that calculates parameter estimates and their variances for a wide variety of capture-recapture models (see White and Burnham, 1999) and was used to calculate the traditional variance estimates that are presented in Table 2. Similar procedures were used for estimates derived from the log-linear models which were fit using the latent class analysis software, LEM (Vermunt, 1997).

Although the ZW model is theoretically similar to the M$_{tAB}$ model, the two models can yield very different estimates of population size as shown below. Further, estimates of N from the L$_{tAB}$ model exhibit even greater differences from the ZW model depending upon the size of γ. To illustrate this, we fit ZW's, the

Table 1 Triple system data from the 1988 Dress Rehearsal Census in Louis, Missouri

A	B	C	Owners, 20–29 years	Renters, 30–44 years
0	0	0	N/A	N/A
0	0	1	59	43
0	1	0	8	04
0	1	1	19	13
1	0	0	31	30
1	0	1	19	7
1	1	0	13	69
1	1	1	79	72

Table 2 Estimates of $\hat{n}_{000}$ for ZW, M$_{tAB}$, L$_{tAB}$ ($\gamma = 0.05$), and L$_{tAB}$ ($\gamma = 0.10$)

	Parameter Estimates			Standard Errors of $\hat{N}$		
Post-stratum	EE	$\hat{n}_{000}$	$\hat{N}$	SE	Sim	Mk
Owners, 20–29 years (ZW)	0	130	358	64	17.6	NA
Owners, 20–29 years(M$_{tAB}$)	0	26	254	4.7	7.6	7.5
Owners, 20–29 years (L$_{tAB}$ at $\gamma = 0.05$)	9	22	241	4.3	6.8	6.4
Owners, 20–29 years (L$_{tAB}$ at $\gamma = 0.10$)	18	18	228	3.9	5.8	5.4
Renters, 30–44 years (ZW)	0	305	565	432	38.8	NA
Renters, 30–44 years (M$_{tAB}$)	0	58	318	9.4	13.8	14.1
Renters, 30–44 years (L$_{tAB}$ at $\gamma = 0.05$)	7	49	302	10.0	13.2	11.1
Renters, 30–44 years (L$_{tAB}$ at $\gamma = 0.10$)	14	40	286	8.5	10.9	8.8

M$_{tAB}$ model, and the L$_{tAB}$ model for two data sets in Table 1 reproduced from ZW's Table 1. The first three columns of Table 1 denote the eight cells of the ABC table with the cell counts displayed in columns 4 and 5 for two groups: home owners aged 20–29 years and home renters aged 30–44 years.

For owners, aged 20–29 years, the AB odds ratio estimate conditioned on $C = 1$ is about 12.9, as is the marginal AB odds ratio. When $C = 0$, the AB odds ratio is about 6.8. Under the M$_{tAB}$ model, all AB odds ratios are about 5.8.

Table 2 provides the estimates of n_{000}, EE, and N for the four estimators shown in column 1, namely, ZW, M$_{tAB}$, and L$_{tAB}$ computed at $\gamma = 0.05$ and $\gamma = 0.10$. The "EE" column gives the number of EEs detected by the model in the 001 cell, and the estimated number of residents given in the 000 cell is given in the column labeled $\hat{n}_{000}$ column. Thus, the estimate of N is $m+\ \hat{n}_{000}$-EE given in the column labeled $\hat{N}$.

Three different standard errors are given for the estimates of N which are shown in the last three columns. The first, expressed by "SE," represents the standard error generated by LEM.[1] The second, expressed by "Sim," is the standard error derived by the simulation experiments described in Sect. 4. The third, expressed by "Mk,"

[1] LEM is a software package for fitting log-linear models with latent variables written by Jeroen Vermunt, Tilburg University, Tilburg, the Netherlands (see Vermunt, 1997).

is the standard error given by program MARK. For the owners, 20–29 years data, $m = 228$; for the renters, 30–44 years data, $m = 260$.

For owners 20–29 years (top half of Table 2), the estimates of N for ZW's estimator is quite discrepant from the MLE estimators; however, the large standard error for ZW's estimate (s.e. $= 64$) suggests that the discrepancies are due more to model instability rather than bias. Also note that changing from $\gamma = 0.05$ and $\gamma = 0.10$ for the L_{tAB} estimator has a small effect on the estimates of N suggesting that the L_{tAB} estimates of N are fairly robust to error in estimates of γ for these data.

The bottom half of Table 2 corresponds to renters, 30–44 years. For these data, the marginal AB odds ratio and the AB odds ratio conditional on $C = 1$ are both approximately 34.0, whereas the AB odds ratio conditional on $C = 0$ is approximately 27.4. Under the M_{tAB} model, the estimates for all odds ratios are 9.9. The large discrepancy in the odds ratio estimates is reflected in the difference between the estimates for n_{000} from the two models (the difference is 247). Again, this difference is small relative to the standard error of ZW's estimate (s.e. = 432).

As we did for owners, the L_{tAB} model was fit twice using 0.05 and 0.10 as plausible values for γ. The estimate for n_{000} from the L_{tAB} model decreased by 15% and 30%, respectively, as compared to the M_{tAB} model. This decrease is expected, as removing EEs from the data will lower the estimate of the population size. Note, however, that the change in $\hat{N}$ is relatively small.

In this example, the L_{tAB} and M_{tAB} models yielded substantially lower estimates for n_{000} than did ZW's model. These smaller estimates of n_{000} appear to be more plausible as they imply census enumeration rates which are more consistent with prior experience for these areas (see, e.g., Hogan, 1993). Our studies of artificial populations such as those described in the next section suggest that in populations where either ZW's or the M_{tAB} model assumptions maintain, estimates of n_{000} based on ZW and M_{tAB} are very close. The standard errors of the M_{tAB} estimate are much smaller in these populations; however, suggesting the M_{tAB} estimate is preferable to ZW's in populations where ZW's model is also appropriate.

4 Assessing Estimation Accuracy Using Artificial Data

One key objective of our research is to investigate the bias and variance of our triple system estimators of N. In particular, we are interested in examining the properties of the M_t, M_{tAB}, L_t, and L_{tAB} models' estimates of N with varying levels of EEs in the estimation process. In addition, we wish to investigate the consequences of misspecifying the estimation model when behavioral interactions between the indicators are present in the data.

Analytical methods for assessing the bias and variance of the estimates from capture-recapture models are quite complex and are often only available for method of moments estimators. Even in that case, the formulas for the mean square error components are often asymptotic expressions (see Seber, 1982). To circumvent these difficulties, current research has focused on numerical methods of parameter estimation. Since numerical estimation methods lack analytical equations for the parameters, estimates of bias are usually obtained by performing simulation experiments.

In one analysis, we generated data deterministically to simulate a situation where the entire population is sampled. Thus, the parameters specified for the population also hold true exactly in the analysis data set. Since the population parameter values are known and pre-specified exactly for the analysis data set, examination of bias and variance components without the effects of sampling variance is possible. The primary goal of this type of analysis is to study model bias when underlying model assumptions have been violated. We refer to this type of simulation as artificial population analysis without sampling.

Section 4.2 summarizes the results of the artificial population analysis without sampling. The four models of interest were compared using the deterministically generated artificial data. The formulas for generating these data are given in the next section. Variance estimates were not calculated for this analysis since there was no meaningful method of testing their validity.

In a second type of analysis, also described in Sect. 4.2, numerous samples were randomly selected from an artificial population. The models or formulas under study are then applied to each sample in order to estimate the population parameters of interest. Since the true parameter values are known, the bias of the parameter estimators can be accurately determined provided a sufficiently large number of samples of a given size are generated. In addition to the estimation of bias, the sampling distributions and the variance of the estimators can be determined so that the coverage properties of interval estimates can also be assessed. We refer to this type of simulation experiment as artificial population analysis with sampling.

The primary purpose of our simulation experiments is to determine the bias for point estimates and validity of variance estimates derived from the four selected models presented in Sect. 3. Additionally, once point and variance estimates have been obtained, the mean square error can be computed to determine which model for producing an estimate of N has smallest total error. An extensive simulation experiment for the four models listed above was conducted, and results are given in the next section.

4.1 Simulation Methodology

Generating the Artificial Data Without Sampling The data consist of the number of individuals in each of the seven observable cells in an ABC table, i.e., all cells except the 000 cell whose count was set to 0. Although the true number of residents in the 000 cell is known for the artificial populations, this information was suppressed in estimation process since it is unobserved in ABC table.

As stated, all values for the ABC table are generated using the deterministic equation for the number of observations in cell (a, b, c) given by

$$m_{abc} = M\pi_x\pi_{a|x}\pi_{b|ax}\pi_{c|x}. \tag{13}$$

In order to narrow the focus of the study, several of the population parameters were held constant over all of the artificial data sets. We chose a population size of $N = 8,000$ corresponding to roughly the size of a census tract in a central city area. We set the enumeration probability for the census at $\pi_{A=1|X=1} = 0.70$ corresponding to a difficult census enumeration area (see, e.g., the estimates for renters in Table 4 or Hogan et al., 2002). The probability of enumeration for the PES given enumeration by the census was set at $\pi_{B=1|A=1,X=1} = 0.90$ which corresponds to a moderately high behavioral dependence. Finally, the probability of being listed on the administrative list was set at $\pi_{C=1|X=1} = 0.50$ which corresponds to a list with fairly poor coverage properties. Some exploration of other values for these parameters has been undertaken within a 25 percentage-point range of these values, and, in general, the results are consistent with those reported here. We make no claim, however, that our results will hold beyond the range investigated.

The remaining two parameters were varied over a fairly wide range of plausible values as determined by previous census experience. The parameter $\pi_{B=1|A=0,X=1} = 0.90$, which specifies the level of behavioral correlation between A and B, was varied over the values 0.40 through 0.90 by increments of 0.10. The parameter $\gamma = \pi_{X=0|C=1}$, which determines the number of EEs in the C-list, was varied over the values 0.0, 0.02, 0.05, 0.10, and 0.15. All possible combinations of parameters are considered with each possible combination yielding one artificial data set.

Generating the Artificial Data with Sampling The data for the artificial data analysis with sampling were derived using the same set of parameters as described for the case without sampling. For each parameter combination, 1000 artificial data sets were generated. Each data set was randomly generated by the following five-step algorithm: (1) calculate the probabilities associated with the eight cells of the ABC table (denoted π_i, $i = 1, \ldots, 8$, say) using the true parameter values of the artificial population; (2) compute the quantities $s_0 = 0$, $s_k = \sum_{i=1}^{k} \pi_i$ for $k = 1, \ldots, 8$, (3) for residents, draw a uniform(0, 1) random number, r, and increment the count in cell k by 1 if $s_{k-1} \le r < s_k$, $k = 1, \ldots, 8$; (4) repeat step 3 for 8000 residents; and (5) add EEs to the 001 cell such that exactly 100γ percent of the enumerations on the C-list were erroneous.

Fitting the Models For each artificial data set, all four models were fit in order to obtain an estimate of N. The M_t, M_{tAB}, and L_t models can be fit using only the data from the ABC tables. As stated in Sect. 3, the L_{tAB} model requires an estimate for γ in order to obtain meaningful inference about N. This shifts the focus of inference for the L_{tAB} model from bias due to violation of model assumptions to bias in the estimate of N due to misspecification of γ.

The parameter estimates were obtained using the unconditional likelihood of the models of interest. The results were compared to estimates obtained using LEM which uses a conditional likelihood estimation approach as described in Bishop et al. (1975). The two estimation methods provided similar inference.

The results from the analysis of the artificial data sets are given in two parts. The first part described in Sect. 4.2 contains the results for the models that do not require knowledge of γ to produce meaningful inference, viz., M$_t$, M$_{tAB}$, and L$_t$. Section 4.2 is devoted solely to the study of robustness of the L$_{tAB}$ model estimates of N to failures of the model assumptions and misspecification of γ.

4.2 Results for the M$_t$, M$_{tAB}$, and L$_t$ Models

In the following tables, γ is the proportion of EEs in the C-list for the artificial population, δ is the error in the value of γ specified in the model, and $\rho_{AB|X=1}$ is the degree of AB interaction in the artificial population, which corresponds to the correlation between the A- and B-list given $X = 1$. In Tables 3, 4, 5, and 6, the models being compared are listed across the top row of the tables. The model estimate of N, the standard error of (SE column), the mean square error (MSE column), and the percent bias of that estimate (%Bias column) are shown for each model. All variances, biases, and MSEs were estimated directly from the simulation results. The tables only report the results from the simulations with sampling since the bias results for the simulations without sampling were essentially the same. As stated previously, for all cases the resident population size being estimated is 8000.

Table 3 explores the level of bias in the M$_t$, M$_{tAB}$, and L$_t$ models when there are EEs, but no AB interaction. As expected, the L$_t$ model is capable of producing an estimate of N that is virtually unbiased when EEs are present in the C-list. Both the M$_t$ and M$_{tAB}$ model yield biased estimates of N; however, the bias of the M$_{tAB}$ estimate is greater than the bias of the M$_t$ estimate. The point estimates for N from the with sampling and without sampling data are similar. The MSEs clearly show that the L$_t$ model performs better than the M$_t$ or M$_{tAB}$ when EEs are present in the data.

Table 4 shows the MSE components for the M$_t$, M$_{tAB}$, and L$_t$ models when there are no EEs in any list, but there is an AB interaction. The values of $\rho_{AB|X=1}$ correspond to the changing levels of $\pi_{B=1|A=0,X=1}$. For example, when $\pi_{B=1|A=0,X=1} = 0.80$, then $\rho_{AB|X=1} = 0.14$. The M$_{tAB}$ model accurately estimates N in the presence of an AB interaction. The other two models show significant bias due to the AB interaction. The L$_t$ model shows considerably more bias and has a larger MSE than the M$_t$ model. It should be noted that the M$_t$ model tends to have the smallest standard error of the three models and the smallest MSE when there is no interaction or EEs present in the data and it behaves poorly when either of these assumptions are violated.

Table 5 reports the MSE components for the M$_t$, M$_{tAB}$, and L$_t$ models when there are EEs and an AB interaction. The AB interaction is set at the highest level explored in this study, $\rho_{AB|X=1} = 0.53$, which corresponds to $\pi_{B=1|X=1,A=0} = 0.40$. All three of the models are substantially biased and produce a large MSE when

Table 3 No AB interaction and EEs given by γ with sampling

	M_t				M_{tAB}				L_t			
$\gamma(\%)$	$\hat{N}$	SE	MSE	Bias(%)	$\hat{N}$	SE	MSE	Bias(%)	$\hat{N}$	SE	MSE	Bias(%)
0	7999	12.28	151	0.0	7999	15.98	257	0.0	7993	15.21	280	–0.1
2	8098	12.41	9773	1.1	8162	16.20	26432	2.0	7999	20.54	425	0.0
5	8257	12.27	66112	3.2	8420	15.99	176588	5.3	8000	18.95	402	0.0
10	8546	14.50	298697	6.8	8887	17.78	786553	11.1	8000	20.20	402	0.0
15	8875	16.31	766748	10.9	9409	19.00	2922785	17.6	8000	19.95	398	0.0

Table 4 AB interaction given by $\rho_{AB|X=1}$ and no EEs with sampling

	M$_t$				M$_{tAB}$				L$_t$			
$\rho_{AB\|X=1}$	$\hat{N}$	SE	MSE	Bias(%)	$\hat{N}$	SE	MSE	Bias(%)	$\hat{N}$	SE	MSE	Bias(%)
0.0	7999	12.29	151	0.0	7999	15.99	257	0.0	7993	15.22	280	−0.1
0.14	7893	16.22	11678	−1.3	7999	22.04	486	0.0	7731	24.10	72652	−3.4
0.25	7785	20.45	46519	−2.7	8000	29.50	870	0.0	7466	28.58	286122	−6.7
0.35	7674	21.81	106732	−4.1	8000	31.94	1020	0.0	7199	30.38	643181	−10.0
0.44	7561	24.54	192972	−5.5	8000	37.34	1394	0.0	6933	33.45	1139608	−13.3
0.53	7444	27.47	309890	−6.9	7998	43.90	1931	0.0	6665	34.42	1785386	−16.7

Table 5 AB interaction with $\rho_{AB|X=1} = 0.53$ and EEs given by γ with sampling

	M$_t$				M$_{tAB}$				L$_t$			
γ(%)	$\hat{N}$	SE	MSE	Bias (%)	$\hat{N}$	SE	MSE	Bias (%)	$\hat{N}$	SE	MSE	Bias (%)
0	7444	27.74	309890	−7.0	7998	43.90	1931	0.0	6664	34.42	1783596	0.0
2	7542	26.84	209834	−5.7	8160	41.98	27513	2.0	6666	34.43	1776659	0.0
5	7700	27.93	90684	−3.8	8420	43.49	178879	5.3	6666	35.76	1790793	0.0
10	7987	29.41	1016	−0.2	8886	45.50	787811	11.1	6666	35.78	1790783	0.0
15	8312	30.53	97989	3.9	9408	49.60	1986191	17.6	6665	34.40	1782739	0.0

an AB interaction and EEs are present in the data. It appears the L_t exhibits the largest MSE of the three models, suggesting that, despite the fact the L_t model can account for EEs, this advantage is negated in the presence of behavioral correlation. These tables highlight the need for a model (e.g., the L_{tAB} model) that is capable of fitting this.

Results for the $\mathbf{L}_{tAB}$ Model Tables 6 and 7 show the key results for the L_{tAB} model. As mentioned previously, identifiability of L_{tAB} can be achieved if information on the number of EEs in the C-list is entered into the model. Therefore, we fit the L_{tAB} model using a known value for y and consider situations where γ is not known exactly. For example, γ may be estimated from a study where a random sample of the persons on the C-list is selected and sent to the field in order to verify their residential statuses. In that case, our estimate of γ would be subject to non-sampling and sampling errors and would not be known exactly (see the next section). In Tables 6 and 7 we consider the effect on the model estimate of N when y is subject to error equal to δ.

In Tables 6 and 7, the value of γ is listed in the first row of the table, and the amount of error in y, denoted by δ, is given in the first column of the table. For example, if $\gamma = 0.10$ and $\delta = 0.20$, then the value of γ used to fit the model is $\gamma = 0.08$. For a given error percentage, the estimate for N along with the percent bias is given in the two columns below the error percentage. The tables consider both positive and negative values of δ.

Table 6 explores the level of bias in the L_{tAB} model when there are EEs but no AB interaction for different values of γ. Table 7 explores the level of bias in the $LtAB$ model when there are EEs and an AB interaction for different values of γ. For this table, $\rho_{AB|X=1} = 0.53$ in all cases.

Both tables illustrate that the L_{tAB} model produces a virtually unbiased estimate of N when y is correctly specified. In addition, the estimate of N appears to be robust to mis-specification of y. For example, even with as much as 50% error, the estimates of N are still within 10% of the true value.

There are differences in the value of between Tables 6 and 7. These differences occur primarily when y is large (10%, 15%) and the amount of error in y is positive and large (i.e., δ in the range of 0.30 to 0.50). This is likely due to the fact that is equal to 0.40 for Table 6 and 0.90 for Table 7. When $\pi_{B=1|A=0,X=1} = 0.40$, fewer individuals tend to be included in the observable cells of the ABC table. Thus, the estimate of N can take on lower values. This is true since the lower bound of the estimate of N is equal to the number of individuals enumerated minus the number of recognized EEs in the data.

5 Summary and Discussion

All four models we considered (M_t, M_{tAB}, L_t, and L_{tAB}) produce virtually unbiased estimates of N when a given model's assumptions are valid. For example, when

Table 6 L_{tAB} model: no AB interaction and EEs given by γ with sampling

	$\gamma = 0\%$				$\gamma = 2\%$				$\gamma = 5\%$				$\gamma = 10\%$				$\gamma = 15\%$			
δ	$\hat{N}$	SE	MSE	Bias (%)	$\hat{N}$	SE	MSE	Bias (%)	$\hat{N}$	SE	MSE	Bias (%)	$\hat{N}$	SE	MSE	Bias (%)	$\hat{N}$	SE	MSE	Bias (%)
−50	n/a	n/a	n/a	n/a	8081	15.9	6795	1.0	8207	16.2	43092	2.6	8441	16.8	194764	5.5	8703	17.5	494514	8.8
−40	n/a	n/a	n/a	n/a	8063	17.0	4203	0.8	8167	15.9	28132	2.1	8354	16.7	125596	4.4	8561	17.2	315017	7.0
−30	n/a	n/a	n/a	n/a	8045	15.6	2259	0.6	8124	15.2	15619	1.5	8263	16.4	69437	3.3	8421	16.5	177512	5.3
−20	n/a	n/a	n/a	n/a	8030	16.1	1160	0.4	8080	16.6	6643	1.0	8173	16.5	30202	2.2	8278	16.4	77554	3.5
−10	n/a	n/a	n/a	n/a	8013	15.6	412	0.2	8037	17.2	1612	0.5	8085	16.7	7503	1.1	8136	15.7	18742	1.7
0	7999	16.3	267	0.0	7997	15.4	252	0.0	7996	16.1	273	0.0	7996	15.5	257	0.0	7994	15.4	272	0.1
10	7960	15.6	1841	−0.5	7981	15.8	604	−0.2	7955	15.7	2268	−0.6	7902	16.4	9873	−1.2	7807	43.5	39138	2.4
20	7919	15.3	6475	−1.0	7965	15.5	1468	−0.4	7909	15.9	8533	−1.1	7773	24.9	52147	−2.8	7762	14.8	56862	3.0
30	7878	16.4	15118	−1.5	7949	15.8	2844	−0.6	7866	17.3	18254	−1.7	7761	14.6	57335	−3.0	7760	14.9	57823	3.0
40	7837	15.6	26803	−2.0	7931	16.0	5004	−0.9	7810	29.3	36958	−2.4	7761	15.3	57355	−3.0	7762	15.0	56869	3.0
50	7789	19.6	44755	−2.6	7916	15.6	7299	−1.1	7765	16.4	55494	−2.9	7762	15.1	56871	−3.0	7761	15.2	57353	3.0

Table 7 L$_{tAB}$ model: AB interaction with $\rho_{AB|X=1} = 0.53$ and EEs given by γ with sampling

	$\gamma = 0\%$				$\gamma = 2\%$				$\gamma = 5\%$				$\gamma = 10\%$				$\gamma = 15\%$			
δ	$\hat{N}$	SE	MSE	Bias (%)	$\hat{N}$	SE	MSE	Bias (%)	$\hat{N}$	SE	MSE	Bias (%)	$\hat{N}$	SE	MSE	Bias (%)	$\hat{N}$	SE	MSE	Bias (%)
−50	n/a	n/a	n/a	n/a	8078	42.3	7875	1.0	8210	42.6	45913	2.6	8440	45.6	195767	5.5	8705	47.9	499321	8.8
−40	n/a	n/a	n/a	n/a	8064	43.2	5966	0.8	8165	43.8	29145	2.1	8356	43.7	128642	4.4	8563	45.7	319058	7.0
−30	n/a	n/a	n/a	n/a	8047	40.1	3818	0.6	8125	39.6	17196	1.5	8265	41.7	71967	3.3	8424	45.1	181811	5.3
−20	n/a	n/a	n/a	n/a	8031	41.9	2720	0.4	8085	42.2	9002	1.0	8173	41.2	31630	2.2	8281	43.0	80810	3.5
−10	n/a	n/a	n/a	n/a	8015	45.4	2283	0.2	8039	42.0	3288	0.5	8089	41.4	9631	1.1	8140	42.7	21423	1.8
0	7999	41.5	1727	0.0	7997	43.2	1876	0.0	7997	43.5	1903	0.0	7996	40.4	1651	0.1	7998	43.0	1850	0.0
10	7958	42.7	3584	−0.5	7980	41.7	2136	−0.2	7956	41.8	3687	−0.6	7912	40.9	9412	−1.1	7857	41.7	22188	−1.8
20	7917	41.2	8586	−1.0	7965	40.3	2852	−0.4	7916	39.6	8624	−1.1	7822	41.1	33376	−2.2	7716	40.5	82294	−3.6
30	7877	40.4	16760	−1.5	7950	38.7	3999	−0.6	7872	39.4	17936	−1.7	7730	39.2	74437	−3.4	7573	42.3	184118	−5.3
40	7841	40.2	26895	−2.0	7933	41.3	6193	−0.9	7827	42.4	31723	−2.4	7639	38.1	131775	−4.5	7430	39.9	326490	−7.1
50	7799	38.5	41880	−2.5	7916	40.3	8681	−1.1	7785	42.7	48051	−2.9	7555	39.7	199598	−5.6	7292	39.1	502789	−8.9

$N = 8000$, $\rho_{AB|X=1} = 0$, and $\gamma = 0$, the M$_t$ model produces an estimate for N of 7999. As the assumptions are violated, all models begin to show biased results. The results of each model will be summarized separately.

The M$_t$ model is the least complex of the four models that we studied in detail. This model's inability to account for either EEs or a behavioral correlation was evident from Tables 4 and 5. EEs induce a positive bias in the estimate of N, while an AB interaction induces a negative bias. When both an AB interaction and EEs are present in the data, the biases due to these conditions tend to offset each other. The result is that the M$_t$ model shows less bias in the estimate of N as compared to the M$_{tAB}$ and L$_t$ models when the effect of EEs in the data is approximately equivalent to the behavior correlation effect. Of course, this is in no way a desirable property of the model since balancing these two errors is not under the control of the experimenter. When correlation bias and EEs are not off-setting, the bias in the M$_t$ estimator can be substantial.

The L$_t$ model is designed to estimate N when there are EEs present in the C-list. As seen from Table 3, the L$_t$ model produces estimates of N that are virtually unbiased when EEs are present in the data and there is no correlation bias. As demonstrated by Table 4, an AB interaction induces a severe negative bias in the estimate of N. This is similar to the negative bias associated with the correlation induced by population heterogeneity discussed in other works (see, e.g., Alho et al., 1993). In addition, the AB interaction induces a much larger bias and MSE for the L$_t$ model than for the M$_t$ model. For the values of $\rho_{AB|X=1} > 0$ presented in Table 4, the MSE for the L$_t$ model is approximately six times larger than that for the M$_t$ model. As illustrated in Table 5, when both EEs and an AB interaction are present in the data, the M$_t$ model will likely have a lower MSE than the L$_t$ model. The exceptions occur when the AB interaction is small and is large. In general, it appears as if the L$_t$ model is not very robust to the presence of an AB interaction.

It is interesting to compare the estimates from the M$_t$ and L$_t$ models. As stated above, the L$_t$ model's estimates of N tend to have more bias and a larger MSE than the M$_t$ estimates when an AB interaction is present. By comparison, the estimate of N from the M$_t$ model appears to be relatively robust when the proportion of EEs on the C-list is small. Therefore, if information on y is not available and the choice is between M$_t$ and L$_t$, we recommend using the M$_t$ model over of the L$_t$ model, particularly if a sizeable AB interaction is expected. The L$_t$ model is preferred when there is a large proportion of EEs in the C-list and γ is unknown. If γ is known, it might be possible to improve the inference obtained by the L$_t$ model by incorporating an estimate of y into the likelihood.

The M$_{tAB}$ model is designed to estimate N when an AB interaction, but no EEs, are present in the data. As shown from Table 4, the M$_{tAB}$ model produces virtually unbiased estimates for N for a range of values for $\rho_{AB|X=1}$. As compared to the M$_t$ model, the presence of EEs induces a large positive bias in the M$_{tAB}$ model. As seen from Table 3, when EEs are present in data, the MSE for the M$_{tAB}$ model is approximately 2.5 times larger than that of the M$_t$ model.

The reason for this increase can be explained by the additional parameter in the M$_{tAB}$ model. The M$_{tAB}$ model has two parameters for enumeration probabilities for the B-list, $\pi_{B=1|A=1,X=1}$ and $\pi_{B=1|A=0,X=1}$. For the M$_t$ model, these two probabilities are equal, $\pi_{B=1|X=1} = \pi_{B=1|A=1,X=1} = \pi_{B=1|A=0,X=1}$, and both the unenumerated and the previously enumerated individuals in the B-list are used to estimate $\pi_{B=1|X=1}$. Thus, more information is used to estimate $\pi_{B=1|X=1}$ and hence N, in the M$_t$ model which, consequently, improves its robustness to EEs.

One nice property of the M$_{tAB}$ model is that the degree of the AB interaction does not effect the bias in N due to EEs. This concept can be seen by comparing the results for the M$_{tAB}$ model given in Tables 3 and 4. Additionally, from Table 5, it appears that the M$_{tAB}$ model outperforms the M$_t$ model when there few to moderate amount of EEs in the data, $\gamma < 0.05$. Thus, it appears to be preferable to use the M$_{tAB}$ model over the M$_t$ model when a moderate number of EEs are expected in the data. Unfortunately, as illustrated in Table 5, the M$_t$, L$_t$, and M$_{tAB}$ models exhibit large biases and MSEs when both EEs and an AB interaction are present in the data. The results from Table 5 highlight the need for the L$_{tAB}$ model.

Of the four models given notable consideration in this study, the L$_{tAB}$ model is the most likely to accurately represent the triple system data. This model can account for both EEs in the C-list and for an AB interaction. Unfortunately, given only the ABC table, the L$_{tAB}$ model is unidentifiable and requires the inclusion of additional information to provide meaningful inferences for N. Our solution to the lack of identifiability is to provide a value for the proportion of EEs in the C-list, γ. By specifying γ, the L$_{tAB}$ model becomes fully identifiable and produces virtually unbiased estimates for N as seen in Table 6. Moreover, as seen in Table 7, the inclusion of an AB interaction does not affect the inference that is obtained from the L$_{tAB}$ model when γ is known. Thus, our results indicate that the L$_{tAB}$ model can accurately represent data with an AB interaction without affecting the nature of the inference.

Another concern for this model is the robustness of the estimate of N from the L$_{tAB}$ model to the misspecification of γ. Both Tables 6 and 7 explore the levels of bias induced in the estimate of N when γ is misspecified. In general, the bias tends to be low, implying that the estimates of N are fairly robust. For example, consider the case when $\gamma = 0.10$ and an $\rho_{AB|X=1} = 0.53$. For the different values of δ presented in Table 7, the MSE for the L$_{tAB}$ model ranges from 1,651 when $\delta = 0$ to 195,767 when $\delta = 0.50$. Similarly, the bias ranges from 0.0% when $\delta = 0$ to 5.5% when $\delta = 0.50$. By comparison, under this scenario, the M$_t$, L$_t$, and M$_{tAB}$ models have MSEs of 1,016, 1,986,191, and 1,782,739, respectively. Even when γ is badly misspecified, the L$_{tAB}$ model appears to outperform the M$_{tAB}$ and L$_t$ model.

In order to fully utilize the L$_{tAB}$ model, a reasonable value for γ must be obtained from a separate data source. One possible method for obtaining an estimate of γ is to conduct a field study by drawing a random sample from the observations in cell 001 of the ABC table. In this situation, the MSE formulas in the present paper can be expanded to include variation in the estimate of N due to estimating γ. Our preliminary investigations of this method suggest that even in situations where there is considerable sampling variability in the estimate of γ, the L$_{tAB}$ model MSE of the L$_{tAB}$ model estimate is still considerably smaller than that of the M$_{tAB}$ model when γ is in the range of 0.05–0.15.

In general, when undetected EEs appear in the C-list and a reasonable estimate of γ is available, the L_{tAB} model performed best for estimating N. If an estimate of cannot be obtained, then the selection of an appropriate model to use for inference about N is less clear. It appears, however, that the M_{tAB} and M_t models outperform the L_t model. The selection of which model to use would depend on the nature of data, specifically on the strength of the AB interaction and the number of EEs in the C-list.

An alternative to using the L_{tAB} model for dealing with EEs in the ARL is to proceed with the M_{tAB} model and use an estimate of γ in post hoc correction of $\hat{N}_{\mathrm{M}_{tAB}}$ for EEs. Our empirical studies suggest that such corrections can produce unbiased results in populations that are also ideal for the L_{tAB} model. One such post hoc estimator (derived in Appendix 2) that appears to produce very good results is

$$\tilde{N}_{\mathrm{M}_{tAB}}(\gamma) = (1-\gamma)\hat{N}_{\mathrm{M}_{tAB}}. \tag{14}$$

In practice, if an unbiased estimate, $\hat{\gamma}$ of γ is available, using (14) after substituting $\hat{\gamma}$ for γ will generally reduce the bias in $\hat{\tilde{N}}_{\mathrm{M}_{tAB}}$; however, it is possible that the MSE of $\hat{\tilde{N}}_{\mathrm{M}_{tAB}}$ could increase depending upon the size of the bias reduction relative to the variance of $\hat{\gamma}$.

An important advantage of using L-models to explicitly account for EEs rather than using post hoc corrections of M-models is the ease with which L-models can be extended to more complex situations. When EE's appear in more than one list, post hoc corrections for EEs are impractical due to their complexity. Latent class analysis provides an integrated structure for modeling much more complicate scenarios than were described in this paper. Thus, L_{tAB} model should be viewed as a foundation for more complex models that involve list-by-list interactions, EEs in all three lists, and four or more lists. The current paper lays the groundwork for dealing with these more complex situations.

Appendix 1: Derivation of M_{tAB} and L_{tAB} Estimators of n_{000}

The likelihood for the M_{tAB} model, denoted by $\ell_{\mathrm{M}_{tAB}} = \ell(N, \pi_a, \pi_{b|a=1}, \pi_{b|a=2}, \pi_c | n_{i,j,k})$ can be written as

$$\begin{aligned}\ell_{\mathrm{M}_{tAB}} = {} & \frac{N!}{\prod\limits_{i,j,k} n_{ijk}!(N-n_{+++})!}\pi_a^{n_{1++}}(1-\pi_a)^{N-n_{1++}}\pi_{b|a=1}^{n_{11+}}(1-\pi_{b|a=1})^{n_{1++}-n_{11+}} \\ & \times\pi_{b|a=2}^{n_{01+}}(1-\pi_{b|a=2})^{N-n_{1++}-n_{01+}}\pi_c^{n_{++1}}(1-\pi_c)^{N-n_{++1}},\end{aligned} \tag{A.1}$$

where n_{ijk} denotes the cell count in cell (i, j, k) of the ABC table, "+" indicates summation over the corresponding index, and the other notation is as defined in Sect. 2. To find the value of the parameters that maximizes (A.1), we take the logarithm of this function and differentiating with respect to the parameters, set these partial derivatives equal to 0, and solve for the parameters. Holding N constant and maximizing with respect to the other parameters produces the following MLEs for π_a, $\pi_{b|a=2}$, and π_c and conditional on N:

$$\hat{\pi}_{a|N} = \frac{n_{1++}}{N}, \quad \hat{\pi}_{a|b=2,N} = \frac{n_{01+}}{N - n_{1++}}, \quad \hat{\pi}_{c|N} = \frac{n_{++1}}{N}. \tag{A.2}$$

Replacing these parameters in (A.1) by their conditional MLEs, the likelihood can be written as a function of N only. Simplifying and removing factors that do not contain N, (A.1) simplifies to

$$\frac{N!}{(N - n_{+++})!} N^{-2N} (N - n_{1++} - n_{01+})^{N-n_{1++}-n_{01+}} (N - n_{++1})^{N-n_{++1}} \tag{A.3}$$

Several approximations will be used in determining an MLE for N. First, N will be treated as a continuous variable, and second, in order to take a derivative of $N!$, Stirling's approximation to the factorial will be used. This yields the following approximation to (A.2):

$$\frac{N^{N+0.5} e^{-n}}{(N - n_{+++})^{N-n_{+++}+0.5} e^{-N+n_{+++}}} N^{-2N}$$
$$\times (N - n_{1++} - n_{01+})^{N-n_{1++}-n_{01+}} (N - n_{++1})^{N-n_{++1}}.$$

Again, eliminating factors that do not involve N yields

$$(N - n_{+++})^{-(N-n_{+++}+0.5)} N^{-N+0.5} (N - n_1 - n_{01+})^{N-n_1-n_{01+}} (N - n_{++1})^{N-n_{++1}}.$$

Taking the natural log of the above expression gives

$$-(N - n_{+++} + 0.5)\log(N - n_{+++}) - (N + 0.5)\log(N)$$
$$+(N - n_{++1} - n_{01+})\log(N - n_{1++} - n_{01+}) + (N - n_{++1})\log(N - n_{++1}). \tag{A.4}$$

Now we can take the derivative of (A.4) with respect to N and set the resulting expression to 0. To obtain the following expression, we use a third approximation, viz., $\log(1 + \alpha) = \alpha$ where α is a small positive constant. Upon simplifying, this yields

$$\log\left[\frac{(N - n_{1++} - n_{01+})(N - n_{++1})}{(N - n_{+++} + 0.5)(N + 0.5)}\right] = 0.$$

Finally, solving for N and further simplifying yields

$$\hat{N}_{\mathrm{M}_{tAB}} = \frac{n_{++1}(n_{1++} - n_{01+}) + 0.5n_{++0} - 0.25}{n_{1++} + n_{01+} + n_{++1} - n_{+++} + 1}. \tag{A.5}$$

Subtracting n_{+++} from the above expression produces estimate of n_{000} that can be compared with the estimator based on (11).

MLEs for the L_{tAB} model can be derived in a similar fashion. Since for the L_{tAB} model, EEs only appear on the C-list, specifying γ is equivalent to specifying the number of EEs, say n_{EE} that occur on the C-list since $\gamma = n_{\mathrm{EE}}/n_{++1}$. Repeating the above steps and approximations for this model yields the following approximate MLE for the L_{tAB} model when γ is known:

$$\hat{N}_{\mathrm{L}_{tAB}} = \frac{(n_{++1} - n_{\mathrm{EE}})(n_{1++} - n_{01+}) + 0.5n_{++0} - 0.25}{n_{1++} + n_{01+} + n_{++1} - n_{+++} + 1}. \tag{A.6}$$

Appendix 2: Derivation of the Estimator $\tilde{\hat{N}}_{\mathrm{M}_{tAB}}(\gamma)$

Using the results of Appendix A, the ratio of $\hat{N}_{\mathrm{M}_{tAB}}$ to $\hat{N}_{\mathrm{L}_{tAB}}$ can be written as

$$\frac{\hat{N}_{\mathrm{L}_{tAB}}}{\hat{N}_{\mathrm{M}_{tAB}}} = \frac{(n_{++1} - n_{\mathrm{EE}})}{n_{++1}} \times \frac{(n_{1++} + n_{01+} + 0.5) + \frac{0.5n_{++0} - 0.25}{(n_{++1} - n_{\mathrm{EE}})}}{(n_{1++} + n_{01+} + 0.5) + \frac{0.5n_{++0} - 0.25}{n_{++1}}}. \tag{A.7}$$

The remainder of this proof will show that the second factor on the right hand side of (A.7), denoted by F, can be approximated by 1 for values of $\gamma < 0.5$. In that case, $\hat{N}_{\mathrm{L}_{tAB}} \approx (1-\gamma)\hat{N}_{\mathrm{M}_{tAB}}$.

To show that $F \approx 1$, for small γ, we multiply and divide F by $(n_{1++} + n_{01+} + 0.5)$. Ignoring the term -0.25 which is negligible compared with $0.5n_{++1}$, we obtain

$$F = \frac{1 + \frac{0.5n_{++0}}{(n_{++1} + n_{01+} + 0.5)(n_{++1} - n_{EE})}}{1 + \frac{0.5n_{++0}}{(n_{1++} + n_{01+} + 0.5)n_{++1}}}. \tag{A.8}$$

Note that $(n_{1++} + n_{01+} + 0.5) > n_{++0}$, which implies that $\frac{0.5n_{++0}}{(n_1 + n_{01+} + 0.5)} = c_1$, for some constant $c_1 < 0.5$. Thus, $F = \frac{1 + \frac{c_1}{(n_{++1} - n_{EE})}}{1 + \frac{c_1}{n_{1++}}}$. Replacing $n_{++1} - n_{EE}$ with $n_{++1}(1-\gamma)$ and simplifying yields

$$F = \frac{\frac{n_{++1}(1-\gamma) + c_1}{(1-\gamma)}}{n_{++1} + c_1} = \frac{n_{++1}}{n_{++1} + c_1} + \frac{c_1}{(n_{++1} + c_1)(1-\gamma)}.$$

When $\gamma = 0$ then this expression is exactly 1. When $\gamma = 0.5$, F reduces to $\frac{n_{++1}+2c_1}{n_{++1}+c_1}$. Since $c_1 < 0.5$, $F \approx 1$, when n_{++1} is reasonably large or, in general, for any value of γ between 0 and 0.5.

References

Alho, J., Mulry, M., Wurdeman, K, & Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association, 88*, 1130–1136.

Biemer, P. P. (1988). Modeling matching error and its effect on estimates of census coverage error. *Survey Methodology, 14*(1), 117–134.

Biemer, P., & Davis, M. (1991a). Estimates of P-sample clerical matching error from a rematching evaluation, evaluation project P7 internal bureau of the census report.

Biemer, P., & M. Davis. (1991b). Measurement of census erroneous enumerations—clerical error made in the assignment of enumeration status. Evaluation Project P10, Internal Bureau of the Census Report.

Bishop, Y. M. M., Fienberg, S. E., & Holland P. W. (1975). *Discrete multivariate analysis: Theory and practice*. MIT Press.

Chao, A., & Tsay, P.K. (1998). A sample coverage approach to multiple-system estimation with application to the census undercount. *Journal of the American Statistical Association, 93*, 283–293.

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics, 48*, 201–216.

Darroch, J.N., Fienberg, S.E., Glonek, G. F. V., & Junker, B. W. (1993). A three-sample multiple recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association, 88*, 1137–1148.

Ding, Y., & Fienberg, S. (1992). Estimating population and census undercount in the presence of matching error. Unpublished manuscript.

ESCAP. (2001). ESCAP II report no. 1 recommendation and report of the executive steering committee for A.C.E. policy (ESCAP II). Bureau of the Census, Washington, D.C.

Fienberg, S. E., Johnson, M. S., & Junker, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, A, 162*, 383–405.

Haberman, S. (1979). Analysis of qualitative data (Vol. 2). New developments. Academic Press.

Hogan, H. (1993). The 1990 post-enumeration survey: Operations and results. *Journal of the American Statistical Association, 88*(423), 1047–1071.

Hogan, H., Kostanich, D., Whitford, D., & Singh, R. (2002). Research findings of the accuracy and coverage evaluation and census 2000 accuracy. In *American Statistical Association joint statistical meetings, proceedings of the section on survey research methods*.

Judson, D. (2000). The statistical administrative records system: system design, successes, and challenges. Internal Census Bureau Report, Nov. 11.

Pollock, K. H., Nichols, J. D., Brownie, C., & Hines, J. E. (1990). Statistical inference for capture-recapture experiments. Wildlife Monographs 107.

Schnabel, Z. E. (1938). The estimation of the total fish population of a lake. *The American Mathematical Monthly, 45*, 348–352.

Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters* (2nd ed.). Griffin.

Sekar, C. C., & Deming, W. E. (1949). On a method of estimating birth and death rates and extent of registration. *Journal of the American Statistical Association, 44*, 101–115.

Vermunt, J. (1997). *LEM: A general program for the analysis of categorical data*. Tilburg University.

White, G. C., & Burnham, K. P. (1999). Program MARK: Survival estimation from populations of marked animals. *Bird Study, 46*(Supplement), 120–138.

Wolter, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association, 81*, 157–162.

Zaslavsky, A., & Wolfgang, G. (1993). Triple-system modeling of census, post-enumeration survey, and administrative-list data. *Journal of Business and Economic Statistics, 11*, 279–288.

Record Linkage in Statistical Sampling: Past, Present, and Future

Benjamin Williams

Abstract Record linkage is a useful tool to match records across datasets when the datasets lack a unique identifier. In this chapter, we examine the past, current, and present uses of probabilistic record linkage with a specific interest in its use in statistical sampling. For example, given the rise in interest and use of non-probability data within sampling, many researchers seek to augment a non-probability sample with a probability sample. Record linkage is a useful method for doing such combining. This chapter will examine the ways record linkage has been used and is currently being researched and implemented, with an emphasis on its current and future use for statistical sampling. The chapter concludes with open research questions for record linkage in the context of sampling, where the questions center around the idea of creating a total error framework for linked data.

1 Introduction

Analysts broadly use the term *record linkage* to define the matching of records existing in two or more datasets. Record linkage is also used for data deduplication, but that is not the focus of this chapter. Here, record linkage encompasses other commonly used terms for data matching, including but not limited to entity resolution, data blending, data combination, document linkage, and record matching. Originally describing the process of combining specific life event records (e.g., birth, graduation, marriage) in a person's "Book of Life" (Dunn, 1946), record linkage has grown in breadth over the past 75 years and is an active area of statistical research. From its humble roots, record linkage has been mathematically formalized, implemented with machine learning, and employed at numerous public and private agencies (Herzog et al., 2007; Christen, 2019; Dong & Srivastava, 2015).

B. Williams (✉)
University of Denver, Denver, CO, USA
e-mail: benjamin.williams@du.edu

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_9

Record linkage is of use when two or more data files refer to the same entity yet lack a unique identifier common among all sources. In this chapter, without loss of generality, assume there are two files to link; call these files A and B. Record linkage relies on comparing *linking variables*, variables present in both A and B which should be equivalent for matching records. Newcombe et al. (1959) note two issues arising from comparing linking variables: (1) two records that *do not* refer to the same entity may have equivalent linking variable values (e.g., Ben Williams and Ben Leonard have equivalent first names, but may be different people), and (2) two records that *do* refer to the same entity may have different linking variable values (e.g., Benjamin Williams and Ben Williams could be the same person, but have different recorded first names). Record linkage can mitigate these issues.

Record linkage has two primary forms: deterministic and probabilistic (Herzog et al., 2007). A deterministic program links records across datasets via strict, predetermined rules concerning linking variables. An example is as follows: only link two entities if the recorded last names are equivalent and the recorded dates are within 2 days of each other. Deterministic record linkage can work well if there are few or no errors in the datasets. Probabilistic record linkage relies on the distribution of the linking variables to determine the likelihood two records match. Probabilistic record linkage is a powerful tool when there are possible errors in the datasets. Errors such as misspellings or incorrect recording of dates are quite common, making probabilistic record linkage popular. For the rest of this chapter, *record linkage* will refer to probabilistic record linkage.

In 1959, Newcombe et al. developed a linking score aggregating estimates of the log-odds that the values of the linking variables agree for each potential link between A and B (Newcombe et al., 1959). Their work was formalized in Fellegi and Sunter (1969). The Fellegi-Sunter implementation is the classic method of record linkage. They derived the linkage score for a pair of potential links by using the probabilities of observing agreement patterns in true matching and non-matching pairs of records. The expectation-maximization (EM) algorithm (Dempster et al., 1977) is often used to estimate the parameters for the score.

Potential links with a score above an upper threshold are called matches, potential links with a score below a lower threshold are called non-matches, and potential links with a score between the upper and lower thresholds are called potential matches. The thresholds, along with prespecified false-positive and false-negative rates, comprise a linking rule. Fellegi and Sunter proved this rule is optimal in the sense that it minimizes the probability a possible link is classified as a potential match as opposed to a match or a non-match. The rigorous method of combining datasets introduced by Fellegi and Sunter opened a new research context for record linkage: statistical sampling.

When a representative sample is drawn at random from a population, inference regarding the population can be made from inspection of the sample (Lohr, 2010). This is a foundational tenet of statistics. However, given the pervasive availability of big data, are large samples drawn not at random (non-probability samples) more useful than small probability samples? See Meng (2018) for a further discussion of this question. Indeed, large non-probability samples are easier than ever to collect, but often at the cost of representativeness and theoretical formulae for sampling variability (Baker et al., 2013). Wiśniowski et al. (2020) examine the trade-offs between non-probability samples and probability samples. They argue combining a small probability sample with a larger non-probability sample allows one to harness the advantages of both. In this, record linkage becomes immensely valuable.

Integrating two samples may require records to be matched between them. If the probability sample adds auxiliary information, records from one sample likely need to be matched to records on the other. One example of this is a capture-recapture framework used to combine the non-probability and probability samples. If the initial capture sample is a non-probability sample and the recapture sample is a probability sample, the records from each sample must be matched for valid estimation (Liu et al., 2017; Stokes et al., 2021). In such cases one may use record linkage for matching. Another example of this is at the US Census Bureau, where smaller secondary samples are gathered after the census which are linked to the original data for additional inference.

In the US Census example, one of the datasets for linking is quite large, the US Census. Since the census is much larger than the second sample, and is nearly a complete register of the population, linking is easier as there is a high probability that respondents to the second sample exist in the census data. If one or both of the data files to be linked are small, relative to the population size, then the likelihood of finding units existing in both samples could be quite small rendering record linkage impractical and not useful.

However, given the pervasive nature in the world today, big data and datasets nearing the size of populations of interest are becoming more common. In cases where one or more of the datasets are relatively large, record linkage is most useful since the probability of a sizeable overlap is higher. The overlapping units are often where the benefit of combining samples comes from. For a treatment of identifying the overlap between a big data source and a smaller probability sample, see Kim and Tam (2021). Record linkage is an important tool to augmenting samples, be they non-probability or probability. This is a critical area of future research in statistical sampling.

This chapter examines the past and current uses of record linkage, along with opportunities for the method in the future. We pay particular attention to the use of record linkage in statistical sampling, especially in the sections on current and future uses. In the coming years, record linkage will play a key role in the analysis of non-probability samples, and open research questions exist which deserve careful consideration. This chapter will thus conclude by laying out these questions, discussing their critical nature, and offering paths toward solutions.

2 Past Uses of Record Linkage

Historically, record linkage has been primarily used to link records of people, businesses, or addresses (Fellegi, 1999). Often the linking variables are comprised

Data File A					Data File B				
Name	City	Birth Year	Marital Status	...	Name	City	Birth Year	Number of Children	...
Ben Williams	Denver	1991	Y		Ben Williams	Dallas	1989	1	
Brian William	Dallas	1990	N		Ben William	Denver	1991	0	
...					...				

Fig. 1 Example of two files to link some variable names which are the same across the files

of words (or strings). An example of two files to link is in Fig. 1. File A and B share the variables *Name*, *City*, and *Birth Year* and those are the linking variables. Suppose it is of interest to combine the files to determine the relationship between *Marital Status* (only in File A) and *Number of Children* (only in File B).

In Fig. 1, a human analyst could reasonably determine the first entry in File A (linking variable values: Ben Williams, Denver, 1991) matches the second entry in File B (linking variable values: Ben William, Denver, 1991) by observing the misspelling of Williams in the File B entry. In this toy example, the values of the *Birth Year* and *City* linking variables are exactly equivalent, but how can the differences in the *Name* linking variable be expressed? String comparator metrics are now well-known, and some resulted from the need to compare strings for matching purposes. Jaro (1989), Jaro (1995), and Winkler (1990) are seminal works which produced the Jaro-Winkler comparator, a metric producing a value between 0 and 1 to determine how similar two strings are. A thorough examination of the Jaro-Winkler comparator is in Herzog et al. (2007), and a deeper examination of more string comparators is in Cohen et al. (2003).

In an early implementation of computer-based record linkage, Newcombe et al. (1959) compared strings using the Russell Soundex Code, which breaks words into phonetic codes of numbers and letters. Those authors used record linkage to determine if health and fertility were affected by exposure to low levels of radiation. Since exposure, marriage, births, and illness information were contained in different files, there was a need to link them with variables common to all files. This is perhaps the earliest example of using computers to implement record linkage, marking a seismic shift in the ability to link large data files, since linking could be done automatically and not solely by hand. Indeed, the advent of computer technology is a key reason for the interest generated for record linkage beginning in the 1960s (Fellegi, 1999).

The work of Newcombe et al. (1959) was a motivator for the formative Fellegi-Sunter method discussed in the Introduction. After the establishment of their method, record linkage surged in popularity. Early use cases included matching insurance claims to medical statistics (Bell et al., 1994), immigration record matching (Copas and Hilton, 1990), and matching records for the Census Bureau (Mulry et al., 2006), to name but a few. If the two files to be linked are not complete enumerations of the populations they represent, inference resulting from the linkage falls under the purview of sampling. For example, if the goal is to examine the relationship between marital status and number of children, as in the toy example from Fig. 1, because there is no complete list of everyone in the world along with their marital status, the files represent samples of people. When inference is made from the matches, the analyst is engaging in estimation resulting from samples. If the files are representative samples, then the inference is valid and well supported. Indeed, most statistical inference results from samples of data, so this is not necessarily an issue for record linkage. However, early record linkage literature lacks discussions regarding the assumption of representativeness in the datasets to be linked.

Another assumption often implicitly made in early record linkage papers is that errors in matching, e.g., false-positive and false-negative matches, do not affect the results of subsequent analyses. In the current research of record linkage, some effort is spent examining how these errors can affect the final analyses. Next, we discuss this along with current research and uses of record linkage.

3 Current Research and Uses of Record Linkage

Record linkage is currently used in medicine (Hallifax et al., 2018) and insurance (Boudreaux et al., 2015), at the Census Bureau (Abowd et al., 2019), and for big data fusion in general (Dong & Srivastava, 2015). Christen (2019) gives a useful and concise treatment of record linkage and includes additional current applications for further reading. Some of these applications have been studied since the inception of record linkage, but over time, research continues to expand the field.

One way the literature is expanding is in the methods used for record linkage, namely, via the introduction of machine learning techniques. The continued improvement in computing power combined with statistical techniques has allowed machine learning methods to be employed across industries and disciplines. Record linkage is no exception, as evidenced by Jurek et al. (2017) who introduced an ensemble learning method for unsupervised record linkage and Christen (2008) who developed a classification technique for record linkage involving support vector machines. There are many examples of machine learning used for record linkage since it can be distilled to a classification problem (match or non-match), a common use for machine learning. In addition to machine learning, Bayesian methods have also been introduced to record linkage. For example, Dalzell and Reiter (2016) took a Bayesian approach and derived a method to concurrently find matches and estimate the regression model.

In another avenue of current work, scholars are studying how the randomness associated with probabilistic linkage affects subsequent analyses. This was discussed in Neter et al. (1965), and it continues to be an area of active research. Recently, Chambers and Diniz da Silva (2020) noted (citing Harron et al., 2016) analysts' abilities to rigorously account for various biases and errors in linked data cannot keep pace with the inception of such datasets. Given the prevalence and availability of big data, this is an important issue for study. Chambers and Diniz da Silva (2020) suggest using paradata (data about the linkage process) to correct for biases resulting from linkage errors.

An important paper regarding analyses done with linked data is Lahiri and Larsen (2005). These authors investigated how errors in linkage affect regression analysis done using the linked data. By handling linking errors as measurement errors, they proposed an unbiased bootstrap regression estimator for use when there are matching errors. Chipperfield and Chambers (2015) similarly derived a parametric bootstrap method for evaluating categorical variables from linked datasets. Chambers (2009) examined ways to remove bias in regression analysis resulting from linking errors and took a specific look at logistic regression as well. Additionally, Zhang and Tuoto (2021) developed a regression approach in the presence of linkage errors and offered a diagnostic hypothesis test for examining assumptions about the linkage errors. Chipperfield (2020) approaches this problem by using bootstrap methods to replicate the linkage procedure in each replicate, along with estimating equations, to make inference in the presence of linkage errors. In both Briscolini et al. (2018) and Salvati et al. (2021), the authors investigate several methods to handle linkage errors when the context is small area estimation. Last, Kim and Chambers (2012) develop ways of correcting for the bias due to linkage errors, including incomplete or missed links, when employing regression after linking sample data to a register (dataset of the entire population), which was discussed in Sect. 1.

Most work in this stream focuses on regression analyses of linked data. However, there are other inferential methods which use linked data, such as sampling estimation. Zhang (2021) recently developed several generalized regression estimators (GREG) (see Särndal et al., 1992) for estimating totals when the sample and the auxiliary information, used in GREG estimators, cannot be perfectly matched. Their work builds on research from Breidt et al. (2017) who examined a difference estimator (type of GREG estimator) when matching between samples is imperfect.

Stokes et al. (2021) similarly attempt to examine the effect of matching errors on estimates of total. In their work, the authors employed capture-recapture methodology where the capture sample was electronic self-reports of fish catch (non-probability sample) and the recapture sample was a randomized dockside intercept sample of anglers (probability sample). Record linkage was used to link the two samples, and then estimates of total were made from the linked data. The authors developed a theoretical model for the probability of linking specific records and derived an expression for the approximate relative bias of an estimator as a function of various levels of matching error (including false-positive and false-negative errors). The works of Stokes et al. (2021), Zhang (2021), and Breidt et al. (2017) discussed here represent a bridge to the future of record linkage in survey sampling.

4 Future Uses of Record Linkage and Open Questions

A bright future of record linkage in survey sampling exists in the combination of non-probability samples with probability samples. As noted in Wiśniowski et al. (2020), the benefits of blending a non-probability sample with a probability sample are substantial. Elliott and Haviland (2007) did this by combining estimators from a probability sample with a web-based non-probability sample. They note the probability sample must be large for useful estimation. Recently, Sakshaug et al. (2019) offered a Bayesian approach for analyzing data from a smaller probability sample blended with a larger non-probability sample. They used the non-probability samples to construct priors for the model and show their approach worked well to reduce mean square error in estimates even when bias was present in the non-probability samples, a usual concern when investigating non-probability samples. These papers, however, do not link specific observations across datasets (samples) but seek to harness the information from both samples to improve the overall estimation.

Often, for inference, the non-probability sample is adjusted or weighted to have similar characteristics as the target population or to be used as auxiliary information (Elliott, 2009; Brus & Gruijter, 2003; Valliant & Dever, 2011). Another framework is to link actual records appearing in two samples, one a probability sample and one a non-probability sample. This occurs if the non-probability sample and the probability sample are subsets of the same population with increased overlap between the two as the non-probability sample size grows.

Specifically, call the population of interest U, the set of observations comprising the probability sample s_p, and the set of observations comprising the non-probability sample s_{np}. Then $s_p \in U$ and $s_{np} \in U$ and as $|s_{np}| \rightarrow |U| \Rightarrow P(s_p \cap s_{np}) = \emptyset) \rightarrow 0$. By examining the overlapping observations between the two samples, inference can be improved. This is how Liu et al. (2017) approached the problem of estimating fish catch in the Gulf of Mexico when they combined a voluntary sample of captains' fishing reports with a random intercept of boats returning to the dock. The overlapping trips, trips both reported and intercepted, provide auxiliary information, namely, measurement error estimates, which is incorporated into the estimator. This is an example of combining samples via matching and is a great application for record linkage.

While Liu et al. (2017) operate in a capture-recapture framework, using record linkage to combine a non-probability and a probability sample need not exist in such a setting. Examining the overlap, the matched set of entities between the samples, can provide accurate and useful auxiliary information to be used along with current non-probability sampling methods such as pseudo-weights or propensity scores. As data from non-probability samples become more available in ever-increasing sizes, linking them to existing or new probability samples will become more and more feasible. Regardless of the final use, record linkage certainly has a role to play.

In the future, assuming record linkage takes an increasing role in non-probability sample inference, there are several research questions which should define the next era of record linkage literature. We present a few open questions which should steer future research regarding record linkage in survey sampling.

The main research question of interest is: "what is the total error framework for linked data?" This question is closely linked to the idea of a total survey error (TSE) framework; see Groves and Lyberg (2010) for a thorough discussion of the TSE framework. The TSE framework decomposes the sources of error and bias when making inferences from surveys. This idea was recently extended in Amaya et al. (2020) for big data. They proposed a total error framework (TEF) for analyzing big data which has specific differences from the usual TSE framework. The authors discuss how certain errors manifest differently when applied to big data, such as coverage error, non-response error, and measurement error, to name a few (Amaya et al., 2020). Meng (2018) adopts a similar framework for making inferences from non-probability samples. He derived a formula to describe the difference between the population and sample averages as the product of measures of data quality, data quantity, and the problem difficulty (standard deviation of the variable of interest). Such previous research informs a TEF for linked data.

When analyzing linked data, a new source of randomness is introduced into the estimation which comes from linking errors. When considering a TEF for linked data, the linkage errors form a new component in the framework. The framework can be expressed as $\textit{Total Error} = \textit{Sampling Error} + \textit{Non-Sampling Error} + \textit{Linkage Error}$. Previous work has been done to examine both sampling error and non-sampling error in both the traditional, big data, and non-probability settings (Groves & Lyberg, 2010; Amaya et al., 2020; Meng, 2018). These three sources of error are broad and encompass many errors within them, e.g., *non-response error* is a subset of non-sampling error. Though these subsets have been investigated for sampling error and non-sampling error, there needs to be a partitioning of linkage error to build the TEF for linked data.

Stokes et al. (2021) started down this path by deriving a model for the effect linking errors have on the approximate relative bias of estimates made from linked data. Their model considers response rates and the discrepancies in the measurements when records are incorrectly linked. The model is generalizable and used to examine the effect of linking errors on the bias when estimating a total. Their work should be extended and further generalized to understand the effect of linkage errors within a total error framework. Linkage errors are especially difficult to partition because each linking scenario is different (Bell, 2017). Additionally, the magnitude of the effect of different linking errors will differ depending on various factors such as the amount of measurement error existing among matched records and if various errors can balance each other out (e.g., false-positive errors vs false-negative errors). Another source of linkage error that deserves further research is coverage error resulting from false-negative or unmatched links. That is, because some records are not linked, error arises. But this error is unique in such a context because the probability of linking two records can depend on the linkage algorithm (e.g., one-to-many linkage or one-to-one linkage) as well as the likelihood that other records link to each other.

A secondary question within the TEF for linked data has to do with estimating matching error if one lacks training data or the ability to perform clerical review. Training data offers a set of true links on which a record linkage algorithm can be tested. Clerical review is the term for manual inspection of potential links to determine if they match or not. Clerical review is usually the gold standard way to evaluate links if the entities refer to people or addresses, such as in the example from Fig. 1.

An example of when clerical review might be impossible is if an analyst links health data from wearable electronic devices to a census probability sample. In that case, manual review of links may prove too difficult to confidently mark links as false-positives, false-negatives, true-positives, or true-negatives. This might be true if the variables used for linking are error prone or if human judgment does not do a good job at determining true match status. Human judgment might also not be useful if no names or strings are used as linking variables, but instead identification numbers or usernames comprise the linking variables. In these settings, a sensitivity analysis for different levels of matching error will prove useful. In the future, a rigorous framework for such sensitivity analyses or methods of expressing confidence in the link states (match vs non-match) deserves careful thought as part of a TEF for linked data.

Another secondary question in this framework manifests when more than two files are to be linked. As stated earlier, the methodologies for linking two files extend to linking three or more files. However, it is likely that the data structures will differ for the different datasets. Each may have distinct and possibly different error sources. It may be that when linking three files (A, B, and C), a record $a \in A$ may be a false-positive link to record $b \in b$ but be a false-negative match to record $c \in C$. If records from one dataset are allowed to link to multiple records from the other datasets (not uncommon in record linkage), the errors and their effects can quickly build up. The implications of linking multiple data files, which likely will be more common in the big data climate of the day, must be considered and included in the TEF for linked data. This issue is under consideration, as seen in Kim and Chambers (2015).

This total error framework is critical for record linkage in survey sampling. Record linkage as a method continues to grow and has its own set of questions deserving inspection, such as issues of privacy (see Vatsalan et al., 2017) and how record linkage can fit into artificial intelligence programs, but we leave those questions to others since that is not in the scope of this chapter.

To conclude, record linkage is a technique which despite being in existence for 75 years continues to thrive. The ubiquitous nature of non-probability data in our world demands rigorous methods to analyze it. In the overlap between big data, non-probability samples, and statistical sampling lies record linkage. This is an exciting time to research record linkage as it will play an important role in statistical sampling in the future.

References

Abowd, J. M., Abramowitz, J., Levenstein, M. C., Mccue, K., Patki, D., Raghunathan, T., Rodgers, A. M., Shapiro, M. D., & Wasi, N. (2019). Optimal probabilistic record linkage: Best practice for linking employers in survey and administrative data. Center for Economic Studies Working Paper Series Working Paper Number CES-19-08.

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology, 8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Baker, R., J. M. Brick, Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). *Report of the AAPOR task force on non-probability sampling*. American Association for Public Opinion Research. www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Bell, R. M. (2017). *Diverse applications of probabilistic record linkage: Schucany lecture series*. Southern Methodist University.

Bell, R. M., Keesey, J., & Richards, T. (1994). The urge to merge: Linking vital statistics records and Medicaid claims. In *Medical care* (pp. 1004–1018).

Boudreaux, M. H., Call, K. T., Turner, J., Fried, B., & O'Hara, B. (2015). Measurement error in public health insurance reporting in the American community survey: Evidence from record linkage. *Health Services Research, 50*, 1972–1995. https://doi.org/10.1111/1475-6773.12308

Breidt, F. J., Opsomer, J. D., & Huang, C.-M. (2017). Model-assisted survey estimation with imperfectly matched auxiliary data. In: *TES 2018: Predictive econometrics and big data, studies in computational intelligence*.

Briscolini, D., Di Consiglio, L., Liseo, B., Tancredi, A., & Tuoto, T. (2018). New methods for small area estimation with linkage uncertainty. *International Journal of Approximate Reasoning, 94*, 30–42. https://doi.org/10.1016/j.ijar.2017.12.005

Brus, D., & Gruijter, J. D. (2003). A method to combine non-probability sample data with probability sample data in estimating spatial means of environmental variables. *Environmental Monitoring and Assessment, 83*(3), 303–317. https://doi.org/10.1023/A:1022618406507

Chambers, R. (2009). *Regression analysis of probability-linked data*. Official statistics research series (Vol. 4). Statistics New Zealand. oCLC: 908449516.

Chambers, R., & Diniz da Silva, A. (2020). Improved secondary analysis of linked data: A framework and an illustration. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 183*(1), 37–59. https://doi.org/10.1111/rssa.12477

Chipperfield, J. (2020). Bootstrap inference using estimating equations and data that are linked with complex probabilistic algorithms. *Statistica Neerlandica, 74*(2), 96–111. https://doi.org/10.1111/stan.12189

Chipperfield, J. O., & Chambers, R. L. (2015). Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. *Journal of Official Statistics, 31*(3), 397–414. https://doi.org/10.1515/jos-2015-0024

Christen, P. (2008). Automatic training example selection for scalable unsupervised record linkage. In *Advances in knowledge discovery and data mining, 12th Pacific-Asia conference PAKDD* (pp. 511–518).

Christen, P. (2019). Data linkage: The big picture. *Harvard Data Science Review* https://doi.org/10.1162/99608f92.84deb5c4

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the 2003 International Conference on Information Integration on the Web* (pp. 73–78).

Copas, J. B., & Hilton, F. J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society Series A (Statistics in Society), 153*(3), 287. https://doi.org/10.2307/2982975

Dalzell, N. M., & Reiter, J. P. (2016). Regression modeling and file matching using possibly erroneous matching variables. arXiv preprint arXiv:160806309.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dong, X. L., & Srivastava, D. (2015). *Synthesis lectures on data management:Big data integration*. Morgan and Claypool. https://doi.org/10.2200/S00578ED1V01Y201404DTM040

Dunn, H. L. (1946). Record linkage. *American Journal of Public Health and the Nation's Health, 36*(12), 1412–1416.

Elliott, M. N., & Haviland, A. (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey methodology, 33*(2), 211–215. http://www.thewitnessbox.com/10498-en.pdf

Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice, 2*(6), 1–7. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.981.4054&rep=rep1&type=pdf

Fellegi, I. P. (1999) Record linkage and public policy—a dynamic evolution. In: *Record Linkage Techniques—1997 Proceedings of an International Workshop and Exposition*. National Academies Press, (pp. 1–12).

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210. https://doi.org/10.2307/2286061

Groves, R. M., & Lyberg, L. (2010). Total survey error: past, present, and future. *Public Opinion Quarterly, 74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Hallifax, R., Goldacre, R., Landray, M. J., Rahman, N. M., & Goldacre, M. J. (2018). Trends in the incidence and recurrence of inpatient-treated spontaneous pneumothorax. *JAMA, 320*. https://doi.org/10.1001/jama.2018.14299

Harron, K., Goldstein, H., & Dibben, C. (Eds.). (2016). *Methodological developments in data linkage*. Wiley.

Herzog, T. N., Scheuren, F., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer. oCLC: ocn137313060.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association, 84*, 414–420.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine, 14*, 491–498.

Jurek, A., Hong, J., Chi, Y., & Liu, W. (2017). A novel ensemble learning approach to unsupervised record linkage. *Information Systems, 71*, 40–54. https://doi.org/10.1016/j.is.2017.06.006

Kim, G., & Chambers, R. (2012). Regression analysis under incomplete linkage. *Computational Statistics & Data Analysis, 56*(9), 2756–2770. https://doi.org/10.1016/j.csda.2012.02.026

Kim, G., & Chambers, R. (2015). Unbiased regression estimation under correlated linkage errors: Correlated linkage errors. *Stat, 4*(1), 32–45 https://doi.org/10.1002/sta4.76

Kim, J., & Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review, 89*(2), 382–401. https://doi.org/10.1111/insr.12434

Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association, 100*(469), 222–230. https://doi.org/10.1198/016214504000001277

Liu, B., Stokes, L., Topping, T., & Stunz, G. (2017). Estimation of a total from a population of unknown size and application to estimating recreational red snapper catch in Texas. *Journal of Survey Statistics and Methodology, 5*(3), 350–371. https://doi.org/10.1093/jssam/smx006

Lohr, S. L. (2010). *Sampling: Design and analysis* 2nd ed.. Brooks/Cole.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics, 12*(2). https://doi.org/10.1214/18-AOAS1161SF

Mulry, M. H., Bean, S. L., Bauder, D. M., Wagner, D., Mule, T., & Petroni, R. J. (2006). Evaluation of estimates of census duplication using administrative records information. *Journal of Official Statistics, 22*(4), 655–679.

Neter, J., Maynes, E. S., & Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association, 60*(312). https://doi.org/10.2307/2283401

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science, 130*(3381), 954–959.

Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., & Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics, 35*(3), 653–681. https://doi.org/10.2478/jos-2019-0027

Salvati, N., Fabrizi, E., Ranalli, M. G., & Chambers, R. L. (2021). Small area estimation with linked data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83*(1), 78–107. https://doi.org/10.1111/rssb.12401

Stokes, S. L., Williams, B. M., McShane, R. P. A., & Zalsha, S. (2021). The impact of nonsampling errors on estimators of catch from electronic reporting systems. *Journal of Survey Statistics and Methodology, 9*(1), 159–184. https://doi.org/10.1093/jssam/smz042

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

Valliant, R., Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research, 40*(1), 105–137. https://doi.org/10.1177/0049124110392533

Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017) *Privacy-preserving record linkage for big data: Current approaches and research challenges*. Springer. https://doi.org/10.1007/978-3-319-49340-4_25

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods American Statistical Association* (pp. 354–359).

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology, 8*(1), 120–147. https://doi.org/10.1093/jssam/smz051

Zhang, L., & Tuoto, T. (2021). Linkage-data linear regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 184*(2), 522–547. https://doi.org/10.1111/rssa.12630

Zhang, L.-C. (2021). Generalised regression estimation given imperfectly matched auxiliary data. *Journal of Official Statistics, 37*(1), 239–255. https://doi.org/10.2478/jos-2021-0010

Part III
Educational and Behavioral Statistics

A Bayesian Latent Variable Model for Analysis of Empathic Accuracy

Linh H. Nghiem, Benjamin A. Tabak, Zachary Wallmark, Talha Alvi, and Jing Cao

Abstract Empathic accuracy (EA), defined as the ability to accurately understand the thoughts and emotions of others, has become a well-studied phenomenon in social and clinical psychology. A widely used computer-based EA paradigm compares perceivers' ratings of targets' feelings or affective states with the ratings of target themselves (the true ratings) and uses correlation or its monotonic transformation as a measure of EA. However, correlation has a number of notable limitations. In particular, perceivers may differ in their rating patterns, but still have similar overall correlations. To overcome the limitations, we propose a Bayesian latent variable model that decomposes EA into two separate dimensions—discrimination and variability. Discrimination measures perceivers' sensitivity in relation to the true ratings, and variability measures the variance of random error in perceiver's perceptions. Similar to the conventional correlation, the Bayesian model is able to measure the overall level of the association between perceiver and target, but more importantly, the Bayesian approach can provide insights into how perceivers differ in their EA. We demonstrate the advantages of the new EA measures in two case studies. The proposed Bayesian model has a simple specification and is easy to use in practice due to its straightforward implementation in popular software. The R code is included in the supplementary material.

L. H. Nghiem (✉)
School of Mathematics and Statistics, University of Sydney, Sydney, NSW, Australia
e-mail: linh.nghiem@sydney.edu.au

B. A. Tabak · T. Alvi
Department of Psychology, Southern Methodist University, Dallas, TX, USA
e-mail: btabak@smu.edu; talvi@smu.edu

Z. Wallmark
Department of Musicology and Ethnomusicology, University of Oregon, Eugene, OR, USA
e-mail: zwallmar@uoregon.edu

J. Cao
Department of Statistical Science, Southern Methodist University, Dallas, TX, USA
e-mail: jcao@smu.edu

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_10

1 Introduction

Empathic accuracy (EA) is defined as the ability to correctly infer the thoughts and emotions of others (Zaki et al., 2009). In addition to the role of EA in the development and maintenance of healthy social relationships (Sened et al., 2017), clinical research has shown that performance in a standard EA video task can differentiate individuals with certain psychiatric disorders from healthy controls (Lee et al., 2011). Thus, the study of EA can help us understand general social functioning and also help identify social cognitive impairment in clinical populations.

There are a number of ways to examine EA, including matching categorical assessments (Schweinle et al., 2002) or continuous real-time assessments of the affective states of people (hereafter referred to as targets) by participants (hereafter, perceivers) (Zaki et al., 2008). Studies of EA that focus on matching categorical assessments between perceivers and targets often use signal detection theory in analyses. However, continuous EA data do not allow for this type of analysis. The focus of this study is on the analysis of EA tasks based on continuous real-time ratings. For example, EA paradigms may include a set of brief video clips in which targets discuss positive or negative events in their lives. Perceivers are asked to rate how negative or positive the target is feeling when discussing autobiographical events in real time using a 9-point scale (e.g., 1 = extremely negative; 9 = extremely positive). Responses from perceivers are captured in 2–5 s epochs throughout each video clip, and these responses are then compared to the responses of the targets, who watched the videos of themselves and completed the same ratings task in order to create a canonical index of "true" responses.

Traditionally, correlational analysis (and its monotonic transformation) is the conventional and arguably most common statistical method used for analysis of the continuous EA data. For example, based on several videos in which social targets discussed emotional events, Zaki et al. (2009) collected ratings averaged across 5-s periods and computed the Fisher transformation of the Pearson correlation coefficient to measure perceivers' EA. Also, in an fMRI validation study of a modified EA task, Mackes et al. (2018) computed the same measure and conducted paired samples t-tests to examine the neural correlates of perceived emotional intensity and mentalizing. However, this one-dimensional correlation approach, which only measures the linear association between two variables, may leave out important patterns in the data. First, unlike weight or height, EA is a latent merit that cannot be directly measured in absolute terms. For example, in a given task, the same rating may mean something different to different perceivers. In addition, although all perceivers are given the same scales (such as from 1 to 9), different perceivers may subjectively choose different ranges of their own ratings (e.g., one person may always give ratings from 4 to 7, while another person may use the whole range from 1 to 9). Second, there are at least two underlying behavioral dimensions that contribute to the discrepancy between perceivers' and targets' ratings, including different interpretations of the scale range and the random error in perceivers' ratings. These two dimensions are distinct, so it is necessary to incorporate both of them when measuring EA. Third, correlation can only be calculated for each stimulus separately, but an EA study typically involves a number of stimuli (such as multiple videos under one condition). Due to all of the issues raised here, statistical analysis based on correlation may limit the amount of information that can be gained from EA studies.

In a broader context of modeling accuracy of human judgment, a few approaches have been proposed as an alternative for correlation, yet these approaches typically require additional data compared to what we have for our applications. For example, West and Kenny (2011) proposed the truth and bias model, in which perceivers' responses to a stimulus are assumed to be influenced by a truth force and a bias force. To use this model, each perceiver is typically asked to provide not only a response toward the target but also a self-judgment response. In our application, we only have the former but not the latter. Biesanz (2010) proposed the social accuracy model, in which accuracy of a judgment is into distinctive accuracy, the extent to which a perceiver can perceive the distinct and unique characteristics of one person, and normative accuracy, a measure of how a perceiver's perception of others corresponds to the same perceiver's perception of an average person. This social accuracy model is commonly used in modeling perception of traits, where a perceiver is asked to rate different traits of other people, and the ratings of these traits for an average person (a normative profile) are available from a larger sample or a meta-study. In our application, a perceiver is asked to provide a continuous rating over time to judge the emotion of a specific target. To the best of our knowledge, the normative profile for these continuous ratings are not available.

As pointed out by an anonymous referee, one may tend to conduct the Bland-Altman analysis (Bland & Altman, 1999) between the perceivers' and targets' ratings. In most of the applications, the Bland-Altman (BA) analysis aims to evaluate whether two different devices give the same measurements of an objective quantity. For example, in Doğan (2018), the BA method is used to evaluate whether a venous blood gas analysis and a biochemistry panel shows the same level of potassium in patients. However, the BA method is not appropriate for measuring EA. First, for one specific task, the perceivers' and target's ratings are not expected to be the same, because they can have different (subjective) interpretation of the rating scale. For example, a rating of “5” for one person is not the same as a rating of “5” for another. Furthermore, just knowing whether perceivers and targets agree with each other may be even less informative than using correlation, since the BA analysis does not quantify the extent to which a perceiver agrees with the target.

In this article, we introduce a Bayesian latent variable approach to model EA response data that is based on the previous work by Cao and colleagues (Cao et al., 2010; Cao & Stokes, 2017). The proposed Bayesian model identifies two latent dimensions of EA—discrimination and variability—that are identifiable when perceivers' ratings differ from the targets' ratings. Discrimination measures a perceiver's ability to distinguish changes in a target's emotions, while a perceiver's variability measures the variance of random error in perceivers' ratings (i.e., the difference between perceiver's and target's ratings due to inconsistency). A smaller variance implies that the perceiver has a higher level of consistency in perceiving the target. Using the proposed Bayesian model, we are able to estimate perceivers' discrimination and variability and hence obtain more valuable information about their EA perception than correlation, which only measures the general association between perceivers' and target's ratings.

We begin by introducing the Bayesian model, including model specification and software implementation. We then describe the advantages of the Bayesian model using two case studies. In the first case study, we re-analyze the dataset in Devlin et al. (2014) that consists of perceivers' ratings of four distinct videos in which targets discuss emotional events in their lives. In this case study, we focus on explaining the underlying dimensions of EA and comparing the Bayesian estimates of discrimination and variability with the standard correlational measure. In the second case study, we analyze perceivers' ratings of 12 original music recordings expressing musician-targets' renderings of four primary emotions (three recordings per emotion), with the focus on how the underlying EA dimensions are associated with the musicality (i.e., level of musical skill and training) of the perceivers. This case study further demonstrates that the new measures can facilitate additional insights on how EA perception is related to perceivers' characteristics.

2 Methodology

In an EA study, suppose that there are n perceivers instructed to provide ratings on J stimuli, where each stimulus has K_j units (i.e., there are K_j points in the sequence of ratings, which can vary among stimuli). Each stimulus corresponds to a specific target. Let x^r_{jk} denote the raw rating given by the corresponding target for the kth unit of the jth stimulus, $j = 1, \dots, J$ and $k = 1, \dots, K_j$. Note that similar to correlational analysis, the mean of target score will not affect the measurement on EA. Hence, to simplify the model specification, the raw ratings x^r_{jk} are centered for each stimulus, where the centered rating is denoted as $x_{jk} = x^r_{jk} - K_j^{-1} \sum_{m=1}^{K_j} x^r_{jm}$ and is treated as the true rating. Similarly, letting y^r_{ijk} denote the rating given by the ith perceiver for the kth unit of the jth stimulus, then the corresponding centered rating is $y_{ijk} = y^r_{ijk} - K_j^{-1} \sum_{m=1}^{K_j} y^r_{ijm}$. We specify the Bayesian latent variable model to measure EA as

$$\begin{aligned} y_{ijk} &= \beta_{ij} x_{jk} + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\sim N(0, \sigma_i^2), \quad \beta_{ij} \sim N(\beta_i, \sigma_\beta^2), \end{aligned} \tag{1}$$

for $i = 1, \dots, n,\ j = 1, \dots, J,\ k = 1, \dots, K_j$. In the model, β_{ij} represents the ith perceiver's discrimination level on the jth stimulus, which is assumed to follow a normal distribution with mean β_i and variance σ^2_β. Note that β_i is the ith perceiver's average discrimination level over all the J stimuli. We allow a perceiver to have different discrimination levels for different stimulus, but assume these discrimination levels are similar by imposing a random-effect structure on all β_{ij}'s.

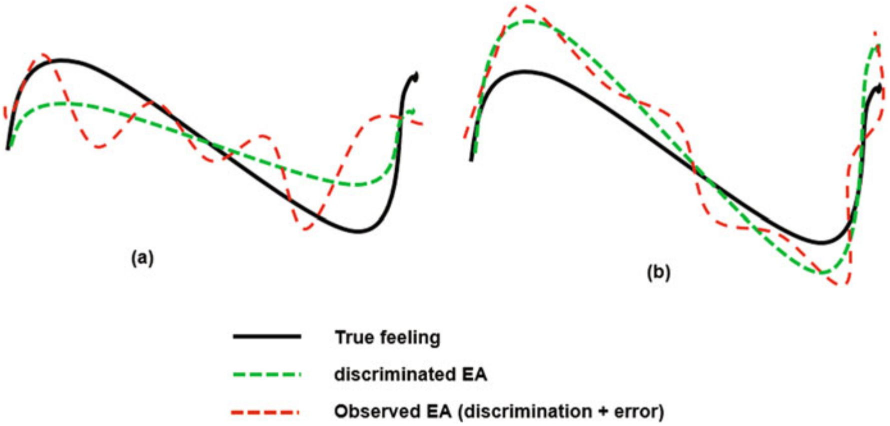


Fig. 1 Illustration of the two latent dimensions in EA. Plot (**a**): a subject has an attenuated discrimination and relatively large variability. Plot (**b**): a subject has a magnified discrimination and relatively small variability

An empathic perceiver's discrimination parameter β_i will be positive, indicating that on average, the perceiver's response has a congruent association with the target. A smaller value β_i suggests that the perceiver's response signal is more attenuated compared to a perceiver with a larger β_i value. Furthermore, a perceiver with a negative β_i has a response that moves in opposite direction compared to the target's ratings, yet these instances are generally rare in EA studies. Additionally, Model (1) contains the random error ε_{ijk}, which is assumed to follow a normal distribution with mean 0 and *perceiver-specific variance* σ_i^2. The smaller the variance, the higher the consistency in the perceiver's ratings, so we refer to σ_i^2 as the measure of the variability in EA of the ith perceiver.

Figure 1 illustrates two examples of concept on how the two latent dimensions of EA (i.e., discrimination and variability) contributes to the actual ratings given by a perceiver. The black line depicts the (observed) target true ratings. The green dashed line represents the (unobserved) expected ratings associated with a certain discrimination level, and the red dashed line represents the (observed) actual ratings after random errors are added to the green dashed line. The plot on the left shows an example of ratings with an attenuated discrimination (i.e., a less distinctive interpretation of true signals) and relatively large variability (i.e., a large deviation between the expected ratings and the actual ratings), and the plot on the right shows an example with a magnified discrimination and relatively small variability.

To complete the Bayesian model specification, the assignment of prior distribution is listed in the following:

$$\beta_i \sim N(1, 100), \quad \sigma_i^2 \sim \text{IG}(2, 1), \quad \sigma_\beta^2 \sim \text{IG}(2, 1), \tag{2}$$

for $i = 1, \ldots, n$. Note that the ratings have a range of 9 points, so the normal prior on β_i has a variance of 100, which is large enough to make the normal prior a non-informative prior. The mean of β_i is 1 because without any prior knowledge, we assume all perceivers have roughly the same interpretation as the true targets. The prior for the variance σ_i^2 is IG(2, 1), which is the inverse gamma distribution with a shape parameter of 2 and a scale parameter of 1, so the corresponding prior variance is infinite. Thus, the assigned priors are conventional conjugate non-informative priors, which facilitates data-driven inference and results in fast Bayesian computation (Sun et al., 2001). The proposed model is referred to as the BDV Model (Bayesian model with the latent dimensions on Discrimination and Variance). Finally, note that when there is only one stimulus in the EA study (i.e., $J = 1$), the BDV Model can be reduced to

$$y_{ik} = \beta_i x_k + \varepsilon_{ik}, \qquad \varepsilon_{ik} \sim N(0, \sigma_i^2), \tag{3}$$

where the priors for β_i and σ_i^2 are the same as in (2) for $i = 1, \ldots, n$.

Note that the general BDV Model with $J \geq 2$ is a random-effect model, with a random slope (perceiver-specific discrimination) and a perceiver-specific variance. The perceiver-specific variance is a novel and indispensable part of the model because it represents a unique EA dimension. In the applications below, we demonstrate that, compared to the same models with constant variance, i.e., $\sigma_i^2 = \sigma^2$ for all $i = 1, \ldots, n$, the incorporation of perceiver-specific variance improves the model fits significantly. After fitting the BDV Model, we use the posterior mean for β_i and σ_i^2 as the estimated discrimination and variability for the EA of the ith perceiver.

To facilitate the implementation of the model, we include the R code in the supplementary material. The code is based on "Just-Another Gibbs-Sampler"(JAGS) model, which is an open-source program designed to run Bayesian hierarchical model using Markov chain Monte Carlo methods (Plummer et al., 2003). With JAGS, users specify a model and its prior specification; then a Markov chain simulation is automatically implemented for the resulting posterior distribution. This frees users from manually deriving the MCMC algorithm, which is the main obstacle for the implementation of Bayesian inference in practice. JAGS is designed to work closely with the R language. Our code uses the `rjags` package (Plummer, 2019) as the interface from R to JAGS. Detailed instructions are annotated in the code.

3 Applications

3.1 Study on Social Empathic Accuracy

In our first application, we consider a study conducted by Devlin et al. (2014) that examined the relationship between perceivers' levels of positive emotion and EA. Their study included $n = 121$ perceivers, who watched four videos of targets discussing emotional events in their lives. These four videos vary in valence (positive or negative) and intensity (high or low), resulting in four non-homogeneous videos, including high-positive, low-positive, high-negative, and low-negative. While watching each video, perceivers provided continuous online ratings of the corresponding target's emotion using a 9-point scale (from 1 = extremely negative to 9 = extremely positive). The ratings from the perceivers were then compared with those from the targets.

To measure the EA of each perceiver, the authors calculated the Fisher transformation to the Pearson correlation between perceivers' ratings and targets' ratings for each video. In other words, each participant had four EA measures, each of which corresponds to one video. For the correlation coefficient r, the corresponding Fisher transformation is defined as $Z = (1/2)\log\{(1 + r)/(1 - r)\}$, where log denotes the natural logarithm. While r ranges from -1 to 1, the Fisher transformed correlation can take any value from the real line, so it is more appropriate to conduct statistical analyses with the normality assumption based on Z than based on r. In this paper, we refer to Z as the "r-to-Z EA estimate" and denote it to be rZ. The data from Devlin et al. (2014) are publicly available at https://doi.org/10.1371/journal.pone.0110470.

Because these four videos varied in valence and intensity, they should be treated as four distinct individual stimuli instead of multiple stimuli under one condition, and we fit the reduced BDV Model (3) to each of the four video stimulus separately. We begin with a graphical demonstration to illustrate how the two latent EA dimensions can provide more insights on EA compared to the conventional measure rZ. Figure 2 shows two plots, each depicting the ratings given by the target and those by three perceivers (selected for illustrative purposes) for the high-negative video. In each plot, the black line represents the true target's rating, and the other lines represent ratings of the selected perceivers watching the same video. The estimated rZs between the target's and perceivers' ratings are listed in the legend, along with the estimated discrimination and variability parameters (abbreviated as D and V, respectively). In the top panel, the three perceivers demonstrated very different rZ, whereas in the bottom panel, the three perceivers had similar rZ.

In the top plot of Fig. 2, the three perceivers (denoted as P1, P2, and P3) demonstrated very different EA levels, indicated by the varying estimated rZ values (1.51, −0.22, and 0.51). However, the correlational analysis does not explain why the three perceivers have such dramatically different EA scores. Based on the Bayesian estimates, we can see that P1's greater EA (red line, $\widehat{rZ} = 1.51$) is due to a higher level of discrimination ($\hat{D} = 0.81$) and smaller variability (i.e., higher consistency, $\hat{V} = 0.16$) when rating the target. The perceiver P2 has a negative correlation (green line, $\widehat{rZ} = -0.22$), which is due to the negative discrimination (i.e., the person perceived the target's emotion in the opposite direction, $\hat{D} = -0.14$). In addition, P2 also has the largest variability among the three perceivers reflecting the more obvious fluctuation of P2's ratings ($\hat{V} = 0.34$). Moreover, P3 (blue line, $\widehat{rZ} = 0.51$) has a moderate EA level: compared to P1, P3 has a lower discrimination (other than the initial drop, P3's ratings are quite flat, not showing the gradual decline in the target's ratings, $\hat{D} = 0.34$) and larger variability (the discrepancy between P3's ratings and the target's ratings are noticeably large in both ends of the series, $\hat{V} = 0.38$).

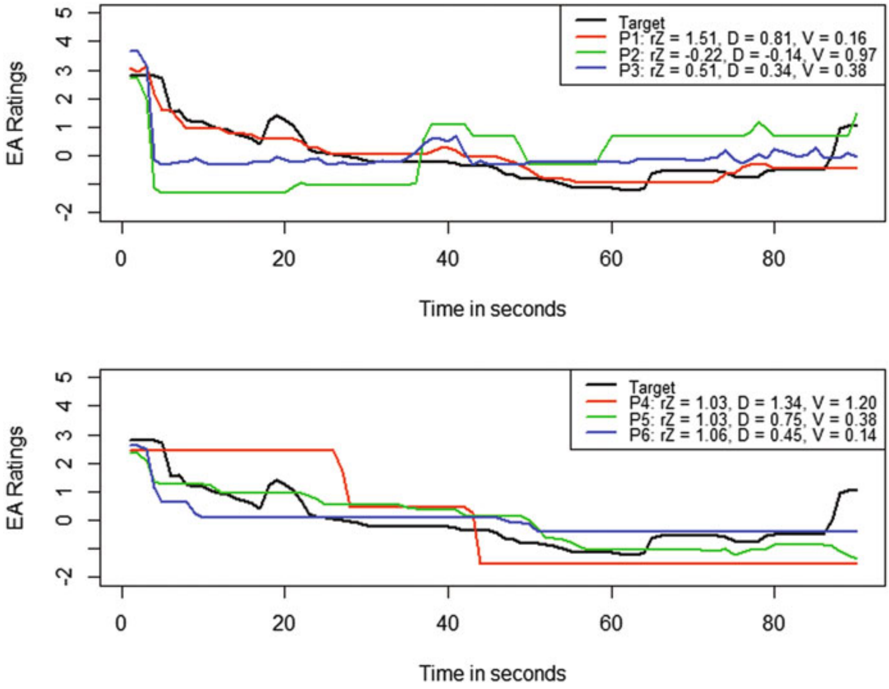


Fig. 2 Comparison of the target's ratings and the six perceivers' ratings in the high-negative video, where rZ denotes the r-to-Z transformed correlation and D and V represent discrimination and variance in the Bayesian model, respectively

In the bottom plot of Fig. 2, we chose data from three different perceivers (P4, P5, and P6) to further demonstrate the advantage of utilizing discrimination and variability to study EA over the rZ measure. In this case, the three perceivers have similar estimated rZ values (1.03, 1.03, and 1.06). Hence, based on the correlational analysis, these perceivers have similar EA. However, the estimates of discrimination and variability show that their underlying EA dimensions have distinct patterns. P4 (red) has a large discrimination value ($\hat{D} = 1.34$), resulting from the fact that P4's ratings have a more dramatic decline than the target's ratings. At the same time, P4 has the largest variability among the three perceivers, reflecting P4's pronounced shift toward negative ratings at around time units 25 and 40. In addition, P6 (blue line) has both the smallest discrimination ($\hat{D} = 0.45$) and smallest variability ($\hat{V} = 0.14$) among the three perceivers. Other than the initial drop, P6's ratings are mostly flat, only spanning a narrow range of scores. Unlike the dramatic decline in P4's ratings and slow change in P6's ratings, P5's (green line) ratings follow the gradual decline in the target's ratings. Because of the inadequate drop in the beginning and opposite change in the end of the series, P5 has a larger variability ($\hat{V} = 0.38$) than P6.

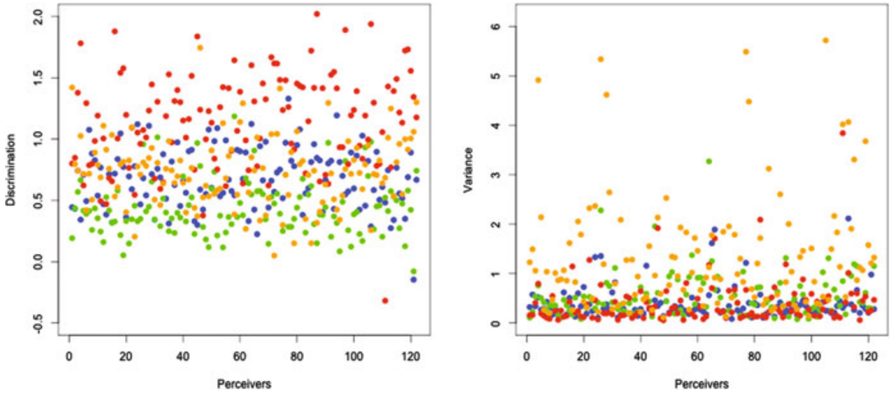


Fig. 3 Perceivers' video-specific discrimination and variance (red = high-positive, orange = low-positive, blue = high-negative, green = low-negative)

Based on the examples included in Fig. 2, we can see that the two latent discrimination and variability dimensions specified in the BDV Model offer unique information regarding EA compared to the correlation analysis. Specifically, the proposed BDV Model is able to explain how perceivers differ in their EA and to identify possible differences in the underlying dimensions in EA when the correlation may show no differences.

Next, we compare the latent dimensions across the four videos. Figure 3 (left panel) shows that perceivers had higher discrimination ability for the high-positive video (red dots) and lower discrimination ability for the low-negative video (green dots). These findings are in agreement with previous studies of showing greater EA for positive videos compared to negative videos in both healthy and clinical samples (Lee et al., 2011). As for the variability, the largest video-specific variances are from the two low-intensity videos (orange and green dots in the right panel of Fig. 3).

As we mentioned in the last section, a novelty of the BDV Model is that it incorporates perceiver-specific variances for random errors instead of assuming a constant variance as in most of the conventional random-effect models. We demonstrate the advantage of this functionality by comparing the model fit between the BDV Model and the following random-effect model with a *constant* variance:

$$y_{ik} = \beta_i x_{ik} + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma^2). \tag{4}$$

Note that Model (4) assumes that all the perceivers have the same variability, i.e., the same consistency level in EA. The model comparison is conducted using the deviance information criterion (DIC), where a small value of is preferred and a difference of more than 10 usually rules out the model with a higher DIC (Spiegelhalter et al., 2002). The results are summarized in Table 1. The evidence is clear and convincing that across all the four video groups, the BDV Model (3) provides much better model fits than Model (4). This data-driven evidence

Table 1 Model comparison between the BDV Model (3) and Model (4) using DIC

	High-negative	Low-negative	High-positive	Low-positive
BDV Model (3)	18288.58	12981.28	9318.62	20196.82
Model (4)	20758.17	14949.47	12063.96	22649.64

Table 2 Correlation between rZ and the Bayesian measures on EA

	High-negative	Low-negative	High-positive	Low-positive
Discrimination	**0.74 (p < 0.01)**	**0.57 (p < 0.01)**	**0.40 (p < 0.01)**	**0.52 (p < 0.01)**
Variance	**−0.63 (p < 0.01)**	−0.08 ($p = 0.41$)	**−0.72 (p < 0.01)**	**−0.52 (p < 0.01)**
B_{EA}	**0.99 (p < 0.01)**	**0.98 (p < 0.01)**	**0.96 (p < 0.01)**	**0.99 (p < 0.01)**

P-values are based on two-tailed tests and included in parentheses. Significant p-values (<0.05) are indicated in bold

supports the inclusion of the perceiver-specific variance, which further confirms that variability is a unique dimension aside from discrimination in EA.

Finally, we compare the results from the Bayesian model with the conventional rZ estimates. For each video, we compute the Pearson correlation between perceivers' estimated rZ estimates and the Bayesian estimates of discrimination and variability, respectively. Furthermore, we investigate the correlation between the rZ estimates and the estimates for β_i/σ_i, $i = 1, \ldots, n$, which is the ratio of the discrimination and the square root of random error's variance, similar to the measure used by Cao et al. (2010). We refer to this ratio as the "Bayesian EA aggregated estimate" and denote it as B_{EA}. Similar to rZ, the measure B_{EA} can take any value from the real line.A high B_{EA} implies that a perceiver has a relatively large discrimination and a relatively small variability.

Table 2 shows that the association between perceivers' rZ and B_{EA} is consistently high, with the correlation being almost 1. However, the association between perceivers' rZ and the latent dimensions on discrimination and variability are weak to moderate (though most of them are statistically significant). This indicates that the conventional correlation, as was used by Devlin et al. (2014) and most existing literature, only provides a valid aggregate measure for EA, but it does not provide much insight into the dimensions underlying the structure of EA.

3.2 *Study on Musical Empathic Accuracy*

In a study of the association between EA and accuracy of emotion recognition in music, Tabak et al. (In press) collected data from 415 undergraduate perceivers enrolled at Southern Methodist University. Perceivers participated in a novel music EA task, in which they listened to and rated 12 brief music recordings expressing the target's (musician's) primary emotions of joy/happiness, sadness, anger, and tenderness (3 recordings per emotion). Stimuli were solo piano pieces created by six composer-pianists. Identical to the video EA task in the previous case study, perceivers listened to the excerpts and provided continuous real-time response evaluations of how negative or positive (1 = very negative to 9 = very positive) they perceived the music to be. Samples were collected every 2 s. The same data collected from the composer-pianist targets provided the "true" target ratings to be compared with perceivers' ratings.

EA research has typically focused on cognitive empathy, i.e., perceivers' understanding of a target's thoughts, feelings, and general mental state (Zaki et al., 2009). However, recently Morrison et al. (2016) included an additional assessment of EA in which they slightly altered the instructions of the task to assess affect sharing or the extent to which a perceiver experiences the same emotion as a target (i.e., affective empathy). To examine the two different kinds of EA, perceivers in this study were randomized into an affective empathy group or a cognitive empathy group. In the affective empathy group ($n = 230$), perceivers were asked to provide their own emotional response when listening to the music, whereas in the cognitive empathy group ($n = 185$), they were instructed to try to understand the emotion being communicated or expressed by the composer-pianist in the recordings.

In this application, our goal is to use the BDV Model to investigate the association between perceivers' EA underlying dimensions and their musical training in both groups. Musical training has been shown to modulate emotion recognition of music (Di Mauro et al., 2018). Our aim here is to examine whether musical training is associated with the accurate perception of musical emotion, as operationalized according to the musician-targets' intent. Musical training is measured by the Goldsmiths Musical Sophistication Index (Müllensiefen et al., 2014), a psychometric tool for the measurement of musical attitudes, behaviors, and skills. For each group, we first compute the correlation between the estimates of each dimension in EA and the Gold-MSI among the perceivers. In addition, we examine the association of EA and Gold-MSI in three conditions: (1) across all 12 music recordings (i.e., $J = 12$), (2) among the 6 positive music recordings (i.e., $J = 6$) which consist of 3 recordings expressing happiness and 3 recordings expressing tenderness, and (3) among the 6 negative music recordings (i.e., $J = 6$) which consist of 3 recordings expressing sadness and 3 recordings expressing anger. Note that there are multiple stimuli under each condition, and it is not straightforward to compute an overall EA measure from the correlational analysis in this setting.

We fit the BDV Model (1) to multiple stimuli for each of the three above conditions. We then compute the Pearson correlation between the Gold-MSI and the Bayesian estimates of discrimination and variability, respectively. Table 3 provides the results for the affective empathy and the cognitive empathy groups. First, for the affective empathy group, none of the association between estimated discrimination nor variance with musical background is statistically significant. On the other hand, for the cognitive empathy group, we find a significant association between perceivers' estimated discrimination and their musicality across all the three conditions. However, the association between the estimated variability and perceivers' musicality is not statistically significant. In other words, higher levels of discrimination in the cognitive assessment of musician/targets' emotions are

Table 3 Correlation between empathic accuracy latent dimensions and musicality

	Affective group		Cognitive group	
	Discrimination	Variance	Discrimination	Variance
All music	−0.00 ($p = 0.98$)	−0.02 ($p = 0.78$)	**0.22 (p < 0.01)**	−0.09 ($p = 0.23$)
Positive music	0.07 ($p = 0.31$)	−0.01 ($p = 0.90$)	**0.19 (p = 0.01)**	−0.14 ($p = 0.06$)
Negative music	−0.00 ($p = 0.50$)	−0.02 ($p = 0.74$)	**0.15 (p = 0.04)**	−0.07 ($p = 0.33$)

P-values are based on two-tailed tests and included in parentheses. Significant p-values (<.05) are indicated in bold

Table 4 Model comparison between the BDV Model (1) and Model (5) using DIC

		All music	Positive music	Negative music
Affective group	Model (1)	183732	70507	107962
	Model (5)	215722	83849	131688
Cognitive group	Model (1)	167756	63654	97049
	Model (5)	189105	72736	116119

associated with perceivers' relative degree of musical ability, while the level of consistency is not. In contrast, the congruence of one's personal emotional responses to the musician's expressive intentions (EA for affective empathy) is not related to one's training and depth of musical knowledge. Finally, in order to confirm the need for including perceiver-specific variance, similar to what was done in the previous application, we compare the model fit between the BDV Model (1) and the following random-effect model with a constant variance:

$$y_{ijk} = \beta_{ij} x_{jk} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \; j = 1, \ldots, J, \; k = 1, \ldots, K_j. \tag{5}$$

The prior specification of model (5), other than the perceiver-specific variance, remains the same as that in (2). The model comparison using DIC is summarized in Table 4. It shows that for all the three conditions and for both the affective group and the cognitive group, DIC for the BDV Model (1) is substantially smaller than that for Model (5). The model comparison results provide strong evidence to show that the incorporation of the perceiver-specific variance improves model fit substantially. Whether examining one stimulus or multiple stimuli, variability, as measured by the perceiver-specific variance of the random error in the model, is a distinctive dimension of EA, which is inherently different from perceiver-specific discrimination. Thus, when looking at perceivers' EA patterns, including both dimensions provides more detailed information about perceivers' perceptions.

4 Conclusion

In this article, we have proposed a Bayesian latent variable model which serves as an alternative to the conventional correlational analysis for empathic accuracy (EA) research using continuous real-time assessments. The proposed BDV model has three main advantages over the correlational analysis. First, it is more sensitive to perceiver-level differences in EA studies, as reflected in varying response behaviors (e.g., using different ranges of the scale). Correspondingly, the BDV Model quantifies two behavioral dimensions of EA, discrimination, and variability. Similar to the correlational analysis, these two dimensions measure the overall EA level for each perceiver, but more importantly, they explain how perceivers differ in EA. Using correlation, many perceivers giving different rating patterns may have a similar EA level, but using discrimination and variability, these differences can be identified. Finally, while correlational analysis must be conducted independently for each individual target, the proposed model is capable of providing an overall EA measure where multiple stimuli are included under one condition of an EA task. Taken together, the Bayesian approach to EA can shed light on distinctions that are not detectable by simple correlational analysis.

There are many areas of research that can benefit from this approach. Broadly speaking, it could be used to increase the analytical precision of any experimental paradigm involving the comparison of sequential measurements on latent perceptual responses, such as research on social cognitive deficits in individuals with autism spectrum disorders and schizophrenia. The association of EA with social functioning in healthy and clinical populations has previously relied on the correlational approach to EA analysis (Lee et al., 2011). With the approach described here, researchers may be able to identify specific dimensions of EA that may be more or less impaired among clinical populations. For example, the discrimination parameter could be used to elucidate the extent to which the amplification of negative information and suppression of positive information that characterize individuals with depression (LeMoult & Gotlib, 2019) . The increased level of specificity could also benefit neuroscientists by examining the extent to which different dimensions of EA are correlated with real-time neural processing (Mackes et al., 2018). Furthermore, the BDV model can be improved in future research by incorporating other covariates that represent perceivers' and targets' characteristics. In general, improving the BDV model requires a consideration of both the quality of the model fit and its interpretability in the context of measuring EA.

In conclusion, the proposed Bayesian EA model is more flexible in handing perceiver-specific parameters than traditional correlational analysis. The model specification is simple, and the computation is efficient. Annotated R code is included to facilitate the implementation of the proposed model.

References

Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research, 45*(5), 853–885.

Bland, J. M., & and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135–160.

Cao, J., & Stokes, L. (2017). Comparison of different ranking methods in wine tasting. *Journal of Wine Economics, 12*(2), 203–210.

Cao, J., Stokes, S. L., & Zhang, S. (2010). A Bayesian approach to ranking and rater evaluation: An application to grant reviews. *Journal of Educational and Behavioral Statistics, 35*(2), 194–214.

Devlin, H. C., Zaki, J., Ong, D. C., & Gruber, J. (2014). Not as good as you think? trait positive emotion is associated with increased self-reported empathy but decreased empathic performance. *PloS One, 9*(10), e110470.

Di Mauro, M., Toffalini, E., Grassi, M., & Petrini, K. (2018). Effect of long-term music training on emotion perception from drumming improvisation. *Frontiers in Psychology, 9*, 2168.

Doğan, N. Ö. (2018). Bland-altman analysis: A paradigm to understand correlation and agreement. *Turkish Journal of Emergency Medicine, 18*(4), 139–141.

Lee, J., Zaki, J., Harvey, P.-O., Ochsner, K., & Green, M. F. (2011). Schizophrenia patients are impaired in empathic accuracy. *Psychological Medicine, 41*(11), 2297–2304.

LeMoult, J., & Gotlib, I. H. (2019). Depression: A cognitive perspective. *Clinical Psychology Review, 69*, 51–66.

Mackes, N. K., Golm, D., O'Daly, O. G., Sarkar, S., Sonuga-Barke, E. J., Fairchild, G., & Mehta, M. A. (2018). Tracking emotions in the brain–revisiting the empathic accuracy task. *NeuroImage, 178*, 677–686.

Morrison, A. S., Mateen, M. A., Brozovich, F. A., Zaki, J., Goldin, P. R., Heimberg, R. G., & Gross, J. J. (2016). Empathy for positive and negative emotions in social anxiety disorder. *Behaviour Research and Therapy, 87*, 232–242.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). Measuring the facets of musicality: The goldsmiths musical sophistication index (gold-msi). *Personality and Individual Differences, 60*, S35.

Plummer, M. (2019). *rjags: Bayesian graphical models using MCMC*. R package version 4-10.

Plummer, M. et al. (2003). Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vol. 124, pp. 1–10). Vienna, Austria.

Schweinle, W. E., Ickes, W., & Bernstein, I. H. (2002). Emphatic inaccuracy in husband to wife aggression: The overattribution bias. *Personal Relationships, 9*(2), 141–158.

Sened, H., Lavidor, M., Lazarus, G., Bar-Kalifa, E., Rafaeli, E., & Ickes, W. (2017). Empathic accuracy and relationship satisfaction: A meta-analytic review. *Journal of Family Psychology, 31*(6), 742.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583–639.

Sun, D., Tsutakawa, R. K., & He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica* 77–95.

Tabak, B. A., Wallmark, Z., Nghiem, L., Alvi, T., Sunahara, C. S., Lee, J., & Cao, J. (In press). Initial evidence for a relation between behaviorally assessed empathic accuracy and affect sharing for people and music. *Emotion*.

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review, 118*(2), 357.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science, 19*(4), 399–404.

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion, 9*(4), 478.

Variance Estimation for Random-Groups Linking in Large-Scale Survey Assessments

Bingchen Liu, Yue Jia, and John Mazzeo

Abstract The random-groups design is frequently used in equating and linking scores from two tests, in which the linking functions are derived from the test scores of two samples of the test-taker population. In this paper, we consider estimating variances of test score population statistics for large-scale survey assessments (LSAs), where the random-groups design is used in linking latent variable test scores. Examples of LSAs include National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Programme for International Student Assessment (PISA). In estimating variances of population statistics in LSAs, the common practice takes into account the uncertainties due to sampling and latency. In this paper, we propose a variance estimation method as an extension of the existing procedure that takes into account the random-groups linking. We illustrate the method using a NAEP dataset for which a linear linking function is used in linking test scores from a computer-based test to those from a paper-and-pencil test. The proposed method can be easily extended when random-groups equating and linking are applied to other assessment contexts, with linking functions being parametric or non-parametric.

B. Liu (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: bliu@ets.org

Y. Jia
Educational Testing Service, Princeton, NJ, USA
e-mail: yjia@ets.org

J. Mazzeo
(emeritus), Educational Testing Service, Princeton, NJ, USA
e-mail: mazzeo123@comcast.net

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_11

1 Introduction

In educational assessments, score linking is a general term that refers to relating scores from different tests or test forms (American Educational Research Association et al., 2014). This paper focuses on the random-groups linking in which one sample drawn from the population is administered one test form, while another sample drawn from the same population is administered a different test form. Based on the two samples selected from the common population, a linking function can be derived to transform the scores of one test form to the scores on the other test form (Kolen & Brennan, 2004).

Large-scale survey assessments (LSAs) are those used to monitor academic performance for populations (e.g., US fourth graders). One of the most important uses of LSAs is to track population statistics at a given time and changes in population statistics over time, such as how countries differ in students' mean scores on reading or how the mean reading scores in a country or a region change over time. Examples of LSAs include US National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Programme for International Student Assessment (PISA).

LSAs apply item response theory (IRT) latent variable regression models to directly estimate score distributional statistics for the population and subpopulations, such as population means and the percentage of students above specified proficiency levels (Mislevy, 1984, 1985). To provide a means to estimate population statistics, the programs also make available plausible values for individuals sampled from the population. Plausible values are random draws, or multiple imputations, from the performance distribution for individuals, conditional on the IRT latent regression model parameter estimates, response data, and contextual information (Mislevy, 1991; Braun & von Davier, 2017). In addition, LSAs make use of sampling weights to draw inferences from the probability-based samples to the population of interest. See, for example, von Davier et al. (2006) and Mazzeo (2018) on the design, sampling, and analysis of LSAs.

For LSAs, standard errors are estimated along with the population statistics. Typically, two general sources of variance are accounted for: sampling of test takers and latency of the test scores. The sampling variance accounts for the variability among the units in the population. The size of the sampling variance is in part a function of the sample design (see, e.g., Johnson & Rust, 1992). The latency variance reflects the uncertainty due to the statistics being estimated from the test-taker performance on a set of test questions and other auxiliary information used in the latent variable regression models. The latency variance is also referred to as between-imputation variance. Details on how these variances are estimated for LSAs are reviewed in Sect. 2.

One approach to estimate the sampling variance of a statistic is to use resampling methods such as the jackknife, balanced repeated replication (BRR), or bootstrap methods. These resampling methods create a number of subsamples and use the variability among the estimates from the subsamples to estimate the variance of the statistic. An alternative approach is to linearize the statistic (e.g., using the delta method or Taylor series expansion) and then estimate the variance of the linearized statistic analytically. Wolter (2007) described both approaches. Kish and Frankel (1974) showed in simulation studies that using a multistage design with two primary sampling units per stratum, both the jackknife and BRR gave acceptably low bias in estimated variance for various statistics. They also showed that these two methods gave results that were similar to those achieved via the Taylor series linearization. Many theoretical and empirical studies have also supported that the resampling methods perform well and result in comparable standard error estimates as the linearization approach (e.g., Krewski & Rao, 1981; Rao & Wu, 1985; Valliant, 1990; Shao, 1996).

In this paper, we consider the random-groups design where a sample of test-takers (referred to as the target sample) is administered assessment *T*, while another sample (referred to as the source sample) is administered assessment *S*. The scores from assessments *S* and *T* are estimated on the two separate latent variable scales. In addition, the scores from assessment *S* are linked to assessment *T* via random-groups linking. One example is to link scores from a paper-and-pencil test to a test given on a computer (Eignor, 2007; Jewsbury et al., 2020). Other examples are the studies in linking scores between two different LSAs (Johnson, 1998; Johnson et al., 2005; Jia et al., 2011).

For the random-groups design, the linking function coefficients are statistics calculated based on the source and target samples and using the test scores that are subject to latency variance. Kolen and Brennan (2004) discussed the use of the bootstrap to estimate the sampling variance of statistics for assessments with the random-groups linking. However, we are not aware of any real-data applications.

When the random-groups design is applied in linking the LSA test scores, uncertainty in the linking function is typically ignored. Mazzeo et al. (in press) offered an approach to approximate the variance associated with the linking function, as an additional source of variance, adding to the sampling and latency variances typically estimated for the population statistics. Jewsbury (2019) derived analytic equations for variance estimation of population statistics such as averages, percentiles, and standard deviations. He suggested that the resampling methods might be more tractable in practice to cover a wide range of statistics. In this paper, we propose a variance estimation method that incorporates the uncertainty of the linking function into the sampling and latency variance estimates. The proposed method can be used to estimate variances for both linear and nonlinear statistics. Further, the method can be used when the two samples used in linking are either dependent or independent from each other.

In Sect. 2, we review the variance estimation approach currently used in LSAs. In Sect. 3, we introduce the new variance estimation method, which is an extension and modification of the existing method. We illustrate the method with a dataset from NAEP in Sect. 4. The conclusion follows in Sect. 5.

2 Variance Estimation in Large-Scale Survey Assessments

For complex survey data, analytical variance estimators for nonlinear statistics are difficult to develop, and some do not have a closed form. For LSAs, one common practice is to use the jackknife repeated replication (JRR) with replicate weights to estimate sampling variance. Several studies (Hansen et al., 1985; Kovar et al., 1988) have shown that JRR provides reasonable variance estimates for both linear and nonlinear statistics. Briefly, a total of H strata are formulated, and each replicate is created by excluding a random set of data in a stratum while keeping the remaining subset from that stratum and all the data in the other $H - 1$ strata. The replicate weights are then calculated for each of the H replicates which reflect the complex sample design. Those replicate weights also help protect the survey participants' information because the more detailed sampling information, such as stratification, primary sampling units (PSUs), clusters, etc., are not needed with the availability of the replicate weights. Details are provided in the next section. Applications include NAEP and TIMSS.

Using NAEP as an example, we now review how the sampling and latency variances are estimated. Let $\boldsymbol{W}_{\mathbf{orig}}$ represent the original sampling weights for the full sample, and $\boldsymbol{W}_{\boldsymbol{j}}$ represent the jth set of jackknife replicate weights, $j = 1, 2, \ldots, N_r$, respectively, for a total of N_r sets of replicate weights. Further, let $\boldsymbol{v}_{\boldsymbol{i}}$ denote the ith set of plausible values which is on an arbitrary IRT scale T, $i = 1, 2, \ldots, M$. Then the population statistic on scale T, denoted as $\hat{t}$ (e.g., population average score), can be calculated as

$$\hat{t} = \frac{\sum_{i=1}^{M} \hat{t}_i}{M} \tag{1}$$

where $\hat{t}_i$ is calculated using $\boldsymbol{v}_{\boldsymbol{i}}$ with weight $\boldsymbol{W}_{\mathbf{orig}}$. The sampling variance of $\hat{t}$ is calculated as $\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_r}\left(\hat{t}_{ij} - \hat{t}_i\right)^2$ where $\hat{t}_{ij}$ denotes the statistic calculated using $\boldsymbol{v}_{\boldsymbol{i}}$ with replicate weight $\boldsymbol{W}_{\boldsymbol{j}}$.

In practice, the sampling variance is often approximated based only on one set of plausible values to reduce computational burden. For example, using the first set of plausible values, the sampling variance can be estimated as

$$\widehat{\mathrm{Var}}_{\mathrm{samp}}\left(\hat{t}\right) = \sum_{j=1}^{N_r}\left(\hat{t}_{1j} - \hat{t}_1\right)^2 \tag{2}$$

Based on the work of Rubin (2004), the latency variance of $\hat{t}$ is estimated as follows:

$$\widehat{\mathrm{Var}}_{\mathrm{lat}}\left(\hat{t}\right) = \left(1 + \frac{1}{M}\right)\frac{\sum_{i=1}^{M}\left(\hat{t}_i - \hat{t}\right)^2}{M - 1}. \tag{3}$$

The total variance for the statistic $\hat{t}$ is the sum of sampling and latency variances:

$$\widehat{\text{Var}}_{\text{total}}\left(\hat{t}\right) = \widehat{\text{Var}}_{\text{samp}}\left(\hat{t}\right) + \widehat{\text{Var}}_{\text{lat}}\left(\hat{t}\right) \tag{4}$$

3 Variance Estimation to Incorporate Uncertainty in Random-Groups Linking

As mentioned in Sect. 1, we consider that a target sample of test-takers is administered assessment T, while a source sample is administered assessment S. Assessment T results are on latent scale T, while assessment S results are on latent scale S. The objective is to apply a linear function to link assessment S results from scale S to scale T by aligning the mean and standard deviation (SD) of the sample taking assessment S to those of assessment T. For example, during the NAEP transition from paper-based assessment (PBA) to digitally based assessment (DBA), the sample who took the PBA is the target sample, and the sample who took the newly implemented DBA is the source sample. The linking function is then derived to link the DBA results to the latent scale for PBA, so that the DBA and PBA results can be compared.

For the source sample statistics that are linked to scale T, we propose a new resampling approach for variance estimation. Under the method, the variance consists of the sampling and measurement variance components, each taking into consideration the random-groups linking. We first discuss the JRR method for the estimation of the sampling variance that involves resampling both the target and source samples simultaneously and then the estimation of the latency variance. The proposed method is an extension of the method discussed in Sect. 2. The method works when the two samples are dependent or independent.

To be more specific, let:

- $\boldsymbol{x_i}$ represent the ith set of plausible values for the target sample on scale T,
- $\boldsymbol{\theta_i}$ represent the ith set of plausible values for the source sample on scale S,
- $\boldsymbol{y_i}$ represent the ith set of plausible values for the source sample that has been transformed to scale T, $i = 1, 2, \ldots,\ M$.

Further, let $\overline{\theta}_S$ and $\widehat{\sigma}_S$ denote the mean and SD of the source sample plausible values on scale S, weighted by $\boldsymbol{W}_{\mathbf{orig}}$, the original student sampling weights of the source sample. Similarly, let $\overline{X}_T$ and $\widehat{\sigma}_T$ denote the mean and SD of the target sample plausible values on scale T, weighted by $\boldsymbol{W'}_{\mathbf{orig}}$, the original student sampling weights of the target sample.

The coefficients $\hat{a}$ and $\hat{b}$ of the linear linking function are calculated as

$$\hat{a} = \frac{\widehat{\sigma}_T}{\widehat{\sigma}_S} \tag{5}$$

and

$$\hat{b} = \overline{X}_T - \hat{a}\overline{\theta}_S \tag{6}$$

Apply $\left(\hat{a}, \hat{b}\right)$ to transform $\boldsymbol{\theta_i}$ from scale S onto scale T:

$$\boldsymbol{y_i} = \hat{a}\boldsymbol{\theta_i} + \hat{b}, i = 1, 2, \ldots, M. \tag{7}$$

Last, we calculate the statistic $\hat{t}$ for the source sample on scale T using Eq. 1, with $\hat{t}_i$ being estimated using $\boldsymbol{y_i}$ with $\boldsymbol{W_{\mathrm{orig}}}$, for $i = 1, 2, \ldots, M$.

In the text below, we describe the procedure in estimating the variance of the statistic $\hat{t}$.

3.1 Estimation of Sampling Variance

In this section, we describe the procedure used in estimating the sampling variance of the source sample statistic $\hat{t}$ as defined in Eq. 1, which is linked to scale T through random-groups linking. We further introduce the following notations $\boldsymbol{W_j}, \boldsymbol{W'_j}$, which represent the jth set of jackknife replicate weights of the source and target samples, respectively, $j = 1, 2, \ldots, N_r$. In the random-groups linking design, it is common that the two samples to be linked have the same number of replicate weights. Therefore, in our method, we assume the source and target samples have the same number of replicate weights (denoted as N_r here). To reduce the computational intensity, we use only the first set of plausible values from both samples for the calculation.

Using the jth pair of replicate weights $\left(\boldsymbol{W_j}, \boldsymbol{W'_j}\right), j = 1, 2, \ldots, N_r$, we conduct the following steps of calculation:

1. Compute $\overline{\theta}_{S_j}$ and $\widehat{\sigma}_{S_j}$, the mean and SD of the first set of plausible values for the source sample on scale S, weighted by $\boldsymbol{W_j}$, as well as $\overline{X}_{T_j}$ and $\widehat{\sigma}_{T_j}$, the mean and SD of the first set of plausible values for the target sample on scale T, weighted by $\boldsymbol{W'_j}$;
2. Calculate the coefficients of the linear linking function $\left(\hat{a}_j, \hat{b}_j\right)$ based on Eqs. 5 and 6, with $\overline{\theta}_{S_j}, \widehat{\sigma}_{S_j}, \overline{X}_{T_j}$, and $\widehat{\sigma}_{T_j}$;
3. Apply $\left(\hat{a}_j, \hat{b}_j\right)$ to transform $\boldsymbol{\theta}_1$ of the source sample from scale S onto scale T of the target sample, i.e. $\boldsymbol{y_1^j} = \hat{a}_j\boldsymbol{\theta}_1 + \hat{b}_j$, where $\boldsymbol{y_1^j}$ is the transformed plausible values for the source sample, $j = 1, 2, \ldots, N_r$;
4. Calculate $\hat{t}'_{1j}$, using $\boldsymbol{y_1^j}$ with replicate weight $\boldsymbol{W_j}$, $j = 1, 2, \ldots, N_r$.

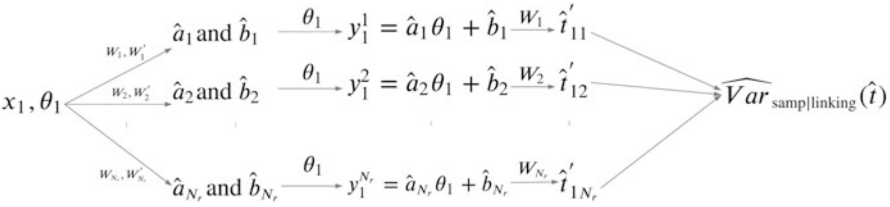

x_1, θ_1
W_1, W_1'
W_2, W_2'
W_{N_r}, W_{N_r}'
$\hat{a}_1$ and $\hat{b}_1$
$\hat{a}_2$ and $\hat{b}_2$
$\hat{a}_{N_r}$ and $\hat{b}_{N_r}$
θ_1
θ_1
θ_1
$y_1^1 = \hat{a}_1\theta_1 + \hat{b}_1$
$y_1^2 = \hat{a}_2\theta_1 + \hat{b}_2$
$y_1^{N_r} = \hat{a}_{N_r}\theta_1 + \hat{b}_{N_r}$
W_1
W_2
W_{N_r}
$\hat{t}_{11}'$
$\hat{t}_{12}'$
$\hat{t}_{1N_r}'$
$\widehat{Var}_{\text{samp|linking}}(\hat{t})$

Fig. 1 The calculation process of sampling variance estimation for the source sample

The sampling variance of statistic $\hat{t}$ can then be approximated as

$$\widehat{\text{Var}}_{\text{samp|linking}}\left(\hat{t}\right) = \sum_{j=1}^{N_r}\left(\hat{t}_{1j}' - \bar{\hat{t}}_1\right)^2, \tag{8}$$

where

$$\bar{\hat{t}}_1 = \frac{1}{N_r}\sum_{j=1}^{N_r}\hat{t}_{1j}' \tag{9}$$

Figure 1 illustrates the calculation process of $\widehat{\text{Var}}_{\text{samp|linking}}\left(\hat{t}\right)$ in Eq. 8. Alternatively, one can approximate the sampling variance of statistic $\hat{t}$ as

$$\widehat{\text{Var}}_{\text{samp|linking}}'\left(\hat{t}\right) = \sum_{j=1}^{N_r}\left(\hat{t}_{1j}' - \hat{t}_1'\right)^2, \tag{10}$$

where $\hat{t}_1'$ is calculated by using the original weights $\boldsymbol{W}_{\mathbf{orig}}$ and the first set of plausible values that are linked to scale T. The scale transformation follows steps 1-3 described above while using the original student weights $\left(\boldsymbol{W}_{\mathbf{orig}}, \boldsymbol{W}'_{\mathbf{orig}}\right)$.

We point out that when calculating $(\hat{a}_j, \hat{b}_j)$, $j = 1, 2, \ldots, N_r$, we pair the replicate weights once and in their corresponding sequential order (i.e., pairing the jth replicate weights from both the source and target samples). For the source and target samples that are dependent, pairing the replicate weights of the two samples in this matter properly accounts for the dependency between the samples. On the other hand, if the source and target samples are independent, then the pairings between the source and target samples can be random. In fact, there are $N_r!$ possible ways to pair the replicate weights between the two samples. In theory, one can calculate the variance estimate for all $N_r!$ sets of pairings and then take an average. In practice, $N_r!$ is usually a very large number. To reduce computational burden, a practical approach is to randomly select a subset from the $N_r!$ sets of pairings. Suppose the N_s ($N_s < N_r!$) sets of random pairings are generated and for ith set of pairing the sampling variance estimate is $\widehat{\text{Var}}_{\text{samp|linking}}^{(i)}\left(\hat{t}\right), i = 1, 2, \ldots, N_s$, which is calculated using Eq. 8. Then the sampling variance is estimated as the average of the N_s estimates:

$$\widehat{\mathrm{Var}}^{*}_{\mathrm{samp|linking}}\left(\hat{t}\right) = \frac{1}{N_s}\sum_{i=1}^{N_s}\widehat{\mathrm{Var}}^{(i)}_{\mathrm{samp|linking}}\left(\hat{t}\right) \tag{11}$$

The choice of the value N_s is a balance between the computation intensity and the stability of the variance estimate.

The above procedure described how to calculate the sampling variance for the source sample statistics only. There are also situations where the statistics are computed based on combining the source and target samples. Next, we show that the procedure can be generalized to estimate the sampling variance for the combined sample as well.

To do that, after getting the transformed plausible values $\boldsymbol{y}_{\mathbf{1}}^{\boldsymbol{j}}$, we concatenate $\boldsymbol{y}_{\mathbf{1}}^{\boldsymbol{j}}$ with $\boldsymbol{x}_{\mathbf{1}}$ of the target sample as the combined set of plausible values, $\boldsymbol{z}_{\mathbf{1}}^{\boldsymbol{j}} = \begin{pmatrix} \boldsymbol{y}_{\mathbf{1}}^{\boldsymbol{j}} \\ \boldsymbol{x}_{\mathbf{1}} \end{pmatrix}$, $j = 1, 2, \ldots, N_r$. Then the statistic of interest based on the combined sample can be calculated using $\boldsymbol{z}_{\mathbf{1}}^{\boldsymbol{j}}$ with weight $\boldsymbol{W}_{\boldsymbol{j}}^{\mathbf{comb}}$, which is the replicate weights for the combined sample. Note that in practice, $\boldsymbol{W}_{\boldsymbol{j}}^{\mathbf{comb}}$ are created specially to the analysis of the combined sample. The rest of the calculation is the same as shown in Eq. 8.

Figure 2 shows the calculation process for statistics of the combined sample. Note that the replicate weights are paired following their corresponding sequential order as (1 to 1), (2 to 2), etc. As discussed earlier, when the source and target samples are independent of each other, the pairing of plausible values from the two samples can be random.

3.2 Estimation of Latency Variance

We now discuss the procedure of calculating the latency variance of the source sample statistics $\hat{t}$ as defined in Eq. 1, which is linked to scale T through the random-groups linking.

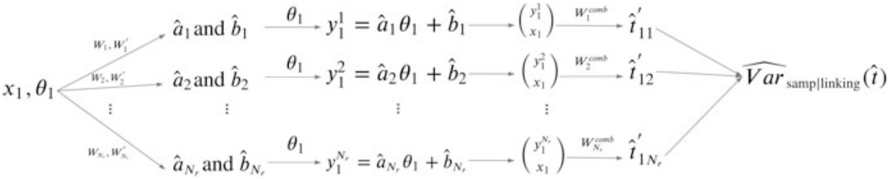

x_1, θ_1
W_1, W_1'
W_2, W_2'
W_{N_r}, W_{N_r}'
$\hat{a}_1$ and $\hat{b}_1$
$\hat{a}_2$ and $\hat{b}_2$
$\hat{a}_{N_r}$ and $\hat{b}_{N_r}$
θ_1
$y_1^1 = \hat{a}_1\theta_1 + \hat{b}_1$
$y_1^2 = \hat{a}_2\theta_1 + \hat{b}_2$
$y_1^{N_r} = \hat{a}_{N_r}\theta_1 + \hat{b}_{N_r}$
$\begin{pmatrix} y_1^1 \\ x_1 \end{pmatrix}$
$\begin{pmatrix} y_1^2 \\ x_1 \end{pmatrix}$
$\begin{pmatrix} y_1^{N_r} \\ x_1 \end{pmatrix}$
W_1^{comb}
W_2^{comb}
$W_{N_r}^{comb}$
$\hat{t}'_{11}$
$\hat{t}'_{12}$
$\hat{t}'_{1N_r}$
$\widehat{Var}_{\mathrm{samp|linking}}(\hat{t})$

Fig. 2 The process of sampling variance estimation for the combined sample

Using the M sets of plausible values from the source sample and the target sample, we conduct the following steps:

1. Calculate $\overline{\theta}_{S_i}$ and $\widehat{\sigma}_{S_i}$, the mean and SD of the scale scores using the ith set of plausible value in the source sample on scale S with $\boldsymbol{W}_{\mathbf{orig}}$;
2. Calculate $\overline{X}_{T_i}$ and $\widehat{\sigma}_{T_i}$, the mean and SD of the scale scores using the ith set of plausible value in the target sample on scale T with $\boldsymbol{W}'_{\mathbf{orig}}$;
3. Calculate the transformation coefficients $\left(\hat{a}_i, \hat{b}_i\right)$ based on Eqs. 5 and 6 with $\left(\overline{\theta}_{S_i}, \widehat{\sigma}_{S_i}\right)$ and $\left(\overline{X}_{T_i}, \widehat{\sigma}_{T_i}\right)$, $i = 1, 2, \ldots,\ M$;
4. Apply $\left(\hat{a}_i, \hat{b}_i\right)$ to transform $\boldsymbol{\theta}_{\boldsymbol{i}}$ from scale S onto scale T, i.e., $\boldsymbol{y}_{\boldsymbol{i}}^{*}{=}\hat{a}_i\boldsymbol{\theta}_{\boldsymbol{i}}{+}\hat{b}_i$;
5. Calculate the statistic of interest $\hat{t}_i^{*}$, using $\boldsymbol{y}_{\boldsymbol{i}}^{*}$ with $\boldsymbol{W}_{\mathbf{orig}}$, $i = 1, 2, \ldots, M$;
6. Calculate the latency variance of the source sample statistics.

$$\widehat{\mathrm{Var}}_{\mathrm{lat|linking}}\left(\hat{t}\right) = \left(1 + \frac{1}{M}\right)\frac{\sum_{i=1}^{M}\left(\hat{t}_i^{*} - \hat{t}^{*}\right)^2}{M - 1} \tag{12}$$

where

$$\hat{t}^{*} = \frac{\sum_{i=1}^{M}\hat{t}_i^{*}}{M} \tag{13}$$

The process of calculating the latency variance is illustrated in Fig. 3.

In the above procedure, the plausible values from the two samples are paired when calculating the linking function coefficients $\left(\hat{a}_i, \hat{b}_i\right)$, $i = 1, 2, \ldots, M$. The plausible values for the source and target samples are multiple imputations that were drawn independently using two latent regression models and therefore are independent regardless whether the two samples are dependent or independent of each other.

There are a total of $M!$ possible sets of pairings of the plausible values from the two samples, with M sets of plausible values for each sample. In practice, we can choose a subset of random pairings to reduce computation intensity. Let's assume

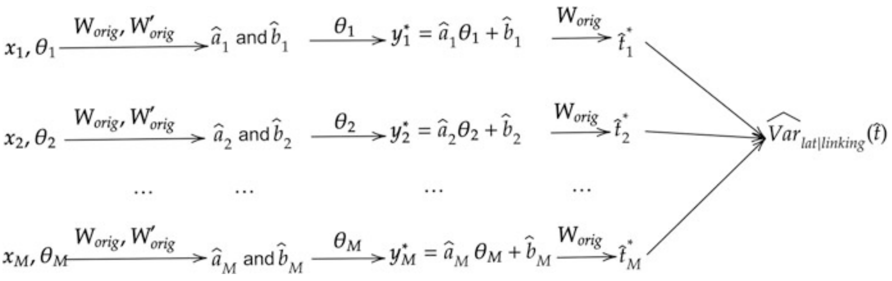
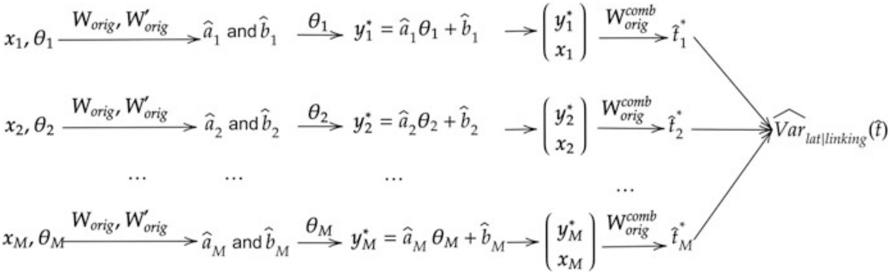

$x_1, \theta_1 \xrightarrow{W_{orig}, W'_{orig}} \hat{a}_1$ and $\hat{b}_1 \xrightarrow{\theta_1} y_1^* = \hat{a}_1\theta_1 + \hat{b}_1 \xrightarrow{W_{orig}} \hat{t}_1^*$
$x_2, \theta_2 \xrightarrow{W_{orig}, W'_{orig}} \hat{a}_2$ and $\hat{b}_2 \xrightarrow{\theta_2} y_2^* = \hat{a}_2\theta_2 + \hat{b}_2 \xrightarrow{W_{orig}} \hat{t}_2^*$
… … … …
$x_M, \theta_M \xrightarrow{W_{orig}, W'_{orig}} \hat{a}_M$ and $\hat{b}_M \xrightarrow{\theta_M} y_M^* = \hat{a}_M\theta_M + \hat{b}_M \xrightarrow{W_{orig}} \hat{t}_M^*$
$\widehat{Var}_{lat|linking}(\hat{t})$

Fig. 3 The calculation process of latency variance estimation for the source sample



x_1, θ_1
W_{orig}, W'_{orig}
$\hat{a}_1$ and $\hat{b}_1$
θ_1
$y_1^* = \hat{a}_1\theta_1 + \hat{b}_1$
$\begin{pmatrix} y_1^* \\ x_1 \end{pmatrix}$
W_{orig}^{comb}
$\hat{t}_1^*$
x_2, θ_2
W_{orig}, W'_{orig}
$\hat{a}_2$ and $\hat{b}_2$
θ_2
$y_2^* = \hat{a}_2\theta_2 + \hat{b}_2$
$\begin{pmatrix} y_2^* \\ x_2 \end{pmatrix}$
W_{orig}^{comb}
$\hat{t}_2^*$
$\widehat{Var}_{lat|linking}(\hat{t})$
…
x_M, θ_M
W_{orig}, W'_{orig}
$\hat{a}_M$ and $\hat{b}_M$
θ_M
$y_M^* = \hat{a}_M\theta_M + \hat{b}_M$
$\begin{pmatrix} y_M^* \\ x_M \end{pmatrix}$
W_{orig}^{comb}
$\hat{t}_M^*$

Fig. 4 The process of latency variance estimation for the combined sample

N_s ($N_s < M!$) sets of random pairings are generated, and for ith set of pairing, the latency variance estimate is $\widehat{\mathrm{Var}}^{(i)}_{\mathrm{lat|linking}}\left(\hat{t}\right)$, $i = 1, 2, \ldots, N_s$ which is calculated using Eq. 12. Then the latency variance can be estimated as the average of the N_s estimates:

$$\widehat{\mathrm{Var}}^{*}_{\mathrm{lat|linking}}\left(\hat{t}\right) = \frac{1}{N_s}\sum_{i=1}^{N_s}\widehat{\mathrm{Var}}^{(i)}_{\mathrm{lat|linking}}\left(\hat{t}\right) \tag{14}$$

The choice of N_s is a balance between the computation capacity and reducing variability of the variance estimation.

The above procedure to calculate the latency variance is for the source sample statistics only. Similar to the estimation of sampling variance, we can extend the method to calculate latency variance for the statistics based on the combined source and target sample. To do that, after transforming the source sample plausible values from $\boldsymbol{\theta_i}$ to $\boldsymbol{y_i^*}$ using $\left(\hat{a}_i, \hat{b}_i\right)$, $i = 1, 2, \ldots, M$, we concatenate $\boldsymbol{y_i^*}$ with $\boldsymbol{x_i}$ as $\boldsymbol{z_i^*} = \begin{pmatrix} \boldsymbol{y_i^*} \\ \boldsymbol{x_i} \end{pmatrix}$. Then the statistic of interest based on the combined sample can be calculated using $\boldsymbol{z_i^*}$ with weight $\boldsymbol{W^{\mathbf{comb}}_{\mathbf{orig}}}$, which is the original weights for the combined sample. The rest of the calculation is the same as shown in Eq. 12.

Figure 4 displays this calculation procedure for the combined sample, with the pairing of the plausible values following a (1 to 1), (2 to 2), etc. fashion.

Finally, the total variance of the statistic $\hat{t}$ is the sum of the sampling and latency variances. When the source and target samples are dependent, the total variance is estimated as

$$\widehat{\mathrm{Var}}_{\mathrm{total|linking}}\left(\hat{t}\right) = \widehat{\mathrm{Var}}_{\mathrm{samp|linking}}\left(\hat{t}\right) + \widehat{\mathrm{Var}}^{*}_{\mathrm{lat|linking}}\left(\hat{t}\right) \tag{15}$$

3.3 Properties of the Proposed Variance Estimation Method

In this study, we consider linear linking in a random-groups design. That is, a linear function is applied to align the mean and SD of the source sample score distribution to the mean and SD of the target sample score distribution. Next, we show that regardless of the sample size and other features of the source sample, its mean and SD are fixed to be the same as those of the target sample as the expected result of the linking. Recall the linear function has the following form:

$$\boldsymbol{y}_i = \hat{a}\boldsymbol{\theta}_i + \hat{b},\ i = 1, 2, \ldots, M. \tag{16}$$

where $\hat{a} = \frac{\widehat{\sigma}_T}{\widehat{\sigma}_S}$ and $\hat{b} = \overline{X}_T - \hat{a}\overline{\theta}_S$, as defined in Eqs. 5 and 6.

Let $\overline{Y}_S$ and $\widehat{\sigma}_S^Y$ denote the mean and SD of the transformed scores of the source sample, then given how the $\hat{a}$ and $\hat{b}$ are constructed, we have

$$\overline{Y}_S = \hat{a}\overline{\theta}_S + \hat{b} = \overline{X}_T \tag{17}$$

and

$$\widehat{\sigma}_S^Y = \hat{a} * \widehat{\sigma}_S = \widehat{\sigma}_T \tag{18}$$

The above property is true when the weights used in the calculation are the original weights or the replicate weights. Therefore, for the estimation of sampling variance discussed in Sect. 3.1, $\hat{t}^{\prime}_{1j}$, $j = 1, 2, \ldots, N_r$, for the source sample are the same as the corresponding statistics of the target sample. According to Eq. 8, the sampling variances of the overall mean and SD for the source sample are the same as those for the target sample, provided the point estimates used in the formula are also the same between the two samples.

Similarly, for the latency variance estimation, $\hat{t}_i^*$, $i = 1, 2, \ldots, M$, of the source sample are the same as the corresponding statistics of the target sample. Following the same logic as for the sampling variance, the latency variances of the overall mean and SD for the source sample are the same as those for the target sample.

Now for the combined source and target sample, we have plausible values $\boldsymbol{z}_i = \begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{x}_i \end{pmatrix}$, $i = 1, 2, \ldots,\ M$. Then the mean of the combined sample

$$\overline{Z} = \frac{\overline{Y}_S n_S + \overline{X}_T n_T}{n_S + n_T} = \frac{\overline{X}_T\,(n_S + n_T)}{n_S + n_T} = \overline{X}_T \tag{19}$$

where n_S and n_T are the weighted sample size of the source and target samples. Similarly, for the SD of the combined sample,

$$\begin{aligned}\widehat{\sigma}_Z &= \sqrt{\frac{\left(\widehat{\sigma}_S^Y\right)^2 n_S + (\widehat{\sigma}_T)^2 n_T}{n_S + n_T}}\\ &= \sqrt{\frac{(\widehat{\sigma}_T)^2 n_S + (\widehat{\sigma}_T)^2 n_T}{n_S + n_T}}\\ &= \widehat{\sigma}_T \sqrt{\frac{n_S + n_T}{n_S + n_T}}\\ &= \widehat{\sigma}_T\end{aligned} \tag{20}$$

That is, the combined sample, after the scale linking, has the same mean and SD as those for the target sample. Moreover, the variances of the overall mean and SD for the combined sample are also the same as those for the target sample. The argument is the same as for the source sample.

In addition, we point out that the variance estimation considering random-groups linking does not necessarily result in a larger estimated value than those procedures in which the uncertainty due to linking is ignored. For example, as described above, the variances of the mean estimates are the same between the source and target samples after linking. The property holds even when the source sample has much smaller sample size than the target sample. For subgroups, as will be shown in the empirical data below, it is possible to obtain a variance estimate that is smaller when considering the uncertainty due to linking.

4 Applications

4.1 Empirical Results

In this section, we use the data from NAEP to illustrate our proposed method. A study with the random-groups design and linear linking was implemented to link the scores from DBA to PBA. The study involved administering the DBA and PBA to two samples of students, respectively, namely, the DBA sample and the bridge PBA sample. A total of 13,400 students were selected in the study to take either the DBA or PBA. The DBA and bridge PBA samples are dependent with comparable sizes.

Table 1 displays the comparison between the DBA and bridge PBA samples. We can see the demographic distributions between the two samples are comparable.

The bridge PBA and DBA samples were analyzed separately using the IRT latent regression models, and the results were expressed on two separate IRT scales. Following the NAEP operational convention, a total of 20 plausible values were imputed for each student in the 2 samples. In addition, for each sample, the original weight and 62 replicate weights were provided for each student. The results for the bridge PBA sample were estimated on the existing NAEP trend scale, where the

Table 1 Weighted percentage of students by subgroup between the bridge PBA and DBA samples: a NAEP dataset

		Bridge PBA	DBA
Gender	Male	51%	51%
	Female	49%	49%
Race/ethnicity	White	49%	49%
	Black	15%	14%
	Hispanic	27%	27%
	Others	10%	10%
School type	Public	91%	93%
	Non-public	9%	7%

Table 2 Sample sizes, standard errors of estimates of means with and without linking error: the combined DBA/PBA sample

Group	*N*	SE	SE*	SE Ratio
All students	13,400	0.87	0.69	1.26
Male	6900	0.95	0.78	1.22
Female	6500	0.92	0.78	1.18
White	5900	1.01	0.93	1.09
Black	2100	1.30	1.18	1.10
Hispanic	3900	1.20	1.01	1.19
Asian	700	1.83	1.77	1.03
American-Indian/Alaska	200	13.89	14.07	0.99
Northeast	2000	1.85	1.78	1.04
Midwest	2300	1.99	2.01	0.99
South	5400	1.10	0.95	1.16
West	3700	1.19	1.09	1.09

mean and SD of the scale were set operationally to be 150 and 35. For the DBA sample, the results were generated on an arbitrary IRT scale with mean 0 and SD 1. The plausible values of these separate analyses were then used to develop a linear linking function (Eqs. 5 and 6) which allowed for the expression of the DBA results on the bridge PBA scale. Since the DBA and bridge PBA samples are dependent, when calculating the sampling variance, we applied the (1 to 1), (2 to 2), ..., (62 to 62) fashion of pairing the replicate weights between the DBA and bridge PBA samples.

Table 2 presents the standard errors of the mean estimates for the combined DBA/PBA sample, using the proposed new method (Eqs. 8, 14, and 15). For comparison purpose, we also include the usual NAEP variance estimates which do not contain linking variance. Column SE contains the standard errors calculated using our proposed methods, and column SE* contains the standard errors without accounting for random-groups linking. For the race/ethnicity variable, the students in the Native Hawaiian/Other Pacific Islander and the Two or More Races categories are not listed in the table.

We can see from Table 2 that for the overall mean estimate, SE $\left(\overline{X}_{\text{all\_student}}\right) = 0.87$, SE* $\left(\overline{X}_{\text{all\_student}}\right) = 0.69$, with a ratio of 1.26. The change in standard errors for subpopulation means, with and without accounting for random-groups linking, is less than the value for the overall population. For the displayed subgroups, the ratios range from about 0.99–1.22.

Furthermore, we observe that the ratios in standard error vary for different subgroups, but have little relationship with the sample size of the group in question. For example, the male and female students are about 50% of the overall population; the ratios in standard error with and without accounting for linking errors are 1.22 and 1.18, respectively. On the other hand, White subgroup is about half the overall population, but the ratio in standard error is 1.09. The linear linking functions were derived based on the overall population, not the subgroups whose results were being transformed by the function. As a result, the ratios in standard error are expected to vary across subgroups. Analytical results of the effect on subgroup standard errors are found in Jewsbury (2019).

4.2 Further Considerations on Latency Variance Estimation

As mentioned in Sect. 4.1, in NAEP, there are 20 plausible values for each student in the source and target samples. When calculating the latency variance, the pairing of the 20 sets of plausible values between source and target sample can be random given the source and target sample plausible values are independent. For example, one way of pairing the plausible values is to follow their corresponding sequential order (i.e., pairing the ith set of plausible values from both the source and target samples). As another example, one could pair the plausible values from the source and target samples following the sequence as (1 to 2), (2 to 3), ..., (20 to 1). In theory, there are 20! possible ways to pair the plausible values between the source and target samples.

We point out that while the latency variance can be estimated based on a single set of pairings of the source and target sample plausible values, averaging the latency variance estimates over multiple sets of pairings, N_s ($N_s < 20!$), is expected to improve the stability of the latent variance estimates. Using the NAEP data, we conducted a simulation study to examine how the latency variance estimates vary with different values for N_s.

In the simulation study, we considered five conditions, with N_s being 1, 5, 10, 25, and 50. For each of the five conditions, we calculated the latency variance 100 times, using the method discussed in Sect. 3. Figure 5 shows the box-plot of the standard errors due to latency for the male students average score for the 100 replications. We can see that as N_s increases, the variation of the standard error estimates decreases. The most noticeable variability reduction is from 1 to 5 random pairings. When N_s equals to 5, the difference between the maximal and minimal standard error estimates among the 100 replications is less than 0.04. In

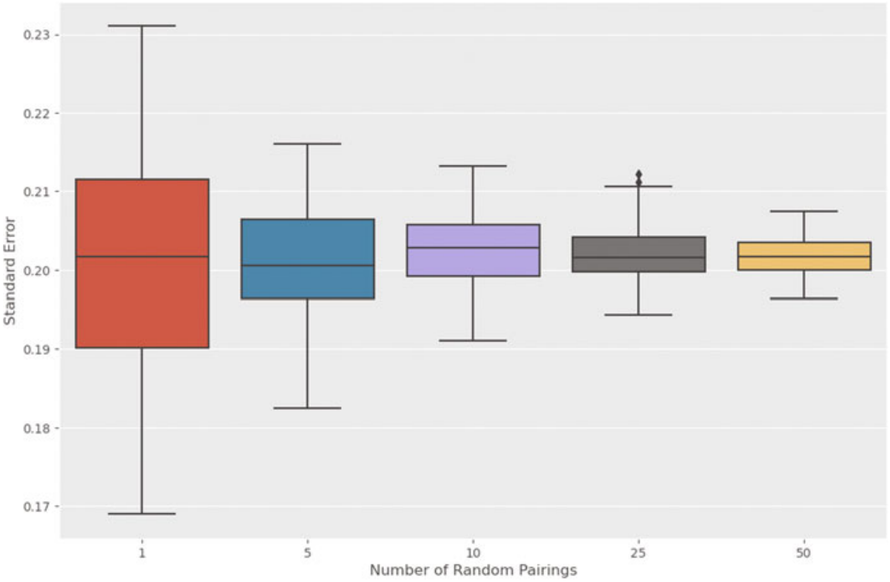


Fig. 5 Box-plot of standard errors due to latency for the male students' average score

this application, we estimated latency error based on five sets of random pairings, considering the latency error estimation is acceptably stable given the magnitudes of subgroup standard errors (as listed in Table 2) and that the latency standard error estimates are typically around 0.2 to 0.4. In practice, similar simulation studies can be helpful to specify the number of random pairings.

5 Conclusion

With complex survey data, it is desirable to have resampling methods that utilize the existing estimation system for variance estimation. For the large-scale survey assessments, the variance of the population statistics is estimated as the sum of two components, the sampling and latency variances. In this paper, we proposed a resampling method for variance estimation when random-groups linking design is applied, incorporating linking error into both the sampling and latency variance estimates. The method is applicable to both linear or nonlinear statistics.

We proposed the estimation procedure in the context of linear linking function. However, the approach applies to both parametric and non-parametric linking functions. Further, it can be applied when the linking sample are dependent or independent.

Acknowledgments The research reported here was supported under the National Assessment of Educational Progress (ED-IES-13-C-0017) to Educational Testing Service (ETS) as administered by the Institute of Education Sciences, US Department of Education. The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education. The authors wish to thank Paul Jewsbury, Daniel F McCaffrey, Mei-Jang Lin, John R Donoghue, and Xueli Xu at Educational Testing Service for contributing to this work. We would also like to thank Keith Rust at Westat and the Design and Analysis Committee of the National Assessment of Educational Progress for useful discussions and advice. We are grateful to the anonymous reviewers for the useful suggestions.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-scale Assessments in Education, 5*(1), 1–16.

Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In *Linking and aligning scores and scales* (pp. 135–159). Springer.

Hansen, M. H., Dalenius, T., & Tepping, B. J. (1985). The development of sample surveys of finite populations. In A. C. Atkinson & S.E. Fienberg (Eds.), *A celebration of statistics* (pp. 327–354). Springer.

Jewsbury, P. (2019). Error variance in common population linking bridge studies. (Research Report No. RR-19-42). Princeton, NJ: Educational Testing Service.

Jewsbury, P., Finnegan, R., Xi, N., Jia, Y., Rust, K., & Burg, S. (2020). 2017 NAEP transition to digitally based assessments in mathematics and reading at grades 4 and 8: Mode evaluation study. https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf

Jia, Y., Phillips, G., Wise, L., Rahman, T., Xu, X., Wiley, C., & Diaz, T. (2011). NAEP-TIMSS linking study: Technical report on the linking methodologies and their evaluations (NCES 2014-461).

Johnson, E., Cohen, J., Chen, W., Jiang, T., & Zhang, Y. (2005). 2000 NAEP–1999 TIMSS linking report.

Johnson, E. G. (1998). *Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): A technical report*. US Department of Education, Office of Educational Research and Improvement.

Johnson, E. G., & Rust, K. F. (1992). Chapter 5: Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*(2), 175–190.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society: Series B (Methodological), 36*(1), 1–22.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.

Kovar, J., Rao, J., & Wu, C. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics, 16*(S1), 25–45.

Krewski, D., & Rao, J. N. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics* 1010–1019.

Mazzeo, J. (2018). Large-scale group-score assessments. In *Handbook of item response theory* (pp. 297–311). Chapman and Hall/CRC.

Mazzeo, J., Liu, B., Donoghue, J., & Xu, X. (in press). Approximate standard errors for NAEP results that incorporate linking error under the random group design.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359–381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*(392), 993–997.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Rao, J., & Wu, C. J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association, 80*(391), 620–630.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. Wiley.

Shao, J. (1996). Invited discussion paper resampling methods in sample surveys. *Statistics, 27*(3–4), 203–237.

Valliant, R. (1990). Comparisons of variance estimators in stratified random and systematic sampling. *Journal of Official Statistics, 6*(2), 115–131.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. *Handbook of Statistics, 26*, 1039–1055.

Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed., Vol. 53). Springer.

Item Response Theory and Fisher Information for Small Tests

Bivin Philip Sadler and S. Lynne Stokes

Abstract Item response theory (IRT) is a comprehensive paradigm for modeling test performance on the item level in contrast to the more general test-level assessment of classical test theory (CTT). Given the added flexibility provided by item-level modeling, IRT has become the predominant theory used in high-stakes tests such as the SAT, LSAT, and GRE. IRT not only provides an estimate of the examinee's ability but also describes methods to estimate the variance (in terms of Fisher Information $I = 1/Var(\hat{\theta})$) of the ability estimate. As will be explained and demonstrated in this chapter, however, these methods are asymptotic and are inadequate for smaller tests with 15 or fewer questions (as might be found in a computer adaptive test). In addition to illustrating the difference between the IRT estimate and the true variance of the ability estimate for smaller tests, an alternative method of variance estimation will be provided and demonstrated.

1 Basics

Although IRT provides a powerful model in which to design and assess tests, its fundamentals are simple. For each item, the probability of a correct response is modeled with a logistic curve (Fig. 1a) in which the x-axis represents the ability range from −3 to 3 and the y-axis represents the probability of a correct response. The curve is known as an item characteristic curve (ICC). The two-parameter logistic version of the model (known as 2PL) describes the probability of a correct response as

B. P. Sadler (✉)
Master of Science in Data Science, Southern Methodist University, Dallas, TX, USA
e-mail: bsadler@mail.smu.edu

S. L. Stokes
Department of Statistical Science, Southern Methodist University, Dallas, TX, USA
e-mail: slstokes@smu.edu

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_12

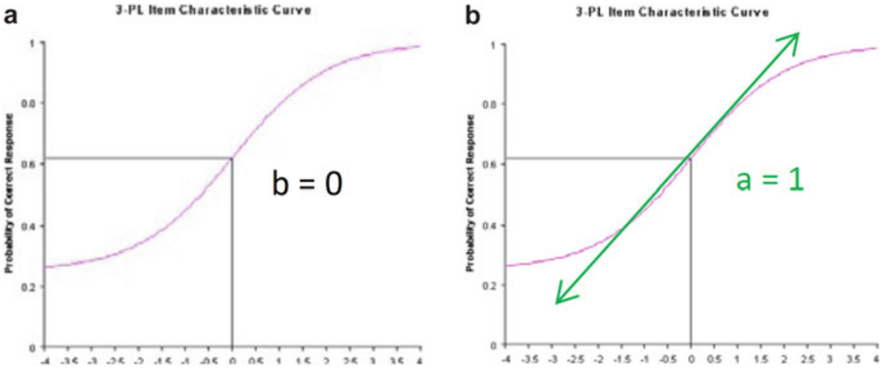


Fig. 1 (**a**) Item characteristic curve (ICC) with difficulty $b = 0$; (**b**) Same ICC showing that the discrimination for the item is $a = 1$

$$p_i(\theta) = \frac{1}{1 + e^{-1.702a_i(\theta - b_i)}}. \tag{1}$$

The parameter b describes the item's difficulty. Specifically, it is the point on the x-axis where the examinee has probability 0.5 to answer the item correctly (Fig. 1a). The parameter a is the discrimination parameter, which represents the slope of the ICC at b. It describes how well the item ascertains the examinee's ability above or below the difficulty of the item (Fig. 1b).

There are other forms of the IRT model for items. Among these are the one-parameter Rasch model, which retains the difficulty parameter but sets $a = 1$. Another version is the three-parameter logistic (known as 3PL) model, which is often used for multiple-choice items, because it includes a guessing parameter. In this chapter, we illustrate our methods with the 2PL IRT model as defined in (1).

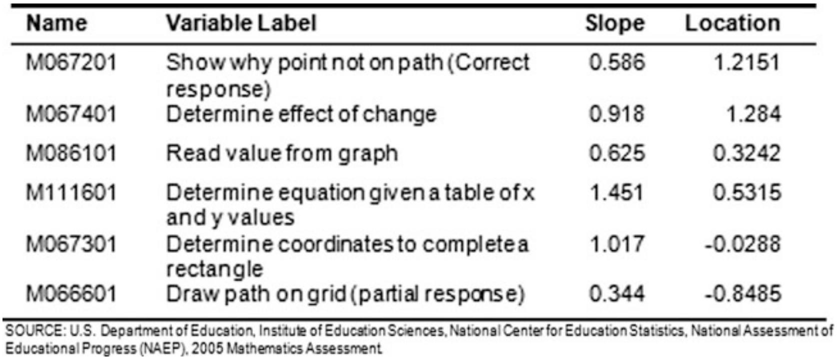
2 Estimation

The IRT model can be used to provide an estimate of the examinee's ability from their responses, when the item parameters are known. If the item parameters are unknown, they can be estimated simultaneously with the ability measures from a sample of examinee responses. For simplicity we assume that the item parameters are known and focus on estimation of ability only.

Maximum likelihood estimation of ability is illustrated with the data from the 2005 National Assessment of Educational Progress (NAEP) Math Assessment. Table 1 displays the slope (a) and location (b) parameters for six actual sample items from the NAEP test (Beaton et al., 2011).

Table 2 shows responses to these items from four fictitious examinees (Beaton et al., 2011). Let z_i denote the indicator of a correct response, i.e.,

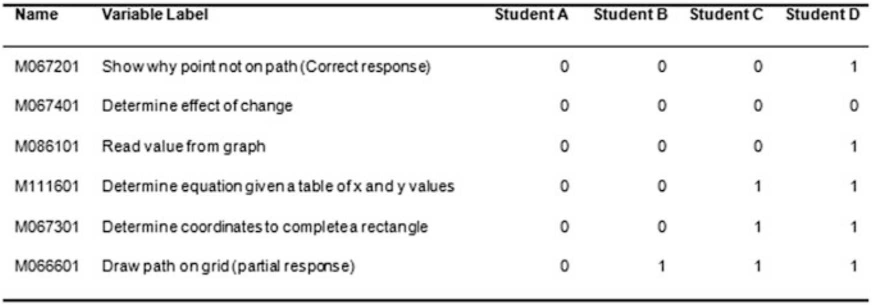
Table 1 Item parameters for the six items referred to in Table 2



Name	Variable Label	Slope	Location
M067201	Show why point not on path (Correct response)	0.586	1.2151
M067401	Determine effect of change	0.918	1.284
M086101	Read value from graph	0.625	0.3242
M111601	Determine equation given a table of x and y values	1.451	0.5315
M067301	Determine coordinates to complete a rectangle	1.017	-0.0288
M066601	Draw path on grid (partial response)	0.344	-0.8485

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Mathematics Assessment

Table 2 Four different students' responses to six different math questions. A correct response is indicated by a "1" and an incorrect response by a "0"



Name	Variable Label	Student A	Student B	Student C	Student D
M067201	Show why point not on path (Correct response)	0	0	0	1
M067401	Determine effect of change	0	0	0	0
M086101	Read value from graph	0	0	0	1
M111601	Determine equation given a table of x and y values	0	0	1	1
M067301	Determine coordinates to complete a rectangle	0	0	1	1
M066601	Draw path on grid (partial response)	0	1	1	1

$$z_i = \begin{cases} 0, & \text{incorrect response to item } i, \\ 1, & \text{correct response to item } i. \end{cases}$$

As an example, Student C answered the first three questions incorrectly and the last three correctly. If the six item responses are independent, the likelihood of Student C's ability given their observed pattern of responses is seen from (1) to be

$$L(\theta|Z) = \prod_{i=1}^{6} \left(\frac{1}{1 + e^{-1.702 a_i (\theta - b_i)}} \right)^{z_i} \left(1 - \frac{1}{1 + e^{-1.702 a_i (\theta - b_i)}} \right)^{1 - z_i} .$$

Student C's likelihood $L(\theta|Z)$ is shown as the bold curve in Fig. 2. The thinner curves show the item characteristic curves of the six items composing the test. Ability is measured on the same scale as the location parameter. On this NAEP test, the range of ability is -3 to 3, with a mean of 0. An iterative Newton-Raphson-type procedure is usually used to maximize this likelihood function to determine the maximum likelihood estimate (MLE) of Student C's ability. Visual inspection

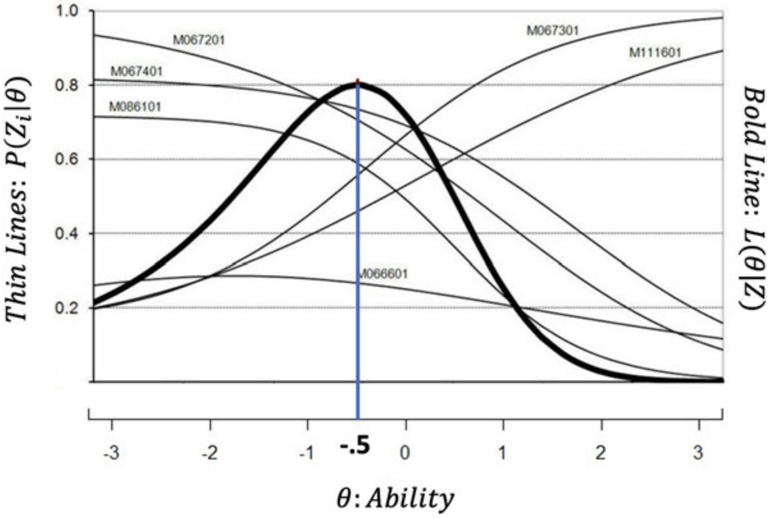


Fig. 2 ICCs and the likelihood (bold) for Student C. The likelihood is calculated by multiplying the student's individual ICCs (Beaton et al., 2011)

shows that Student C's ability would be estimated by maximum likelihood to be about -0.5.

Estimation of ability at the extreme ends of the ability scale is difficult, especially for short tests. Consider Student A in Table 2, who answered all questions incorrectly. His likelihood is shown in Fig. 3 (Beaton et al., 2011). No MLE exists in this case because the likelihood has no maximum. One method for handling estimation for this situation is to assign pre-specified values to examinees who answer no or all questions correctly. This is the method used by the STAAR test in Texas (STAAR, 2004). We will adopt this convention by assigning an ability of -4 to examinees who provide all incorrect responses and an ability of 4 to those who provide all correct responses.

3 Test Information

The test information function (TIF) is defined as the Fisher information of the entire test as a function of ability. One can show that the TIF for the 2PL model, where $p_i(\theta)$ defined in 1, is as defined below:

$$TIF(\theta) = \sum\nolimits_{i=1}^{n} a_i^2 p_i(\theta)(1 - p_i(\theta)). \tag{2}$$

Two examples of TIFs are presented in Fig. 4a and b. These two curves represent TIFs for tests of ten items that measure ability on a scale that is symmetric around

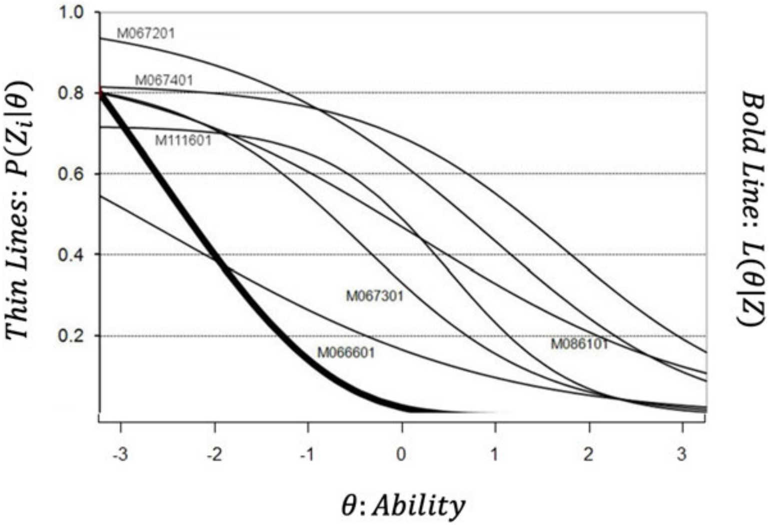


Fig. 3 This plot pictures the ICCs and the likelihood (bold) for Student A. The deficiency of the MLE is exposed in this plot as the student has answered every question incorrectly, and thus the likelihood has no maximum

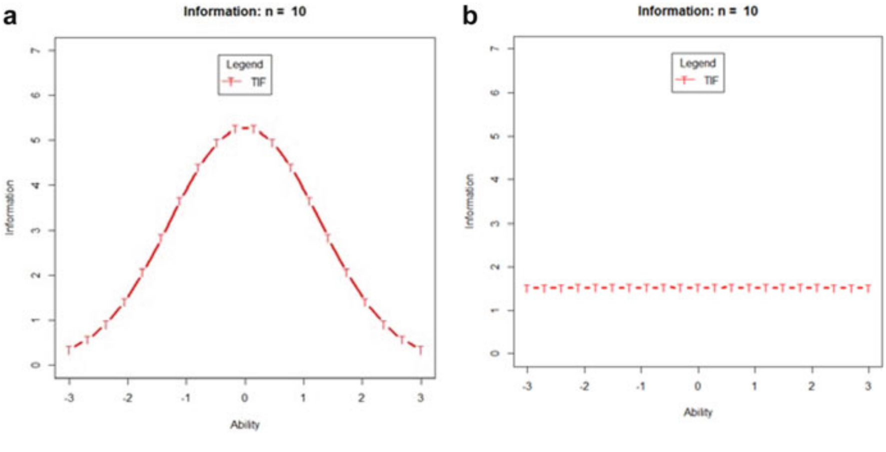


Fig. 4 (**a**) "Peaked" information function; (**b**) Rectangular information function

0, and both will produce some information of examinee ability for those with ability between −3 and +3. However, the tests differ greatly in the shape of their TIFs.

4 Shapes of TIFs

It is common for tests to contain more information about abilities close to the average than at the extremes. The TIF for such a test with ten items[1] is shown in Fig. 4a. It is often desirable that a test maximize information for abilities in the center of the scale, where examinees may be most numerous. This shape is referred to as “peaked." On the other hand, when a population of examinees contains a substantial number at the extremes of the scale, it may be desirable to consider tests with other TIF shapes, such as the “rectangular” one shown in Fig. 4b.

A peaked test information function can be formed through a variety of combinations of items. For instance, a test whose a (discrimination) parameters are similar and whose *b* (difficulty) parameters are grouped near the center will have this shape. On the other hand, a peaked TIF would also result from a test whose *b* parameters are uniformly distributed across the scale and whose a parameters are larger for the items in the center of the range than for those near the tails. Figure 5a displays the discrimination and difficulty parameters of such a test along with its corresponding TIF. Note the increase in item discrimination (a) as the difficulty (b) approaches 0. Figure 5b shows an alternative ten-item test in which the discriminations are nearly constant across the uniformly distributed difficulties which have had a “flattening” effect on the TIF. The tests in Figs. 5a and b will be known as Test 1 and Test 2, respectively, and will be used in examples later in the chapter.

Similar to the peaked TIFs, a rectangular TIF may also be formed through a variety of item parameter combinations. For example, they may have items that have similar a’s and uniformly distributed b’s (Fig. 5c), or they may have more normally distributed b’s, with the items with extreme difficulty having higher a’s

Item	a	b
99	0.79	-2.1
101	0.92	-1.11
14	0.96	-0.61
8	1.02	-0.52
138	1.06	0.15
18	1.68	0.2
161	1.06	0.46
121	1.07	0.49
47	0.97	1.14
155	0.82	2.76

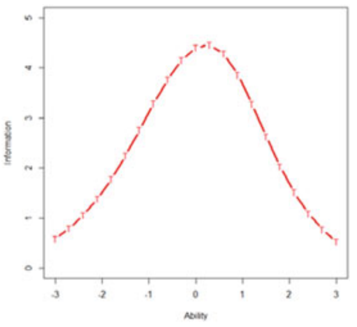


Fig. 5a A ten-item peaked test (Test 1) with uniformly distributed *b* parameters and *a* parameters greater for *b* parameters near 0

[1] These 10 items were real items from the 2004 NAEP Math Exam.

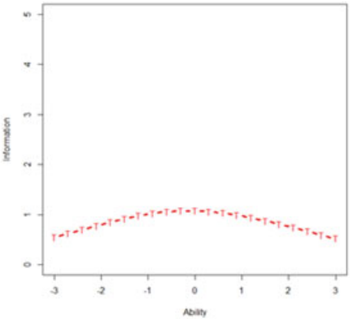
Item	a	b	
82	0.45	-1.73	
147	0.49	-0.72	
17	0.4	-0.42	
32	0.35	-0.259	
117	0.47	-0.25	
118	0.34	0.13	
105	0.41	0.22	
25	0.4	0.87	
136	0.4	1.28	
110	0.46	1.94	

Fig. 5b A ten-item peaked test with uniformly distributed *b* parameters and *a* parameters with less magnitude and nearly uniform across their *b* parameters

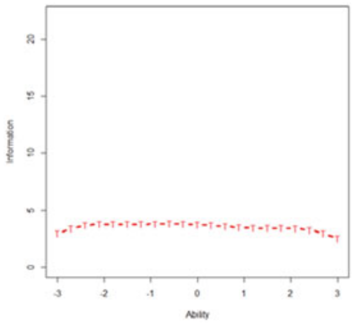
Item	a	b	
85	0.99	-3.1	
92	0.92	-2.11	
15	0.96	-1.61	
76	.902	-0.52	
163	.906	0.15	
144	.968	0.2	
103	.906	0.46	
123	.907	1.09	
87	0.97	2.14	
33	0.87	3.06	

Fig. 5c A 10-item rectangular test (Test 3) with uniformly distributed *a* and *b* parameters

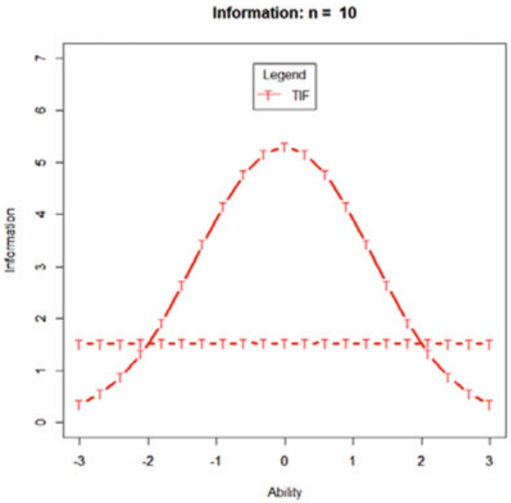
than those near the center. In general, grouping item difficulties and/or increasing item discrimination create peaks in the TIF, while spreading the difficulties and/or decreasing the item discrimination will flatten the TIF. Again, a test with a peaked shaped TIF will be described as a "peaked test," while a test with a flat (rectangular) shaped TIF will be referred to as a "rectangular test."

5 Uses

5.1 Standard Error

An advantage of an IRT model is that its TIF provides an approximate measure of precision for the estimated ability conditional on its value θ:

Fig. 6 Peaked and rectangular TIFs superimposed for comparison



$$SE(\theta) = \frac{1}{\sqrt{TIF(\theta)}}.$$

For example, we can see from TIF for the “peaked test" in Fig. 4a that the information provided by the test for an examinee with ability $\theta = 1$ is approximately $I(1) = 4$, yielding an approximate standard error of the ability estimate of $1/\sqrt{4} = 0.5$. However, for a subject of ability $\theta = 2$, $I(2) = 1$ yielding an approximate standard error of $1/\sqrt{1} = 1$. Therefore, this peaked test has less uncertainty for estimated ability of examinees of ability near $\theta = 1$ than for those with ability near $\theta = 2$.

5.2 *Test Construction and Selection*

Another use of the TIF is in item selection and test construction. A test constructor may use the TIFs to choose among tests that measure best for the targeted range of abilities. Figure 6 displays the TIFs from Fig. 4a and b superimposed on one another. If the test constructor is most interested in extremely low or high ability subjects, a rectangular test may be preferred where the information for those examinees is higher. On the other hand, if subjects in the middle of the ability scale make up the target population, the peaked test may be deemed more useful.

6 Small Sample Information of Ability Estimates from IRT Models

As mentioned above, Fisher information measures the asymptotic precision of the maximum likelihood estimator. Therefore, the TIF is a useful tool for standard error estimation and item selection for large tests. An aim of this chapter is to investigate how well it works for that purposes in short tests. Figure 7 shows the TIFs for tests of 10 to 100 items. Each figure shows two curves:

(1) The solid curve is the "actual" test information, defined as the reciprocal of the variance of the MLE and estimated via simulation using the following steps:

Simulation Method for True Information Estimation

(a) An array of quadrature points was created from $\theta = -3$ to $\theta = 3$.
(b) For each quadrature point, a third-party software named MSTSIM5[2] is used to generate 100,000 subjects of that ability as well to simulate each subject's responses to the test of interest.
(c) Each subject's MLE of ability ($\hat{\theta}$) was calculated using MSTSIM5, producing 100,000 estimates of θ for each quadrature point.
(d) The variance of these 100,000 $\hat{\theta}$s ($\widehat{Var}(\hat{\theta})$) was then estimated for each θ in the set of quadrature points.
(e) The true information for each θ in the set of quadrature points was estimated as $\hat{I} = 1/\widehat{Var}(\hat{\theta})$. We will denote this as the actual test information function ($ATIF_{Sim}$).

(2) The dotted curve is the TIF described earlier in (2). This again is the theoretical test information based on an infinitely long test:

As the number of items decrease, the true test information becomes more discrepant from the TIF. In this example, tests of 100 items have information close to what is indicated by the TIF, especially near the center of the curve, but the difference between the two is considerable for smaller tests and for ability levels significantly distant from the center.

However, the discrepancy between the asymptotic and small test size performance is not present for all tests. Figure 8 compares the TIF and the true test information for a rectangular test of ten items. The figure shows that the small sample performance of estimators of ability from this test nearly matches that predicted from asymptotic theory.

To review, we have seen that when a test comprises a large number of items, the TIF is an accurate assessment of its performance. In that case, the asymptotic theory for IRT models is useful and effective for many practical purposes, from assessing

[2] The FORTRAN routine MSTSIM5 (Jodoin, 2003) was used to simulate student responses and calculate the corresponding MLEs for the given IRT models. R was then used to calculate summary statistics (variance, bias, MSE) for these MLEs.

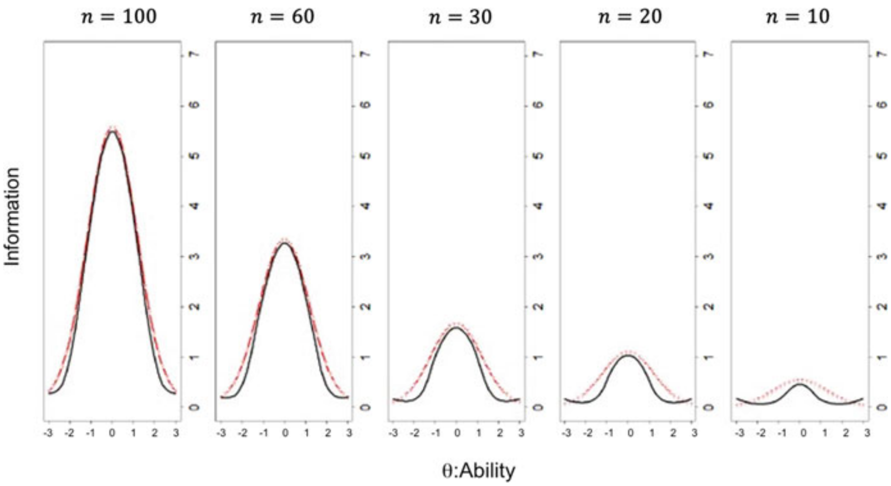
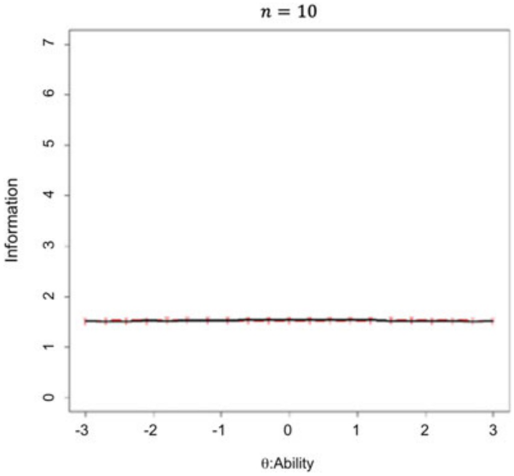


Fig. 7 This figure illustrates how the actual test information (solid black line) increasingly diverges from the theoretical test information (dashed red line) as the test size decreases from $n = 100$ to $n = 10$

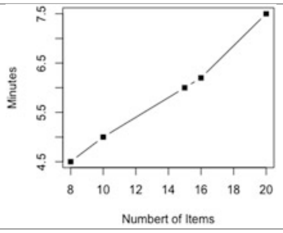
Fig. 8 This plot displays the TIF and empirical information for a ten-item test. Compared to Fig. 7, the empirical information is much closer to the TIF which is expected as the TIF is an asymptotic bound of the information



uncertainty in examinee scores to efficient construction of tests. However, there are practical situations when only a few items can be presented to an examinee. One such example is in large-scale assessment, such as the NAEP, where the testing time available is limited. A second example is in multistage testing, where examinees are routed to subsequent stages of varying difficulty based on their performance on earlier stages of the test (Van der Linden & Glas, 2010). Each stage must necessarily consist of a relatively small number of items, after which an ability estimate must be made to facilitate routing. Finally, some tests produce scores on multiples subscales, so that each one may have only a few items. These are the applications in which we

Table 3 Computation times for the simulation method with scatterplot of computation time versus number of items

Simulation method	
Number of Items	Computing Time
8	4.5 min
10	5.0 min
15	6.0 min
16	6.2 min
20	7.5 min



are interested. For "small tests," which we will formally define in a moment, we have seen that the asymptotic theory often overestimates the true test information especially for peaked tests.

We have seen that the method based on simulation can estimate the actual information of the test although it comes with a considerable cost: time. Table 3 shows the computing time of the simulation method to estimate the actual information with 100,000 simulated subjects. All computing was performed on a 4 GB 2.2 GHz Intel i7 processor Apple MacBook Pro for various test sizes and 30 quadrature points. While wait times are subjective, we see that they are at least 4.5 min for an 8-question test and increase linearly with the number of questions at a rate of .24 min per additional item.

7 Exact Method for Information Calculation

Here we provide an alternative to the asymptotically developed TIF and the time-consuming simulation method described above. This method, which we refer to as the exact method, can be broken down into five steps:

1. Generate all possible response patterns given the number of items.
2. Find the unique MLE for each response pattern.
3. For each true ability (discrete number of quadrature points)
   (a) Find the probability for each unique MLE.
   (b) Make a probability distribution given the MLE and corresponding probability from step 3a.

MLE	Probability
$\hat{\theta}_1$	$P(\hat{\theta}_1\|\theta)$
$\vdots$	$\vdots$
$\hat{\theta}_{n-1}$	$P(\hat{\theta}_{n-1}\|\theta)$
$\hat{\theta}_n$	$P(\hat{\theta}_n\|\theta)$

4. Compute the conditional variance using the equation

$$\sigma_{\hat{\theta}}^2 = \sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i^2 P(\hat{\theta}_i|\theta) - \left[\sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i P(\hat{\theta}_i|\theta) \right]^2$$

5. Calculate the conditional information as $I(\theta) = \frac{1}{\sigma_{\hat{\theta}}^2}$.

Example Consider a test with the following three items:

Item	*a*	*b*
1	1	−2
2	0.5	0
3	0.5	1

Step 1. Generate all possible response patterns given the number of items.

Response pattern	Item 1	Item 2	Item 3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1
5	1	1	0
6	1	0	1
7	0	1	1
8	1	1	1

Step 2. Find the unique MLE for each response pattern. (From MSTSIM5)

Response pattern	Item 1	Item 2	Item 3	MLE $\hat{\theta}$
1	0	0	0	−4
2	1	0	0	−1.75
3	0	1	0	−2.71
4	0	0	1	−2.71
5	1	1	0	0.92
6	1	0	1	−1.74
7	0	1	1	0.92
8	1	1	1	4

Step 3. For each true ability (discrete number of quadrature points)

(Assume the quadrature points are −3, −2.5, −2, −1.5, −1, −.5, 0, .5, 1, 1.5, 2, 2.5, 3.) We will demonstrate the process for the first quadrature point, $\theta = -3$, and this process would be repeated for each of the remaining 12 quadrature points above.

(a) Find the likelihood (probability) for each unique MLE.

For $\theta = -3$, the probability of response pattern one (missing all three questions) is calculated as

$$P(Z|\theta = -3) = \prod_{i=1}^{3} \left(\frac{1}{1 + e^{-1.702 a_i(\theta - b_i)}} \right)^{z_i} \left(1 - \frac{1}{1 + e^{-1.702 a_i(\theta - b_i)}} \right)^{1-z_i}$$

$$= \left(\frac{1}{1 + e^{-1.702\times 1\times(-3-(-2))}} \right)^{1-0} \times \left(\frac{1}{1 + e^{-1.702\times .5\times(-3-(0))}} \right)^{1-0}$$

$$\times \left(\frac{1}{1 + e^{-1.702\times .5\times(-3-(1))}} \right)^{1-0}$$

$$= 0.84580 \times 0.92777 \times 0.96783 = 0.7595.$$

The probabilities for the remaining 12 quadrature points are found in a similar fashion.

(b) Make a probability distribution given the MLE and likelihood (conditional probability) from step 3a.

For $\theta = -3$,

MLE	$P(\hat{\theta}\|\theta)$
−4	0.7595
−1.75	0.1385
−2.71	0.0591
−2.71	0.0252
0.92	0.0108
−1.74	0.0046
0.92	0.0020
4	0.0004

Step 4. Compute the conditional variance using the equation

$$\sigma^2_{\hat{\theta}} = \sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i^2 P(\hat{\theta}_i|\theta) - \left[\sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i P(\hat{\theta}_i|\theta) \right]^2$$

For $\theta = -3$, we have

MLE	$P(\hat{\theta}\|\theta)$	$\hat{\theta}_i^2 P(\hat{\theta}_i\|\theta)$	$\hat{\theta}_i P(\hat{\theta}_i\|\theta)$
−4	0.7595	12.152	−3.038
−1.75	0.1385	0.42415625	−0.242375
−2.71	0.0591	0.43403631	−0.160161
−2.71	0.0252	0.18507132	−0.068292
0.92	0.0108	0.00914112	0.009936
−1.74	0.0046	0.01392696	−0.008004
0.92	0.0020	0.0016928	0.00184
4	0.0004	0.0064	0.0016

$$\sigma_{\hat{\theta}}^2 = \sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i^2 P(\hat{\theta}_i|\theta) - \left[\sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i P(\hat{\theta}_i|\theta)\right]^2$$

$$= 13.226 - (-3.5034)^2 = 0.952.$$

Step 5. Calculate the conditional information as $I(\hat{\theta}|\theta) = 1/\sigma_{\hat{\theta}}^2$.

For $\theta = -3$, $I(\hat{\theta}|\theta = -3) = 1/0.952 = 1.05$.

Note: in order to find the exact value for a particular ability (i.e., for use in a confidence interval or as a standard error of an estimate), simply follow the steps above and make the quadrature point in step 3 the desired ability.

7.1 Constraint on the Use of the Exact Method

While the exact method yields the exact information/variance for the MLE of ability for any test for which item parameters are known, time is still an important factor. Since the method entails calculating the MLE for every possible response pattern, the number of MLEs to calculate doubles for each item added to the test. This equates to an exponential increase in computation time as the number of items increase. Table 4 shows the computing time of the exact method versus simulation time to estimate the same value with the simulation method. Again, all computing was performed on a 4 GB 2.2 GHz Intel i7 processor Apple MacBook Pro for various test sizes and 30 quadrature points. With a computation time of 2 h, the exact method is practically limited to tests under 20 items. However, since pure simulation is quicker than the exact method beginning at 16 items, we will select the exact method for tests of individual ability with 15 items or fewer.

Table 4 The number of response patterns and computation time for the exact method in calculating the true variance of estimates of individual ability

	Exact method		Simulation method
Number of items	Number of response patterns	Computing time	Computer time
8	$2^8 = 512$	8 s	4.5 min
10	$2^{10} = 1024$	13 s	5 min
15	$2^{15} = 32{,}768$	**3.5 min**	**6 min**
16	$2^{16} = 65{,}536$	**7.33 min**	**6.2 min**
20	$2^{20} = 1{,}048{,}576$	2 h	7.5 min

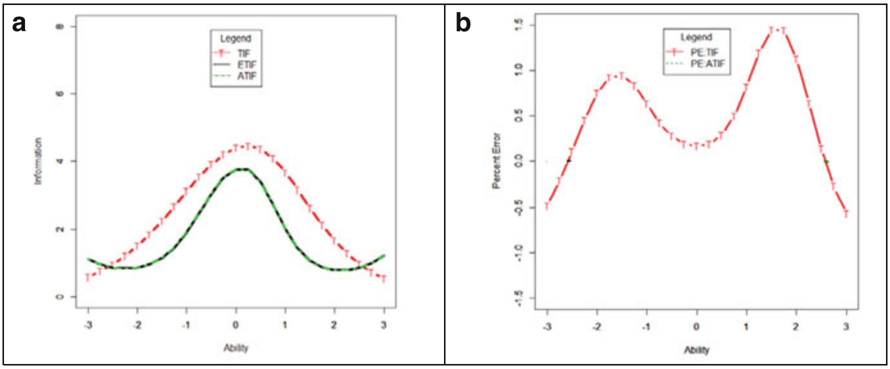


Fig. 9 (**a**) TIF, ATIF, and ETIF for Test 1. The $ATIF_{Exact}$ and $ATIF_{Sim}$ overlap completely; (**b**) PE for the TIF and the ATIF

7.2 *Example: Standard Errors*

Recall that the square root of the reciprocal of the test information function (TIF) is the asymptotic conditional standard error of the MLE of ability (Hambleton et al., 1991). Some standardized tests, such as the STAAR test in Texas and the CST in California, use square root of the reciprocal of the TIF to report standard errors for their estimates (STAAR, 2004). As we showed above, however, there can be a considerable difference between the TIF and the actual test information. This difference could result in standard errors and confidence intervals that incorrectly represent the variability in the MLE, a particularly troubling problem if the intervals are too narrow.

Figure 9a displays the TIF and the actual information for Test 1 constructed in Fig. 5a. The actual test information is defined as the reciprocal of the true variance of the MLE and was computed by the exact method and is referred to as $ATIF_{Exact}$. For confirmation, the simulated value of the actual information was computed as well, using the simulation method described in the introduction. This function, the $ATIF_{Sim}$, is also shown in Fig. 9a and matches the $ATIF_{Exact}$.

Table 5 The PE with respect to the true SE of Test 1 from the small item bank when the goal is to estimate individual ability

θ	TIF	ATIF	*PE*
−3	0.58	1.12	**−0.48**
−2	1.50	0.86	**0.74**
−1	3.11	1.91	**0.63**
0	4.40	3.77	**0.17**
1	**3.67**	**2.02**	**0.82**
2	1.68	0.79	**1.13**
3	0.54	1.23	**−0.56**

An important note concerns the tails of the $ATIF_{Exact}$ and $ATIF_{Sim}$ in Fig. 9a. As mentioned in the introduction, fixed values are assigned to subjects who obtain perfectly correct and incorrect scores ($\theta = -4$ and $\theta = 4$ were adopted for this study). Therefore, as a subject's ability increases (decreases), a larger percent of them begin to obtain perfectly correct (incorrect) scores and therefore receive an MLE of 4 (−4). This in turn causes a decrease in variance as the true ability approaches 4 (−4), thus resulting in an increase in information. The inflection point of the $ATIF_{Exact}$ and $ATIF_{Sim}$ is the ability level at which subjects begin to obtain perfectly correct (incorrect) scores.

We now examine the difference between the TIF and the ATIF more closely by calculating the percent error (PE) between them:

$$PE = \frac{TIF - ATIF_{Exact}}{ATIF_{Exact}}.$$

Figure 9b displays the PE for the TIF and ATIF (exact and from simulation) in Fig. 9a. Table 5 displays the numerical results. Interestingly, the PE of the TIF is as high as 113%, indicating that the TIF is calculating the information to be 113% higher than it actually is! In a practical setting, the exact method would be used to find the desired standard errors which may then be used in the calculation of confidence intervals.

As an example, consider a fictional subject (Sammy) who was trying to qualify for admission to SMU, where the minimum requirement on the entrance exam is a $\theta = 2.1$.

On a 15-question computer adaptive exam, he received a $\hat{\theta} = 1.0$ and was faced with the decision of whether to retake the exam. Being an asymptotic upper bound on the information, the margin of error using the TIF is smaller than the actual margin of error, thus leading Sammy to believe his true ability is between −0.02 and 2.02 (Table 6); he thus abandons his SMU dream and looks at other schools. However, using the exact method ($ATIF_{Exact}$), we are able to calculate the actual standard error which yields a margin of error of 1.38 (Table 7). Sammy would now be led to believe that his true ability is in the interval (−0.38, 2.38), which contains 2.1 and therefore gives him hope! Although he did not pass the first time, given the actual confidence interval facilitated by the $ATIF_{Exact}$, Sammy receives a more accurate measure of the test's uncertainty and, because he believes passing is now possible, may decide to try the entrance exam a second time.

Table 6 Calculations of the margin of error and 95% confidence limits using the TIF to calculate the SE

Name	Margin of error TIF	95% Confidence interval TIF
Sammy	$1.96 \times \sqrt{3.67} = 1.02$	$1 \pm 1.02 \rightarrow (-0.02, 2.02)$

Table 7 Calculations of the margin of error and 95% confidence limits using the exact method to calculate the exact SE

Name	Margin of error TIF	95% Confidence interval TIF
Sammy	$1.96 \times 1/\sqrt{2.02} = 1.38$	$1 \pm 1.38 \rightarrow (-0.38, 2.38)$

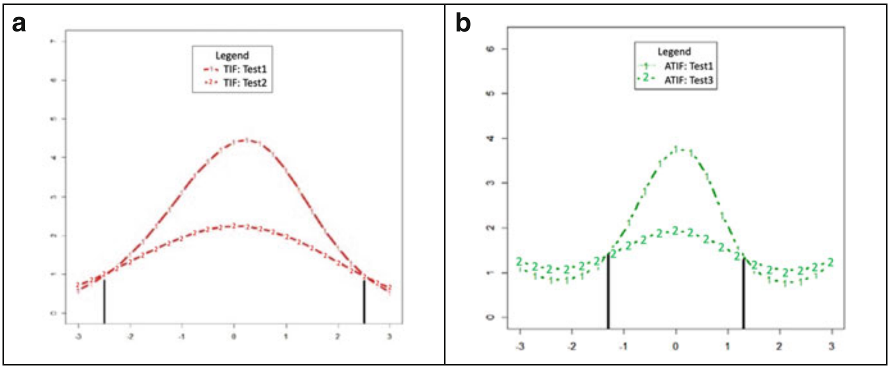


Fig. 10 (**a**) TIFs for Test 1 and Test 2. Test 1 clearly has the higher TIF for the majority of the ability range; (**b**) ATIFs for Test 1 and Test 2. Actual superiority of Test 1 is reduced when the actual information is used

7.3 Example: Test Construction/Selection

This example assumes a practitioner would like to compare two tests, both constructed from the NAEP item bank: Test 1 (very peaked from Fig. 5a) and Test 2 (less peaked from Fig. 5b). Figure 10a displays the TIFs from both tests and could be used as a diagnostic tool to decide between them. Assume the practitioner would like to identify students for a remedial math program and has thus been tasked with finding the best test for estimating abilities between −2.5 and −1.5. Judging from the TIFs in 10a, the practitioner would conclude that Test 1 will provide more accurate results because the TIF (the information) is higher over the target range of abilities. We will show, however, that this is not the right conclusion.

We have established that the TIF is an asymptotic target, but this test is only ten items in length. Thus, the practitioner elects to use the exact method to calculate the variance of the estimator and plots the results for both the tests in 10b. The results show that Test 3 is the more accurate test for his target population, as it is superior for $\theta < -1.3$ and $\theta > 1.3$.

8 Conclusion

Calculation of the asymptotic information of estimates of ability in item response theory is useful for tests with a sufficient number of questions. For tests with few items, however, the difference between the theoretical information and the actual information can be substantial. This chapter focused on the practical scenario in which tests have 15 items or fewer. In these cases, the asymptotic estimate can significantly exceed the truth, leading to significant underestimation of the variability of an individual's estimated ability. A relatively quick, exact method of calculating test information can inform test construction and lead to more accurate confidence intervals for individual ability.

References

Beaton, A. E., Rogers, A. M., Gonzalez, E., Hanly, M. B., Kolstad, A., Rust, K. F., Sikali, E., Stokes, L., & Jia, Y. (2011). The NAEP primer (NCES 2011–463), U.S. Department of Education, National Center for Education Statistics, Washington, DC.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamental of item response theory*. Sage Publications.

Jodoin, M. (2003). MSTSIM5 [computer software]. University of Massachusetts, School of Education, Amherst, MA.

STAAR2004. (2004). *Technical digest chapter 14: Reliability*. http://www.tea.state.tx.us/student.assessment/techdigest/yr0405/

Van der Linden, W. J., & Glas, C. (2010). *Elements of adaptive testing*. Springer.

Statistical Evaluation of Process Variables: A Case Study on Writing Tool Usage in Educational Survey Assessment

Yue Jia and Yi-Hsuan Lee

Abstract For educational and psychological research and practice, the electronic log of test takers' interactions with test questions is referred to as *process data*. Process data hold promise for making inferences about test takers' skills, attributes, and test-taking activities and behaviors. A key to process data analysis is to establish procedures for evaluating the statistical properties of variables of interest derived from such data. Using a real writing test dataset as a case study, this chapter demonstrates the use of a three-step evaluation procedure—including descriptive data analysis and data visualization, statistical modeling, and expert feedback—to evaluate the distributions of process variables for their intended purposes across multiple writing tasks and within persons.

1 Introduction

Administering mental tests using computers and other digital devices has become a common practice. Apart from recording answers to questions, the digital test delivery system can log the action-by-action record of test takers' progress through a test. This information is referred to as *process data*. Bergner and von Davier (2019) framed the distinction between the process data and product data of a test, in that the product data capture where a test taker has ended up (e.g., the correct or incorrect solution or the response to an open-ended question), whereas the process data describe the means to that end. Process data can then be summarized and/or aggregated into variables of interest, such as the total time spent on a test question or the frequency of using a digital tool (e.g., zoom, text highlight color, or digital calculator). In this chapter, we use the term *process variables* to describe summaries of a series of actions recorded in the log.

Y. Jia (✉) · Y.-H. Lee
Educational Testing Service, Princeton, NJ, USA
e-mail: yjia@ets.org; ylee@ets.org

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_13

In practice, a testing program typically designs its own platform to render a test. Process data are used to inform the design of the testing platform (i.e., the user interface and digital tools), to support data quality control, and to monitor test security (Yamamoto & Lennon, 2018). Moreover, process data are used to investigate test-taking behaviors (e.g., Lee & Jia, 2014; Lee & Haberman, 2016; Wise, 2019). For example, Lee and Jia (2014) proposed a method of identifying response-time thresholds that separate rapid-guessing behavior from solution behavior as "the right end of a cluster of short response times whose response accuracy fluctuated around chance level" (p. 8) before response accuracy moved toward that of solution behavior. Expanding further from describing test-taking behaviors, process data provide opportunities to address issues such as validity of the test score interpretation (Kane & Mislevy, 2017). Last, in some applications, the test designers might be interested in considering the process in answering a test question as the product to measure latent traits (Mislevy et al., 2014).

As process variables are more widely used, it is necessary to establish sound evaluation procedures for their intended inference. When assessing individuals with test questions, one might be interested in understanding how test takers interact with a specific question or how test takers' response processes are related to their scores on that question. These types of inferences are question-specific and are not required to be generalizable across different test questions. On the other hand, if one is interested in considering process variables as measures of test-taking behaviors, there should be within-person consistency on performing those actions across the test questions. Further, if test takers are given different test questions, then we should seek to establish comparability of the values of process variables obtained from individual questions.

In this chapter, we focus on the type of inference that considers process variables as measures of specific activities a student performs with the intent to compare them across test questions. We address the following research questions in evaluating process variables:

1. Is there a theoretical argument to meaningfully define and interpret process variables independently from the test questions assigned to test takers?
2. Across-question comparability: Is there empirical evidence that the values of process variables are comparable across test questions, or do they vary by test question?
3. Within-person consistency: Do individuals show consistency in performing actions described by process variables on test questions assigned to them, or do they show random actions on the test questions?

To create and analyze process variables, it is often preferable to start with a research-based theory or cognitive model as it provides hypotheses to construct and evaluate the variables. In our review of the literature, when researchers describe process variables as measures of activities by a subject, there appears to be lack of discussion on procedures for evaluating the statistical properties of such variables. Thus, the focus of the chapter is on research questions (2) and (3), described above.

For illustrative purposes, we describe a case study using a dataset of US 4th-grade writing test collected by an educational survey assessment program. In the next section, we describe the study data, the assignment of the writing tasks to students, and the two process variables of interest and their intended inferences. The description of data and process variables introduces the important idea of establishing a statistical evaluation procedure with proper analysis and modeling approaches. The results of the evaluation application appear in Sect. 3. Conclusions and recommendations follow.

2 Method

2.1 Participants

We used a writing test dataset collected in 2017. A national representative sample of 23,900 US 4th-grade students was selected, of whom 51% were female and 49% male. A plurality (43%) was non-Hispanic white, 17% were Hispanic, 30% were black, 5% were Asian, and 5% belonged to other groups. English language learners (ELs) constituted 12% of the sample, 2% were originally ELs but were reclassified to English-proficient, and 86% were English-proficient. For simplicity, we eschew the use of sampling weights in this analysis. The test is considered low-stakes for the students and schools, as no decisions about the students, teachers, or schools depend on its results.

2.2 Assignment of Writing Tasks to Sampled Students

A total of 22 tasks were used in forming 44 test forms through a partial balanced incomplete block (pBIB) design. More specifically, the design was “balanced,” as each writing task appeared exactly four times, pairing with four other tasks across the 44 forms. Further, the design assigned writing tasks to students in a manner that “balanced” the positioning (first vs. second task given in a test form) of any task across all forms. The forms were “incomplete,” as only two tasks out of the total 22 tasks could fit into a test form. The forms were “partial” in that a task was paired with some but not all other tasks.

Of the 23,900 students, about 550 were assigned to each test form and about 2,200 to each writing task. The pBIB design resulted in comparable groups of students among test forms and writing tasks.

2.3 Process Variable Definitions

The writing tasks were presented to students in a variety of ways including text, audio, video, and images. Before taking the writing tasks, students were asked to take a tutorial to familiarize themselves with the way material is presented on the computer screen and how to use the custom-developed software program, which is similar to common word processing programs. Students' writing responses were scored by trained human readers from 1 to 6 against rubrics defining the relative strengths and weaknesses of the response in relation to specified criteria, with 1 being low and 6 being high.

For the writing assessments, there is a rich body of research on creating and interpreting writing process variables, as well as on connecting the process variables to cognitive theory and models. For example, Baaijen et al. (2012) analyzed several keystroke-logging measures including pause duration, length of burst, and revision. They further inferred from those keystroke-logging measures to cognitive models of writing. Research has also described differences in the writing process (e.g., fluency in text production, frequencies in editing and revision, and extent of between-sentence pauses) between stronger and weaker writers (Bennett et al., 2020; Guo et al., 2018; Sinharay et al., 2019).

In this chapter, we consider two process variables—Formatting and Reviewing. Both variables summarize frequency counts of digital tool usage defined at the task level. They reflect test takers' digital tool usage strategies, whether those strategies aid test takers during the writing process (e.g., Anderson-Inman & Knox-Quinn, 1996) and how the strategies might be associated with writing scores. For this study, we frame the substantive research question as whether the Formatting tool usage or Reviewing tool usage variables can be compared among the 4th-grade students in the study, regardless of which writing forms or tasks the students received.

Formatting tool usage included the following actions—bold (i.e., make text bold), highlight (i.e., highlight text), italic (i.e., italicize text), underscore (underline text and change font size), as well as indent (i.e., move paragraph farther away from the margin) and outdent (i.e., move paragraph closer to the margin). The digital test delivery system counted the number of such events, regardless of whether the student later undid them. The *Formatting* variable was defined as the sum of frequencies of the Formatting keyboard and mouse actions performed on a writing task. For example, if in a writing task a student changed font five times, italicized text two times, and indented text three times, then the value of the Formatting variable would be 10.

The *Reviewing* variable represented use of the spell-checking tools and the Thesaurus. Students accessed the spell-checker by clicking on the tool bar icon or the drop-down menu or by right-clicking on a word identified as misspelled. Students accessed the Thesaurus via the tool bar icon or drop-down menu. Counted actions include the number of spell-checks by any mode, the number of Thesaurus accesses by any mode, and the numbers of spell-check corrections and Thesaurus replacements accepted. For example, if the student opened the spell-checking tool, used it to identify and correct three misspellings, and exited the tool, there would be five actions in total.

2.4 Data Analysis Procedures

When evaluating the statistical properties of process variables, it is helpful to consider the four basic elements suggested by Hoaglin et al. (1983): Resistance, Revelation, Re-expression, and Residuals. *Resistance* refers to insensitivity to localized mis-behavior or abnormal data. *Revelation* relies on visualization to uncover unexpected patterns and behaviors along with familiar regularities. *Re-expression* suggests transforming variables (e.g., by the log or square root) to place them on a scale that is easier to analyze. *Residual* concerns statistical model comparisons and assessment of model fit. DiCerbo et al. (2015) applies the four elements in the discovery of evidence identification rules from both product and process data on system thinking. Here, we propose a three-step data analysis procedure to evaluate the two writing tool usage variables relative to the intended inference of comparing students' test-taking activities across tasks:

(i) Descriptive data analysis;
(ii) Statistical modeling approaches that estimate the test question effect and within-person consistency on the values of process variables;
(iii) Feedback from question-content experts with regard to findings from the first two steps.

Step 1: Descriptive Data Analysis

Computing descriptive statistics and preparing graphs allows us to become familiar with the data, revealing patterns and identifying outliers for further examination. A thorough descriptive analysis can help us generate or refine hypotheses.

Revelation relies heavily on this step. For process variables that are frequency counts or otherwise not normally distributed, some visualization approaches and descriptive statistics are more resistant than others. A common strategy is to consider rank-based measures: the median, trimmed mean, and winsorized mean for location and the interquartile range (IQR) for spread.

It is important to recognize that student actions might reflect factors unrelated to the process one seeks to observe. For example, a high frequency of Formatting actions might result from equipment failure or from the student using Formatting tools for different reasons than the strategy researchers seek to discover. In such situations, trimming extreme frequency counts would reduce irrelevant noise.

Step 2: Statistical Modeling

This step enables data analysts to describe substantive questions in terms of statistical hypotheses about model parameters. The Formatting and Reviewing variables are both frequency counts. As each student had two writing tasks, we consider the two observations from a student as repeated measurements. That is, writing task is the between-person factor (individual writing tasks as categories of the factor), while the position of a task within a form (first vs. second position) is the within-person factor. One can measure within-person consistency by modeling the correlation between the observations from the two tasks.

In this study, we have applied Poisson regression via generalized estimation equations (GEE) (Liang & Zeger, 1986) to model the relationship between the writing tool usage variables and writing task, task position, and the correlation between the two measurements of the tool usage for students. GEE extends the generalized linear model (GLM) to correlated data. We often use it when the outcome of interest is binary or count data. The GEE approach specifies the first two moments of the joint distribution but not its full form. One maximizes a quasi-likelihood function to produce regression estimates that are consistent under mild assumptions; see Agresti (2003, Chapters 4 and 11).

Let y_{ik} denote measurement k ($k = 1, 2$) on subject i. Further, let $x_1, \dots, x_{21}$ represent a vector of indicator variables for 21 of the 22 tasks, treating task "WV" as the reference task, and let the dummy variable x_{22} represent the effect of the first position (relative to the second position) in a test form. For a model describing both task effects and task position effect, the Poisson regression with a log link function can be specified as

$$\mathrm{E}(y_{ik}) = \mu_{ik},\ i = 1, \dots, N,\ k = 1, 2, \tag{1}$$

with

$$\log(\mu_{ik}) = \beta_0 + \sum_{j=1}^{21} \beta_j x_{j,ik} + \beta_{22} x_{22,ik}, \tag{2}$$

where β_0 is the intercept, $\beta_1, \dots, \beta_{21}$ are the regression coefficients associated with the task effects, and β_{22} is the regression coefficient associated with the position effect. Note that the regression coefficients are on the log scale of y_{ik} (counts for tool usage). Equation (2) may be extended to include the interactions of task effects and position effect or reduced to model either task effects or position effect only.

In practice, the observed variability in frequency counts often exceeds that would be predicted by the Binomial or Poisson distribution, a phenomenon referred to as *overdispersion* (Agresti, 2003, Chapter 1). The GEE approach to Poisson regression addresses overdispersion by estimating an additional scale parameter.

With GEE one is also required to assume a working correlation structure. GEE regression coefficient estimates are consistent under all working correlation models and efficient when the assumed model is correct. Several such models are in common use: The independent correlation matrix is specified as

$$\mathrm{Corr}(y_{i1},\ y_{i2}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and a common within-person unstructured correlation matrix is

$$\mathrm{Corr}\,(y_{i1},\ y_{i2}) = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

When there are more repeated measurements per subject, more complicated models are possible. The SAS GENMOD procedure (SAS Institute Inc., 2020) estimates the scale and correlation parameters by the method of moments.

In applications, one may wish to consider a range of models involving different predictors together with working correlation structures. A few evaluation criteria are useful in identifying a best-fitting model. For instance, the QIC (Quasilikelihood under the Independence model Criterion) statistic (Pan, 2001), a generalization of the Akaike Information Criterion (AIC) (Akaike, 1974), can be used for comparing relative model fit in GEE, with a smaller value indicating a better fit. The adjusted QIC, termed QICadj, is a further adaptation that represents the amount of information per observation. Similar adjustments are considered in psychometrics when using information criteria (e.g., AIC, BIC, and log penalty) to compare different item response models (e.g., Haberman et al., 2008). These criteria depend on the number of observations; without the adjustment, small differences in their values across different models may be exaggerated in large samples.

To examine fixed effects in a GEE under an assumed correlation structure, one can use the generalized score test (Boos, 1992; Rotnitzky & Jewell, 1990). For example, the generalized score statistic for the task effects in Eq. (2) is χ^2-distributed with 21 degrees of freedom (DF), and the corresponding generalized score statistic for the task position effect has a χ^2 distribution with 1 DF.

Hardin and Hilbe (2002) recommend three methods for choosing a working correlation structure in GEE:

1. Consider a correlation structure that reflects how the data were collected.
2. Choose a correlation structure that minimizes the QIC statistic.
3. Choose a correlation structure for which the empirical sandwich SE estimates most closely approximate the model-based SE estimates computed under the working correlation structure.

Considering the first method, recall that the working correlation describes the 4th graders' behavior in terms of their tool usage. We anticipate some level of within-person consistency and intend to estimate the extent. Consequently, an unstructured correlation may be a more sensible choice in our study. Considering the third method, it is noteworthy the sandwich SE estimate is the default estimate for the SE of Poisson regression parameters in many statistical packages, including the SAS GENMOD procedure (SAS Institute Inc., 2020). The sandwich SE estimate adjusts the SE estimate based on the working correlation structure (which is referred to as the model-based SE estimate) using a correction that is established from the model residuals. The sandwich SE estimate is a consistent estimate of the SE even if the working correlation structure is mis-specified. For these reasons, it makes sense to consider Hardin and Hilbe (2002)'s third method when comparing GEE models that differ only in the choice of the working correlation structure. The model whose sandwich SE estimates most closely resemble its model-based SE estimates is the model that best represents the data.

With a chosen GEE model, the estimated task effects are generally of primary interest. To assess the practical importance of the estimated task effects across all tasks, we suggest two new measures:

- Range of Prediction (RP_k), which is the range (i.e., difference between the lowest and the highest values) of the predicted task effects at position k ($k = 1$ or 2) among the 22 tasks. We treat task WV as the reference task and position 2 as the reference position. The predicted task effects for the 22 tasks are transformed to the frequency count scale by exponentiating the estimated model parameters:

$$\begin{aligned}\text{RP}_1 = \text{Range}\big[&\exp(\widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_{22}), \exp(\widehat{\beta}_0 + \widehat{\beta}_2 + \widehat{\beta}_{22}), \ldots, \\ &\exp(\widehat{\beta}_0 + \widehat{\beta}_{21} + \widehat{\beta}_{22}), \exp(\widehat{\beta}_0 + \widehat{\beta}_{22})\big], \end{aligned} \tag{3}$$

and

$$\text{RP}_2 = \text{Range}\big[\exp(\widehat{\beta}_0 + \widehat{\beta}_1), \exp(\widehat{\beta}_0 + \widehat{\beta}_2), \ldots, \exp(\widehat{\beta}_0 + \widehat{\beta}_{21}), \exp(\widehat{\beta}_0)\big]. \tag{4}$$

- Relative Range of Prediction at position k (RR_k) is the ratio of Range of Prediction RP_k at position k ($k = 1$ or 2) to the observed IQR of the frequency counts across the 22 tasks:

$$\text{RR}_1 = \frac{\text{RP}_1}{\text{IQR}} \text{ and } \text{RR}_2 = \frac{\text{RP}_2}{\text{IQR}}. \tag{5}$$

Note that RR_k is a global effect size measure of the task effects at position k ($k = 1$ or 2). In this study, the smaller the RR_k values, the stronger indication that the tool usage variables are comparable across all tasks.

Step 3: Feedback from Content Experts

As discussed earlier, it is preferable to design and create process variables from behavioral or cognitive models. Data analysis and modeling approaches are effective ways to check the properties of the variables for their intended uses. Communicating empirical findings with content and cognitive experts informs the interpretation of results and may suggest further research questions for analysis. Empirical data discovery moreover enables content experts to refine their theoretical models.

In the case study, we aim to explore the possibility of treating the two writing tool usage variables as measures of students' test-taking activities. Accordingly, the descriptive statistics and GEE models suggested in steps 1 and 2 largely focus on evaluating cross-task comparability and within-person consistency. If the empirical evidence suggests that a task effect is statistically or scientifically significant, a necessary next step would be to obtain expert opinion and consider models that include a task-by-tool usage interaction.

3 Results

As discussed in Sect. 2, we derived two digital tool usage variables, Formatting and Reviewing, from the writing task data. Each of the 23,900 students in the study sample was assigned two writing tasks. Listwise deletion of individuals with missing values on one or both tasks yielded 46,400 observations from 23,200 students for both variables in the analyses. In this section we describe the results of our analyses.

Step 1: Descriptive Data Analysis
We first examined the descriptive statistics on the frequency distributions of the process variables across the 22 writing tasks. Table 1 shows that the frequency distribution of Formatting tool usage is highly skewed, with a high proportion of subjects never or infrequently using the tools during either 30-min writing task. The situation was similar for Reviewing tool usage. For both variables, the standard deviations (SD) of the frequency counts substantially exceeded their corresponding means.

Table 2 disaggregates the Formatting and Reviewing tool usage data by the position of the writing task in the test form. For the Formatting tools, there was a modest decrease in usage when the writing tasks appeared second. For the Reviewing tools, the frequency counts were comparable across the two positions.

To evaluate whether the students had consistent Formatting or Reviewing tool usages on their two assigned writing tasks, we computed both the Pearson and Spearman correlations of the two sets of frequency counts per variable. The former measures the linear relationship between the students' two frequency counts, and the latter assesses how well the relationship between the students' two frequency counts can be described using a monotonic function. Table 3 suggests that by both

Table 1 Descriptive statistics for Formatting and Reviewing: frequency counts

Variable	Mean	SD	Min	Median	Max	IQR
Formatting	5.6	11.1	0	2	280	7
Reviewing	7.8	10.3	0	4	148	12

Table 2 Descriptive statistics for Formatting and Reviewing by position: frequency counts

Variable	Position	Mean	SD	Min	Median	Max	IQR
Formatting	1	6.4	11.5	0	3	280	8
	2	4.9	10.7	0	1	196	5
Reviewing	1	7.9	10.5	0	4	148	12
	2	7.8	10.1	0	4	141	12

Table 3 Results of Pearson correlation and Spearman correlation between the students' two frequency counts for Formatting and Reviewing

Variable	Pearson correlation	Spearman correlation
Formatting	0.27	0.39
Reviewing	0.60	0.71

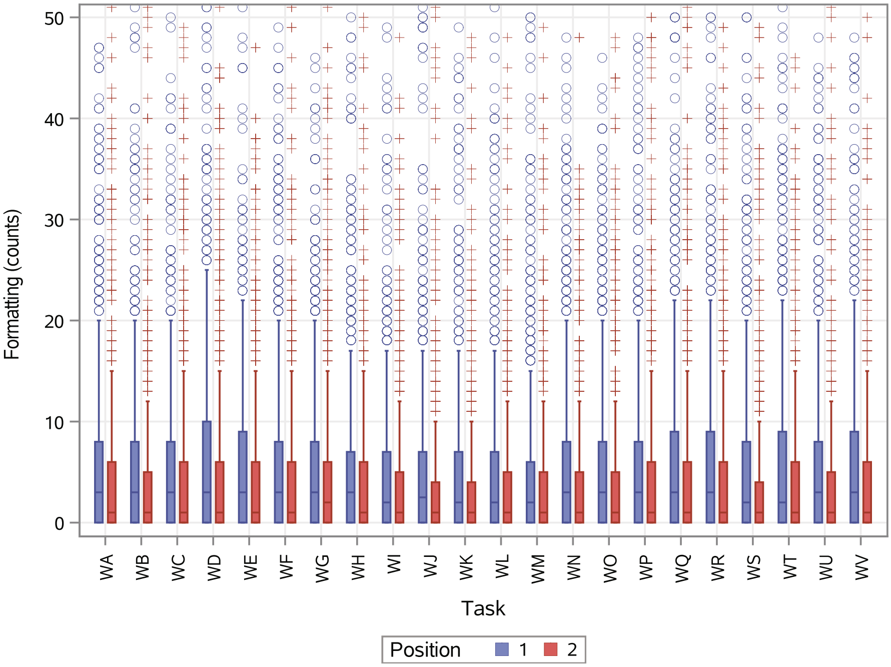


Fig. 1 Boxplots by writing task and position for Formatting, truncated at the 99th percentile (counts=53)

correlation measures the 4th graders used the Reviewing tools more consistently than the Formatting tools.

We visually evaluated the consistency in the two variables across all tasks with boxplots by writing task and position. See Fig. 1 for Formatting and Fig. 2 for Reviewing; vertical axes are truncated at the 99th percentile of the variables for ease of demonstration. The boxes show some variability in IQR across the 22 tasks. Formatting showed a consistent position effect, but Reviewing did not. The patterns agree with what the descriptive statistics suggested; for example, compared to Reviewing, the Formatting tools were less frequently used with respect to the median, but the observed counts had greater variation.

Step 2: Statistical Modeling and Testing

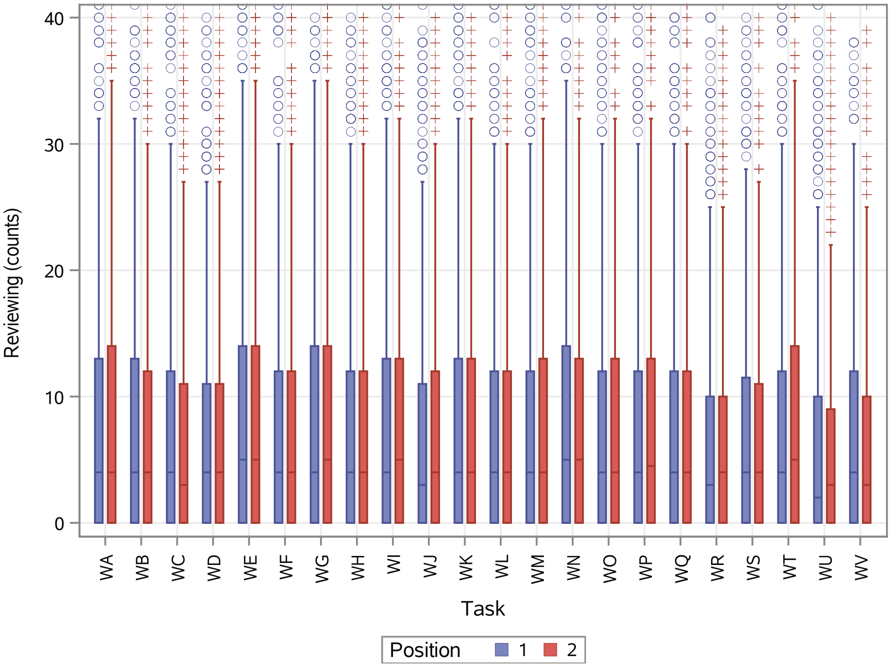


Fig. 2 Boxplots by writing task and position for Reviewing, truncated at the 99th percentile (counts=44)

The descriptive statistics and boxplots from step 1 showed that, for both of the Formatting and Reviewing variables, the spreads of the frequency counts were greater than the corresponding medians or means. Given that the dispersion of the empirical data distribution was greater than a Poisson distribution would be, an overdispersed Poisson regression was used to model both variables. For each variable, four Poisson regression models with a log link function were used to model the frequency counts with two working correlation structures (independent or unstructured) in the GEE approach:

- The first set of models focused on the task effects and included 21 dummy variables $x_1, \ldots, x_{21}$ as predictors for 21 of the 22 tasks (the last task WV was treated as the reference task). They were referred to as Model 1 for independent correlation and Model 5 for unstructured correlation.
- The second set of models focused on the task position effect and included one dummy variable x_{22} as predictor for position 1, assuming that position 2 was the reference position. They were Model 2 and Model 6 for independent and unstructured correlations, respectively.
- The third set of models considered the dummy variables ($x_1, \ldots, x_{21}$) for tasks and position (x_{22}) as two main effects, as expressed in Eq. (2). They were Model 3 and Model 7 for independent and unstructured correlations, respectively.

- The fourth set of models added the interaction terms for tasks and positions to Model 3 and Model 7, referred to as Model 4 and Model 8, respectively, for independent and unstructured correlations.

We estimated these models in the SAS GEMMOD procedure (SAS Institute Inc., 2020). Sample code appears in the Appendix.

We compared the fit of the eight models using the evaluation criteria discussed in Sect. 2.4. The generalized score tests were assessed at the $\alpha = 0.05$ level. For choosing a correlation structure in GEE, the second and third methods of Hardin and Hilbe (2002) were applied to our study, with the expectation that unstructured correlation is likely to be a more sensible choice to model the writing tool usage (i.e., first method of Hardin & Hilbe, 2002). Comparing the results from the two correlation structures (independent vs. unstructured) made it possible to empirically investigate the impact of allowing the correlation to be nonzero on model fit.

Tables 4 and 5 present the results for Formatting. For either correlation type, Table 4 shows that employing different predictors in the Poisson regressions did not affect the estimated working correlation from GEE. Including both task effects and position effect improved both QIC and QICadj, but adding the interaction terms added little. The score test results in Table 5 confirmed that main effects were statistically significant but interactions were not. Thus, we selected Models 3 and 7 for further comparison.

Table 4 also reveals that using the two correlation types led to very close QICadj values for Models 3 and 7 (QICadj: −0.3830 vs. −0.3827). To compare the empirical sandwich SE estimates with the model-based SE estimates, we examined the differences in SE (method-based SE minus empirical SE) for the estimated regression coefficients in Models 3 and 7. Figure 3 shows that the differences in SE were modest (from −0.005 to 0.006) for both correlation types, but those for the unstructured correlation were always closer to 0 (with the exception of task WP).

Table 4 GEE model comparison for Formatting: QIC, QICadj, and estimated working correlation

Model	Corr type	Predictors	QIC	QICadj	Estimated working corr
1	Independent	Task	−17617.23	−0.3797	0.00
2	Independent	Position	−17470.95	−0.3765	0.00
3	**Independent**	**Task, Position**	**−17771.38**	**−0.3830**	**0.00**
4	Independent	Task, Position	−17813.09	−0.3839	0.00
		Task*Position			
5	Unstructured	Task	−17608.28	−0.3795	0.27
6	Unstructured	Position	−17455.44	−0.3762	0.27
7	**Unstructured**	**Task, Position**	**−17760.50**	**−0.3827**	**0.28**
8	Unstructured	Task, Position	−17801.51	−0.3836	0.28
		Task*Position			

Note: Model highlighted in boldface indicates the selected model for each correlation type

Table 5 GEE model comparison for Formatting: score test results

Model	Corr type	Predictors	DFs	χ^2 statistics	p-values
1	Independent	Task	21	125.72	<0.0001
2	Independent	Position	1	282.11	<0.0001
3	**Independent**	**Task, Position**	**21, 1**	**126.81, 284.41**	<**0.0001**, <**0.0001**
4	Independent	Task, Position Task*Position	21, 1, 21	127.54, 295.8, 21.03	<0.0001, <0.0001, 0.46
5	Unstructured	Task	21	139.98	<0.0001
6	Unstructured	Position	1	284.25	<0.0001
7	**Unstructured**	**Task, Position**	**21, 1**	**142.90, 289.35**	**<0.0001, <0.0001**
8	Unstructured	Task, Position Task*Position	21, 1, 21	137.67, 297.61, 26.33	<0.0001, <0.0001, 0.19

Note: Model highlighted in boldface indicates the selected model for each correlation type

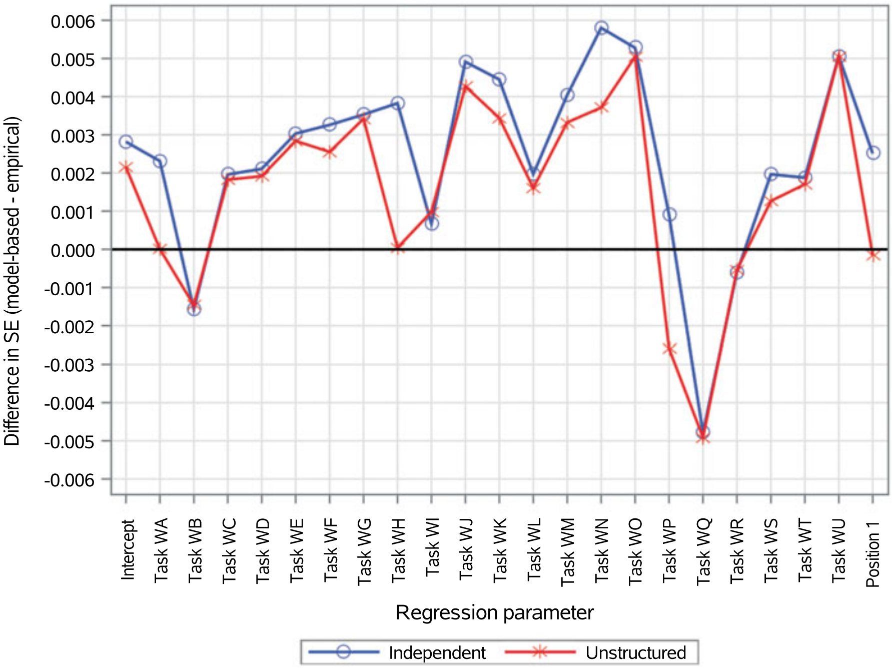


Fig. 3 Comparison of model-based SEs and empirical SEs for Model 3 (independent) and Model 7 (unstructured), Formatting

Figure 4 further shows that the estimated regression coefficients and empirical SEs for the two correlation types were comparable. (We omitted intercepts, where values are much larger, for ease of display.) Some tasks clearly had stronger task effects on the frequency counts of Formatting than others. Because the model-based

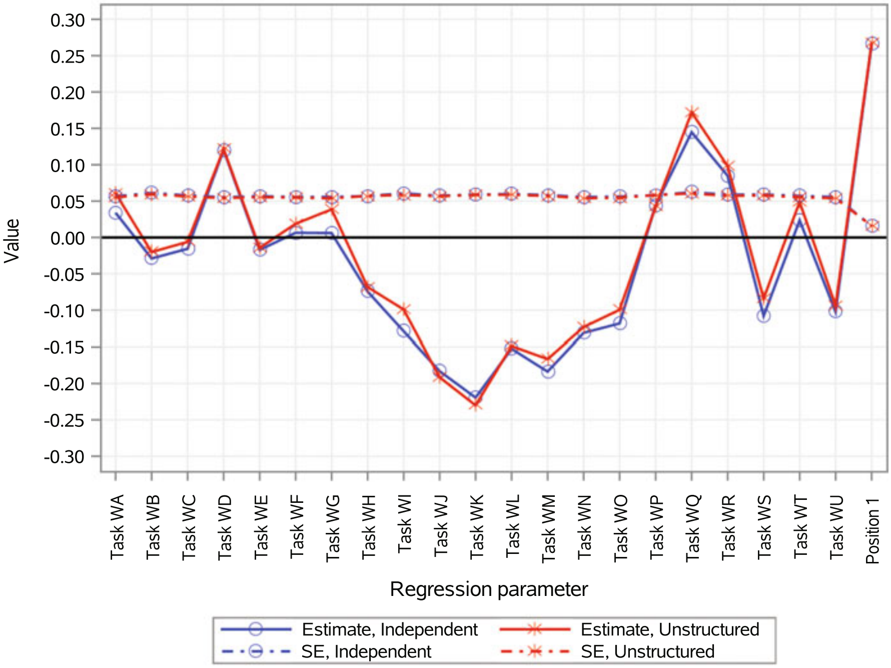


Fig. 4 Comparison of regression coefficient estimates and empirical SEs for Model 3 (independent) and Model 7 (unstructured), Formatting. The regression coefficient estimates are on the log(count) scale

SEs based on the unstructured correlation approximated the empirical SEs slightly better, we chose Model 7 as the final GEE model for Formatting.

Tables 6 and 7 present the results of a similar analysis applied to the Reviewing counts. For either correlation type, considering different predictors in the Poisson regressions did not affect the estimated working correlation from GEE. Task position effect was statistically insignificant, and adding it to the models solely with task effects made a marginal difference in QIC and QICadj, which is not surprising given the findings from the descriptive data analysis at step 1. As a result, we chose Models 1 and 5 for further comparisons.

To compare the empirical sandwich SE estimates with the model-based SE estimates, the differences were again computed as method-based SE minus empirical SE for the estimated regression coefficients in Models 1 and 5. Figure 5 shows that the differences in SE for the unstructured correlation were typically closer to 0 than those for independent correlation. In Fig. 6, the empirical SEs for the unstructured correlation were uniformly smaller than those for the independent correlation, which indicates unstructured correlation was a more efficient choice than independent correlation.

Table 6 GEE model comparison for Reviewing: QIC, QICadj, and estimated working correlation

Model	Corr Type	Predictors	QIC	QICadj	Estimated Working Corr
1	**Independent**	**Task**	**−57963.13**	**−1.2491**	**0.00**
2	Independent	Position	−57233.97	−1.2334	0.00
3	Independent	Task, Position	−57966.40	−1.2492	0.00
4	Independent	Task, Position, Task*Position	−57967.39	−1.2492	0.00
5	**Unstructured**	**Task**	**−57776.88**	**−1.2451**	**0.60**
6	Unstructured	Position	−57062.46	−1.2297	0.60
7	Unstructured	Task, Position	−57782.14	−1.2452	0.60
8	Unstructured	Task, Position Task*Position	−57786.80	−1.2453	0.60

Note: Model highlighted in boldface indicates the selected model for each correlation type

Table 7 GEE model comparison for Reviewing: score test results

Model	Corr type	Predictors	DFs	χ^2 statistics	p-values
1	**Independent**	**Task**	**21**	**275.90**	**<0.0001**
2	Independent	Position	1	0.47	0.49
3	Independent	Task, Position	21, 1	275.90, 0.41	<0.0001, 0.52
4	Independent	Task, Position Task*Position	21, 1, 21	277.02, 0.65, 18.94	<0.0001, 0.42, 0.59
5	**Unstructured**	**Task**	**21**	**370.18**	**<0.0001**
6	Unstructured	Position	1	0.97	0.33
7	Unstructured	Task, Position	21, 1	370.34, 1	<0.0001, 0.32
8	Unstructured	Task, Position Task*Position	21, 1, 21	326.98, 1.12, 18.6	<0.0001, 0.29, 0.61

Note: Model highlighted in boldface indicates the selected model for each correlation type

All but two tasks involved more Reviewing tool usage than the reference task, and the task effects were more comparable across tasks for Reviewing than for Formatting (Fig. 4). Again, the estimated intercepts for the two correlation types were large compared to the rest of the estimates and are not shown.

In summary, using model selection criteria, significance tests, and comparisons of model-based to robust SEs, we chose Model 5 as the final GEE model for Reviewing.

With the final models selected for Formatting and Reviewing, we proceeded to compare their estimated within-person correlations and task effects; see Table 8 for a summary of results. The estimated working correlations between the students' two frequency counts were equal to 0.28 and 0.60 for Formatting and Reviewing, respectively. That implies greater within-person consistency for the 4th graders' Reviewing tool usage than their Formatting tool usage. The estimated scale parameters that were used to address overdispersion in data were equal to 4.66

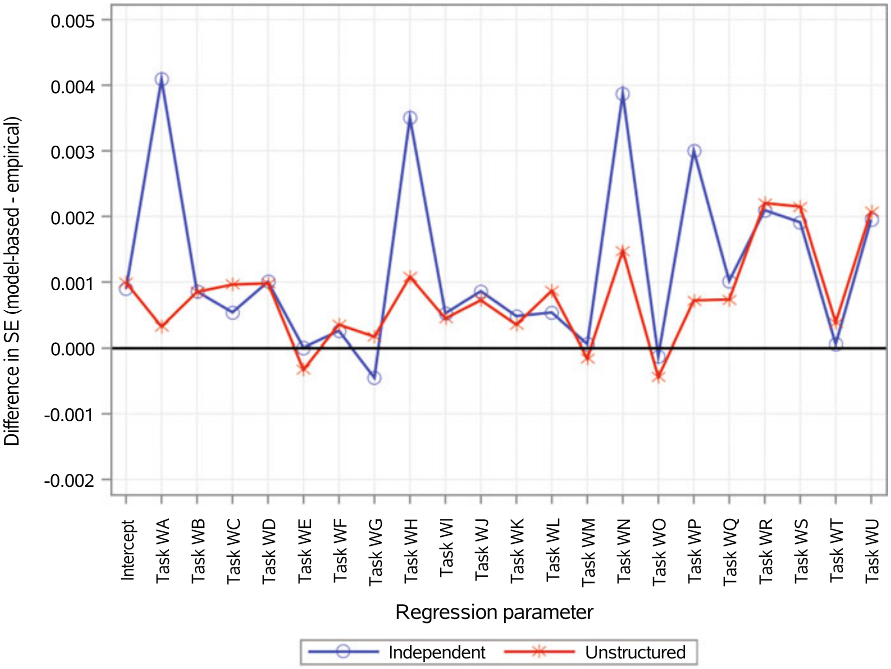


Fig. 5 Comparison of model-based SEs and empirical SEs for Model 1 (independent) and Model 5 (unstructured), Reviewing

for Formatting and 3.66 for Reviewing. Thus, the frequency counts for Formatting were more dispersed than those for Reviewing relative to expectations from the Poisson regressions. Both of these results are consistent with the descriptive results in Tables 1 and 2 and the Pearson correlations in Table 3.

We computed the Range of Prediction (RP_k) and the Relative Range of Prediction (RR_k) for both positions using Eq. (3) through (5). Since both RR_1 and RR_2 for Formatting were greater than their respective values for Reviewing, we concluded that Reviewing tool usage was more consistent across tasks than Formatting tool usage.

Step 3: Feedback from Content Experts

For both variables, descriptive statistics and boxplots clearly showed that the spreads of the frequency counts were greater than the corresponding medians or means. Compared to the Reviewing tools, the Formatting tools were less frequently used with respect to the median, but the observed counts had greater variation. Further, the relative range of predicted task effects (RR_1 and RR_2) from GEE suggested that Reviewing tool usage was more comparable across tasks than Formatting tool usage. The tasks associated with relatively larger predicted task effect would be candidates for a content expert's input.

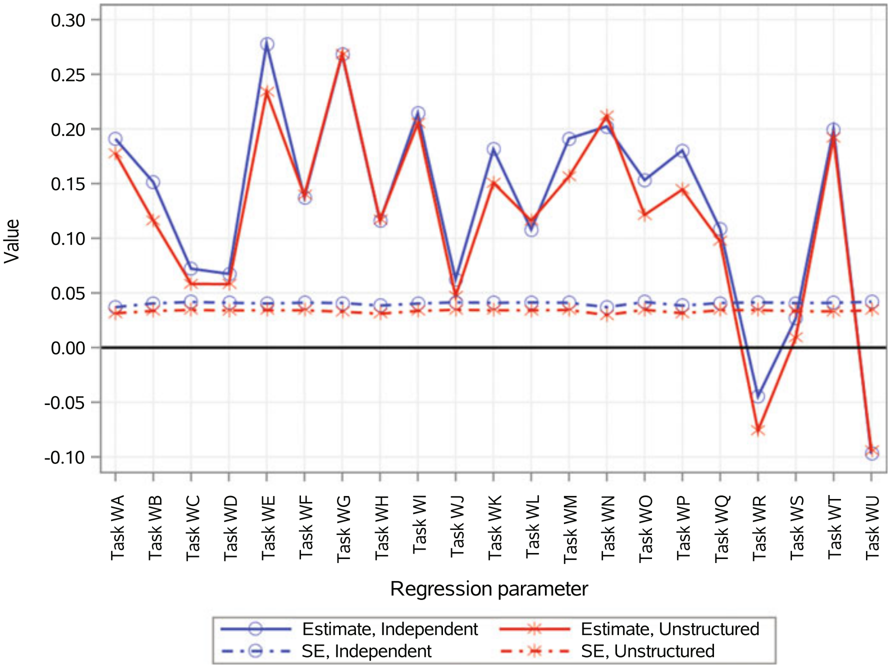


Fig. 6 Comparison of regression parameter estimates and SEs for Model 1 (independent) and Model 5 (unstructured), Reviewing. The regression coefficient estimates are on the log(count) scale

Table 8 Summary of estimated effects based on the final models for Formatting and Reviewing

Variable	Estimated scale	Estimated correlation	IQR (counts)	RP_1 (counts)	RP_2 (counts)	RR_1	RR_2
Formatting	4.66	0.28	7	2.57	1.97	0.37	0.28
Reviewing	3.66	0.60	12	2.77	2.77	0.23	0.23

No task-specific content features were identified to potentially explain the observed distributional differences among the tasks. Nonetheless, we believe it is a necessary step in the procedure to communicate the empirical data findings with content experts and to seek feedback and validate the observations.

4 Summary and Discussion

We have discussed the novel area of analysis and modeling of process data. We have proposed a three-stage procedure whose goal is to systematically evaluate the distributional properties of the process variables for their intended use as measures of students' test-taking activities.

Our case study presented process variables from 22 writing tasks that were assembled into 44 test forms with two tasks per form and with test forms assigned to comparable groups of students. We observed that within-person consistency was good for the Reviewing tool usage variable, which exhibited a within-person correlation of 0.60. For all of tasks, the frequency counts were comparable across the two positions in the 60-min test forms. Boxplots suggested that the frequency counts were comparable across tasks, which was confirmed by the modest value of 0.23 for the Relative Range of Prediction. For the Formatting tool usage variable, there was less within-person consistency (within-person correlation of 0.28), and there was a modest decrease in usage counts when the writing tasks appeared second in a test form. Thus Reviewing tool usage appears to be a more reliable measure of subject test-taking characteristics.

We proposed the Relative Range of Prediction as an effect size measure of the variability of task effects—the smaller the Relative Range of Prediction, the more comparable the variable is across tasks. A potential further research topic is to develop reference values to calibrate the Relative Range of Prediction.

We emphasize that one should consider cognitive theories in defining process variables and in interpreting their analysis results (Provasnik, 2021). Although this chapter has largely focused on the statistical elements of analysis, collaboration with content experts is an essential part of any such study.

The analysis methods that we have demonstrated are in common use across a range of scientific disciplines. They are sufficiently flexible to deal with different types of process variables and are effective in quantifying within-person consistency and potential task effects. The three-step procedure can serve as a precursor and address basic, important questions that researchers should pose early in a study. The four elements of analysis suggested by Hoaglin et al. (1983)—Resistance, Revelation, Re-expression, and Residuals—provide a principled way for practitioners to choose their data visualization and statistical modeling tools.

5 Some Final Comments

We are grateful to Lynne Stokes, who provided valuable guidance to our work.

Acknowledgments This research was funded by the National Center for Education Statistics (NCES) within the Institute of Education Sciences (IES) of the US Department of Education under Contract Award No. ED-IES-13-C-0017. The opinions expressed in this chapter are solely those of the authors and not of ETS, NCES, or any of their affiliates. We would like to thank Daniel F McCaffrey, Gary Feng, and Sooyeon Kim at Educational Testing Service and the anonymous reviewers for valuable comments and edits that improved the clarity and quality of the chapter.

Appendix

This appendix provides sample SAS codes for fitting GEE models for the Formatting variable. Two specific models, Model 3 for independent correlation and Model 7 for unstructured correlation, are chosen for illustrative purposes. It is straightforward to extend the sample codes to other models considered in this chapter.

```
/* Destination of SAS data sets */
libname libref "my_file_path";  *path ends with "\";

/* Define variable for analysis */
%let varlabel=Formatting;
%let var=fmt;

**Model 3: Main effects for task and position,
   independent model;
%let model=M3;
proc genmod data=libref.mydata;
class personid task position;
model &varlabel.=task position/dist=poisson link=log
   type3;
repeated subject=personid/within=position type=ind
   corrw MODELSE ;
ods output GEEEmpPEst=&var._paramest_&model. GEEWCorr=&
   var._WCorr_&model. GEEFitCriteria=&var._QIC_&model.
   GEEModPEst=&var._ModPEst_&model. geemodinfo=&var._
   modinfo_&model. type3=&var._type3_&model. nobs=&var.
   _nbos_&model.(keep=Label N);
run;
quit;


**Model 7: Main effects for task and position,
   unstructured model;
%let model=M7;
proc genmod data=libref.mydata;
class personid task position;
model &varlabel.=task position/dist=poisson link=log
   type3;
repeated subject=personid/within=position type=un corrw
    MODELSE;
ods output GEEEmpPEst=&var._paramest_&model. GEEWCorr=&
   var._WCorr_&model.  GEEFitCriteria=&var._QIC_&model.
    GEEModPEst=&var._ModPEst_&model. geemodinfo=&var._
   modinfo_&model. type3=&var._type3_&model. nobs=&var.
   _nbos_&model.(keep=Label N);
run;
quit;

/* Combine results and compare models */
%macro setfiles(varlabel,var,model);
proc transpose data=&var._QIC_&model. out=&var._QIC_&
   model._v1;
id Criterion;
var value;
run;

data &var._WCorr_&model._v1; set &var._WCorr_&model.;
where RowName="Row1";
rename Col2=WCorr;
drop col1;
run;

data &var._modinfo_&model._v1;
set &var._modinfo_&model.;
where Label1 contains ("Structure");
rename cvalue1=Correlation;
drop nvalue1;
run;

data &var._nbos_&model._v1; set &var._nbos_&model.;
where Label contains ("Used");
rename N=Nused;
run;

proc transpose data=&var._type3_&model. out=&var._type3
   _&model._v0
prefix=predictor;
var source;
run;
proc transpose data=&var._type3_&model. out=&var._type3
   _&model._v1
prefix=DF;
var DF;
run;
proc transpose data=&var._type3_&model. out=&var._type3
   _&model._v2
prefix=ChiSq;
```

```
var ChiSq;
run;
proc transpose data=&var._type3_&model. out=&var._type3
   _&model._v3
prefix=ProbChiSq;
var ProbChiSq;
run;

data &var._type3_&model._v4;
merge &var._type3_&model._v0(drop= _NAME_) &var._type3_
   &model._v1(drop= _NAME_) &var._type3_&model._v2(drop
   =_LABEL_ _NAME_) &var._type3_&model._v3(drop=_LABEL_
    _NAME_);
predset=catx(’,␣’,of predictor:);
run;

data &var._&model.;
merge &var._modinfo_&model._v1 &var._type3_&model._v4 &
   var._nbos_&model._v1 &var._QIC_&model._v1 &var._
   WCorr_&model._v1;
Model="&model.";
variable="&varlabel.";
run;

data &var._parmest_&model.;
merge &var._ModPEst_&model.(rename=(estimate=mod_est
   Stderr=mod_Stderr LowerCL=mod_LowerCL UpperCL=mod_
   UpperCL Z=mod_Z ProbZ=mod_ProbZ))
     &var._paramest_&model.(rename=(estimate=emp_est
   Stderr=emp_Stderr LowerCL=emp_LowerCL UpperCL=emp_
   UpperCL Z=emp_Z ProbZ=emp_ProbZ));
D_Stderr=mod_Stderr-emp_Stderr;
model="&model.";
variable="&varlabel.";
run;
%mend setfiles;
%setfiles(&varlabel.,&var.,M3);
%setfiles(&varlabel.,&var.,M7);


**Model comparison results;
data libref.&var._allmodels;
length predictor1 $13. predictor2 $13. predictor3 $13.;
set &var._M3 &var._M7;
```

```
QIC_perobs=QIC/Nused;
QICu_perobs=QICu/Nused;
run;

**Parameter estimates;
data &var.corrstr; set libref.&var._allmodels;
keep correlation model;
run;
data &var._parmest_allmodels;
length parm $13.;
set &var._parmest_M3 &var._parmest_M7;
run;
data libref.&var._parmest_allmodels;
merge &var._parmest_allmodels &var.corrstr;
by model ;
run;
proc datasets library=libref nolist;
modify &var._parmest_allmodels;
attrib _all_ label='';
quit;

**Export results;
proc export data=libref.&var._allmodels
outfile="%str(my_file_path.GEE results for &varlabel..
   xlsx)"
dbms=EXCEL LABEL REPLACE;
newfile=YES;
sheet="ModelComp";
run;
proc export data=libref.&var._parmest_allmodels
outfile="%str(my_file_path.GEE results for &varlabel..
   xlsx)"
dbms=EXCEL LABEL REPLACE;
newfile=No;
sheet="ParmEst";
run;
```

References

Agresti, A. (2003). *Categorical data analysis*. Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Anderson-Inman, L., & Knox-Quinn, C. (1996). Spell checking strategies for successful students. *Journal of Adolescent & Adult Literacy, 39*(6), 500–503.

Baaijen, V. M., Galbraith, D., & De Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication, 29*(3), 246–277.

Bennett, R. E., Zhang, M., Deane, P., & van Rijn, P. W. (2020). How do proficient and less proficient students differ in their composition processes? *Educational Assessment, 25*(3), 198–217.

Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, Present, and Future. *Journal of Educational and Behavioral Statistics, 44*(6), 706–732.

Boos, D. D. (1992). On generalized score tests. *The American Statistician, 46*(4), 327–333.

DiCerbo, K. E., Bertling, M., Stephenson, S., Jia, Y., Mislevy, R. J., Bauer, M., & Jackson, G. T. (2015). An application of exploratory data analysis in the development of game-based assessments. In *Serious games analytics* (pp. 319–342). Springer.

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement, 55*(2), 194–216.

Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions. *ETS Research Report Series, 2008*(2), 1–25.

Hardin, J. W., & Hilbe, J. M. (2002). *Generalized estimating equations*. Chapman & Hall/CRC.

Hoaglin, D., Mosteller, F., & Tukey, J. (Eds.) (1983). *Understanding robust and exploratory data analysis*. Wiley.

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In *Validation of score meaning for the next generation of assessments* (pp. 11–24).

Lee, Y.-H., & Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing, 16*(3), 240–267.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education, 2*(1), 1–24.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13–22.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A. A., & Hao, J. (2014). *Psychometric considerations in game-based assessment*. GlassLabGames.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics, 57*(1), 120–125.

Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest? *Large-Scale Assessments in Education, 9*(1), 1–17.

Rotnitzky, A., & Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika, 77*(3), 485–497.

SAS Institute Inc. (2020). *SAS/STAT 15.2 user's guide*. SAS Institute Inc.

Sinharay, S., Zhang, M., & Deane, P. (2019). Prediction of essay scores from writing process and product features using data mining methods. *Applied Measurement in Education, 32*(2), 116–137.

Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education, 32*(4), 325–336.

Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education, 26*(2), 196–212.

Index

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4